Research Master's programme:
Methodology and Statistics for the Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

Msc Thesis Thijs Carrière (5843545)
Using Process Data in the Detection and Explanation of Differential Item Functioning
May 2022

Supervisors:
Dr. Remco Feskens (CITO & University of Twente)
Dr. Dylan Molenaar (University of Amsterdam)

Second grader:
Dr. Ir. Mirjam Moerbeek (Utrecht University)

Preferred journal of publication: Applied Measurement in Education
FETC number: 21-1994
Word count: 7107

Using Process Data in the Detection and Explanation of Differential Item Functioning

Thijs C. Carrière[a]

[a]Utrecht University, the Netherlands

ARTICLE HISTORY
Compiled May 9, 2022

ABSTRACT
Differential Item Functioning (DIF) is a broadly studied property of test items. DIF can be an unwanted form of bias, but can also indicate a meaningful difference. To distinguish between the two, the source of the DIF needs to be explained. With the growth of computerized testing, process data have become more widely available. Process data are data that are collected during the response process but are not the response itself, such as response times. These data might hold additional information that is not captured in the response itself, and might therefore be useful in explaining DIF. In this study, a new model based on the MIMIC model is suggested that can be used to investigate whether DIF can be explained with process data. The model is first tested with simulations, where the sample size, strength of the DIF, latent differences between groups, and the number of DIF items are altered. Next, the model is applied to empirical Trends in International Mathematics and Science Study (TIMSS) data to see whether DIF in TIMSS can be explained with response times. Based on the findings it is discussed how the proposed model can be used to explain DIF with process data, but that the current state of process data might form a hindrance for wider use. Further research with a wider variety of process data is recommended.

KEYWORDS
Differential Item Functioning; Process data; Response times; MIMIC; TIMSS

1. Introduction

Over the last decades, educational assessment is transitioning from paper-based assessment (PBA) to computer-based assessment (CBA) (Burkhardt & Pead, 2003). More recently, international large-scale assessments (ILSAs) started using more and more of the full possibilities of digital assessment, gaining access to a new type of data additional to response data: process data. Process data are data that are collected during the process leading to the response, such as response time, number of clicks or whether an optional help tool is used (Molenaar, 2015; Wools et al., 2019). Process

data could consequently be useful in creating more understanding of the process leading to a response of a student (Han et al., 2021), which can hold didactically important information (Wools et al., 2019) and could allow for improved measurements of tests (Molenaar, 2015; Zhang et al., 2021).

Currently, measurement models occasionally include process data. This is the case for example for response times, which are a popular and commonly collected type of process data (Molenaar et al., 2015). Response times are assumed to contain information about the latent trait measured that is not covered by the responses itself. Several psychometric measurement models have been developed to include both responses and response times in one model (Entink, 2009; Molenaar et al., 2015; van der Linden, 2007).

The rise of CBA also opens the possibility for more complex test items, such as technology enhanced items (TEIs). TEIs are test items that make use of the (digital) assessment platform itself by including aspects of technology such as media, interactivity, or response methods that go beyond traditional assessment methods (Bryant, 2017; International Test Commission & Association of Test Publishers, in press). TEIs make the collection of more and newer process data possible, such as the order in which sub-questions are answered or whether a certain strategy was used for getting to the response. With both response times and newer types of process data however, the use and correct way of analyzing these data are not always clear (Qiao & Jiao, 2018; Tang et al., 2021; von Davier et al., 2019).

One test property that is important for both traditional test items and TEIs is that probabilities of giving a correct response on an item should be stable over groups, given an equal level of a latent trait. When groups have unequal probabilities of responding correctly on a test item, despite an equal score on the latent trait intended to be measured, the test item is subject to Differential Item Functioning (DIF; Hambleton and Rogers, 1989). DIF can either indicate a meaningful difference between groups or an unwanted nuisance factor (Ercikan, 2002; Kalaycioğlu & Berberoğlu, 2011). To distinguish between these two, it is important that the source of the DIF can be explained. An example of when DIF can be explained and forms a meaningful difference could be the case when one group is more proficient on a subtopic of a test than another group because of policy choices that differ over the groups. In that case, the probability of a respondent answering items related to this subtopic correctly might be higher while the overall proficiency scores are equal, but a fair comparison can still be made between the groups. However, when DIF forms a nuisance factor the DIF items would be biased and cannot be used to compare groups fairly. These items should either be deleted from the test or should be changed so the bias disappears (Hagquist, 2019). When items are deleted because of DIF, the test loses part of the measurement quality, validity and reliability (Hambleton, 2006). Adjusting test items and item parameters to deal with DIF is therefore preferred, so the test stays intact. However, these adjustments need to be made carefully, in order to not introduce additional bias. When DIF cannot

be explained, it is unclear whether the DIF is unwanted. Therefore, entanglement of DIF sources is important to ensure fair comparisons (Sireci & Rios, 2013). In ILSAs, the groups compared on average proficiency level are often countries and important policies are based on these outcomes. Therefore, DIF detection and explanation form important steps of the analyses of these studies. Biased items must be identified and dealt with to ensure valid comparisons.

The domain of DIF has been studied extensively (Zumbo, 1999, 2007) and multiple techniques to identify items with DIF have been developed (Bechger & Maris, 2015; Gao, 2019; Mellenbergh, 1989). All these methods are based on investigating the relation between group membership and response probability conditional on the measured latent trait. However, all current detection techniques do not distinguish between meaningful DIF and DIF as undesired bias. One way to explain DIF is by including covariates on the person or group level that can explain the difference in response probability (Choi et al., 2014; Hagquist, 2019). This way of explaining DIF is in line with the way Ackerman (1992) sees DIF. He states that DIF is the problem of measuring additional latent traits beside the one intended to measure. Process data might contain information on these additional traits. For example, response times might correlate with reading ability in a mathematical test item. Process data might therefore be useful in the explanation of DIF, and a new use for process data might be found in making a distinction between useful DIF and DIF as bias.

The current study will address the link between process data and DIF, which to the best of our knowledge has not been studied yet. The objective is to investigate the possible use of process data in making the distinction between meaningful DIF and DIF as a form of bias. This will be studied based on the research question "*To what extend can process data be used in the explanation and detection of Differential Item Functioning?*". Since no earlier studies have explained DIF with process data, a new method needs to be developed in order to answer the research question. This will be done within this study. The research question will be studied with the use of the data of the ILSA *Trends in International Mathematics and Science Study* (TIMSS; Fishbein et al., 2021). It is important to note that before DIF can be explained, it must be detected first. Therefore, DIF detection will be executed on the TIMSS test items before process data is used to explain the DIF. In the past, DIF has already be found in ILSA test items (Avşar & Emons, 2021; Choi et al., 2014; Feskens et al., 2019; Özdemir, 2015), and it is therefore expected that some of the TIMSS items will be subject to DIF as well. Since covariates have been used to explain DIF in the past, process data are expected to function in a similar way and it is expected that process data will indeed explain detected DIF to a certain extent.

The outline of this paper will be as follows: In section 2 models on DIF detection and process data modeling will be discussed. The MIMIC model will be central, since it is the DIF detection technique used in this study. Based on the discussed models, a new model is proposed that can be used to model process data as an explanation for

DIF. In sections 3 and 4, the proposed model is evaluated with the use of simulated data. Additionally, to see if DIF in TIMSS can be explained with process data, and to evaluate the use of the model on an empirical example, the proposed model is used on TIMSS data. We will discuss the implications of the findings in the discussion.

2. Models for DIF and process data

In order to investigate process data as an explanation for DIF, a new method needs to be developed. Such a method should be based on current DIF detection frameworks, so explanation can be done in a same model as the detection. Furthermore, the way of modeling the relation between process data and responses should be considered when developing a new method. Several measurement models that model these relations already exist. In this section, both DIF detection techniques and measurement models are discussed before a new model is proposed that can be used to investigate process data as an explanation for DIF.

2.1. Models for detection of DIF

As mentioned earlier, a wide variety of methods for detection of DIF exists, where each method has advantages and disadvantages over other detection methods (Gao, 2019). For example, methods differ on the number of groups they can compare and the types of DIF that they can detect (uniform DIF and non-uniform DIF; Mellenbergh, 1983). A consideration must be made in what features of DIF detection are important in the current study to determine what DIF detection method is most suitable to use. Since ILSAs often compare multiple countries, it is desirable to have a detection method that can handle more than two groups simultaneously. Examples of DIF detection methods that can only compare two groups at the time are the Mantel Heanszel procedure (Mantel & Haenszel, 1959), the logistic regression procedure (Swaminathan & Rogers, 1990), and Lord's Chi-square test (Lord, 1980; Lord, 1977). Examples of detection methods that can compare multiple groups on the other hand, are the MIMIC model (Jöreskog & Goldberger, 1975), the Generalized Mantal Haenszel procedure (Landis et al., 1978; Penfield, 2001), and Bayesian approximate measurement invariance testing (Muthén & Asparouhov, 2013; Van De Schoot et al., 2013). Additionally, since the DIF needs to be explained, a model in which process data can be added is preferred. Of the mentioned techniques above, the MIMIC model is one that can incorporate covariates to explain DIF (Chun, 2014; Woods, 2009). Because of these two requirements it is decided to use the MIMIC model for the detection of DIF within this study. Although we chose for the MIMIC model, other frameworks, such as Bayesian approximate measurement invariance testing, could have been used in this study as well.

2.2. The MIMIC model

Multiple Indicators Multiple Causes models (MIMIC; Jöreskog and Goldberger, 1975) are a form of structural equation models (SEM) that can be used to detect DIF. In these models, a number of items (i) are loading on a latent variable (θ). The latent variable is then regressed a grouping variable (z). Lastly, the item under investigation for DIF is regressed on the grouping variable. This model can be written as:

$$y_i^* = \lambda_i \theta + \beta_i \gamma z + \beta_i z + \epsilon_i. \tag{1}$$

Here, y_i^* is the underlying latent variable for the response on item i (Wirth & Edwards, 2007). λ_i is the factor loading of item i on the latent variable θ, which is analogous to the discrimination parameter used in IRT models (Bulut & Suh, 2017), z is the grouping variable, β_i is the regression coefficient from item i on grouping variable z, and γ is the regression coefficient from the regression of the latent variable on grouping variable z, indicating the mean differences between groups on the latent trait. Lastly, ϵ_i is the random error for item i. A general assumption of the MIMIC model is that the variance of θ, ζ, is equal over the different groups under investigation. This assumption is taken as given within this study. The investigated item i is subjected to DIF when $\beta_i \neq 0$. As explained by Wirth and Edwards (2007), the latent underlying variable y_i^* is a normal continuous variable and its response category (c) in item i is determined based on threshold τ, where

$$y_{ij} = c \text{ if } \tau_{jc} \leq y_{ij}^* < \tau_{jc+1}. \tag{2}$$

In this basic MIMIC model, grouping variable z shows the case where two groups are compared. If more than two groups are compared, multiple grouping variables z should be added to the model, following dummy coded groups. Figure 1 shows a MIMIC model with n groups. Furthermore, this MIMIC model can be used to investigate DIF in multiple test items at the same time, by estimating multiple β-parameters. However, this can not be done for all test items at the same time. At least one anchor item needs to be identified where the β-parameter is set to 0 and is not estimated.

The basic MIMIC model has received several additions over the years. Features added to the MIMIC model are the ability to detect DIF over more than two groups (Dell-Ross, 2021) and the possibility to add covariates to the model to account for DIF (Chun, 2014; Woods, 2009). So far these covariates have often been on person level, such as gender or Socio Economic Status (Chun, 2014). Using covariates on item level, such as process data is uncommon.

Another addition to the MIMIC model is its ability to detect both uniform and non-uniform DIF, in comparison to the original MIMIC model that only can detect uniform

DIF (Woods & Grimm, 2011). A test item has uniform DIF when the difference in the logit of the probability of giving a correct answer is constant over the latent trait. On the other hand, an item is subject to non-uniform DIF when there is an interaction between group membership and the latent ability. For example, when one group has a higher probability of answering an item correctly when having a low score on the latent ability, while the other group has a higher probability of answering correctly when having a high score on the latent ability, the item has non-uniform DIF (Güler & Penfield, 2009). The MIMIC model that is able to detect both non-uniform and uniform DIF can be written as:

$$y_i^* = \lambda_i \theta + \beta_i \gamma z + \beta_i z + \omega_i \theta z + \epsilon_i. \tag{3}$$

Here, ω_i is the regression coefficient from interaction term θz between grouping variable z and latent trait (θ) on item i, and the other terms are equal to equation 1. The investigated item has non-uniform DIF when $\omega_i \neq 0$. The MIMIC model that both detects uniform and non-uniform DIF is shown in Figure 1.

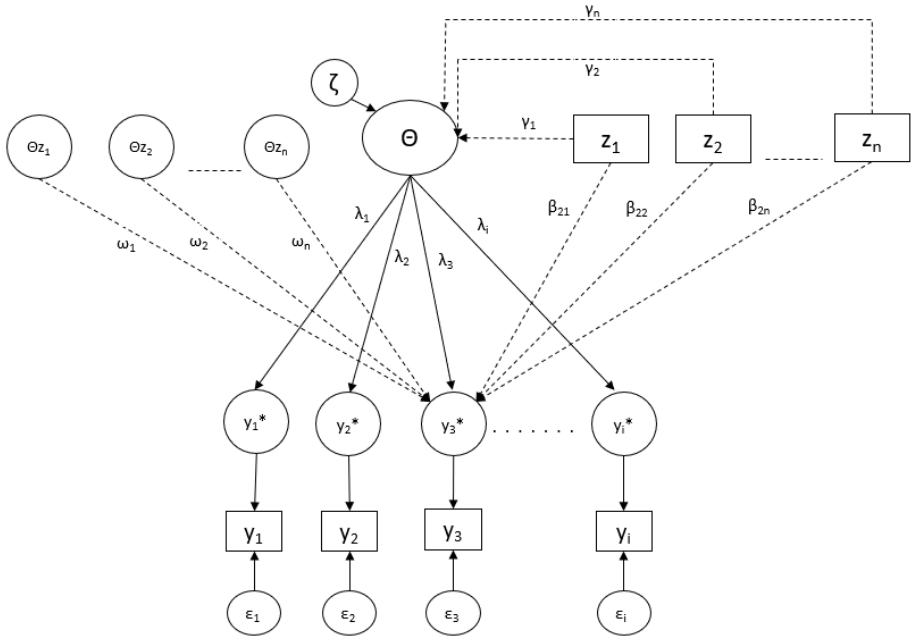

Figure 1.: *MIMIC model to investigate uniform and non-uniform DIF. In this model, DIF is checked for item 3. Dashed arrows indicate regression.*

2.3. Models on response and response time

In the past, various models have been developed that combine both information on the response and response times in their estimation of the latent trait, using the additional information that response time can hold about the measured trait (Entink, 2009; Molenaar et al., 2015; van der Linden, 2007). Van der Linden (2007) gives a summary of some models and develops his own hierarchical framework to model both responses and log-response times. In this framework, for test taker j taking a test both a vector of responses ($\mathbf{U}_j$) and a vector of log-response times ($\mathbf{T}_j$) are obtained. These vectors lead to separate latent variables for the responses (θ_j) and for the response times (τ_j). On a second level of the framework, the two latent variables are modeled jointly on a population level.

Molenaar et al. (2015) developed a frequentist adaptation of the framework by Van der Linden which is referred to as the bivariate generalized linear item response theory framework (B-GLIRT framework). Here as well, θ_j and τ_j are latent variables for the responses and response times respectively, forming a two factor model. The two factors are then linked by cross-relations, which can take multiple forms. Like van der Linden (2007), this framework models log-response times. This frequentist framework has several advantages (Molenaar et al., 2015): it allows for common used models to be used in this framework, and multilevel structures can be modeled easily, which is desirable for educational data that often has a multilevel structure. In addition, the framework allows for covariates to be modeled as well, since it is a form of SEM modeling. This last advantage is especially useful in this study. Therefore, when response times are modeled in this study, the B-GLIRT framework is used.

2.4. A new model for DIF and process data

As explained, a new method to explain DIF with process data needs to be based on both DIF detection techniques and measurement models. In this study a model is proposed that adds process data to a standard MIMIC model in the form of a item-level covariate. Within the model, the item flagged with DIF is regressed on both the grouping variable and the process data of that specific test item. Since the model tries to explain DIF with individual difference on item level, possible differences on group level should be accounted for. This can be accounted for in two different ways. The first way is to regress the process data on the grouping variable. Adding this relation accounts for any group level differences in process data. The other way is by centering the process data on person means. In this study, the method of person-mean centering is used.

DIF is (partly) explained by the process data if the regression of item on the process data is significant, and the β-parameters and ω-parameters become non-significant, compared to the MIMIC model where DIF for the item was detected. The proposed model is shown in Figure 2. In this figure, the method to account for group-level

difference by regressing the process data on the grouping variables is shown.

Since the proposed new model is based on measurement models specifically developed for response times, this study will be limited to the use of response times as process data. However, the model can be used for other process data as well, sometimes with little adjustments to apply measurement models for other types of process data.

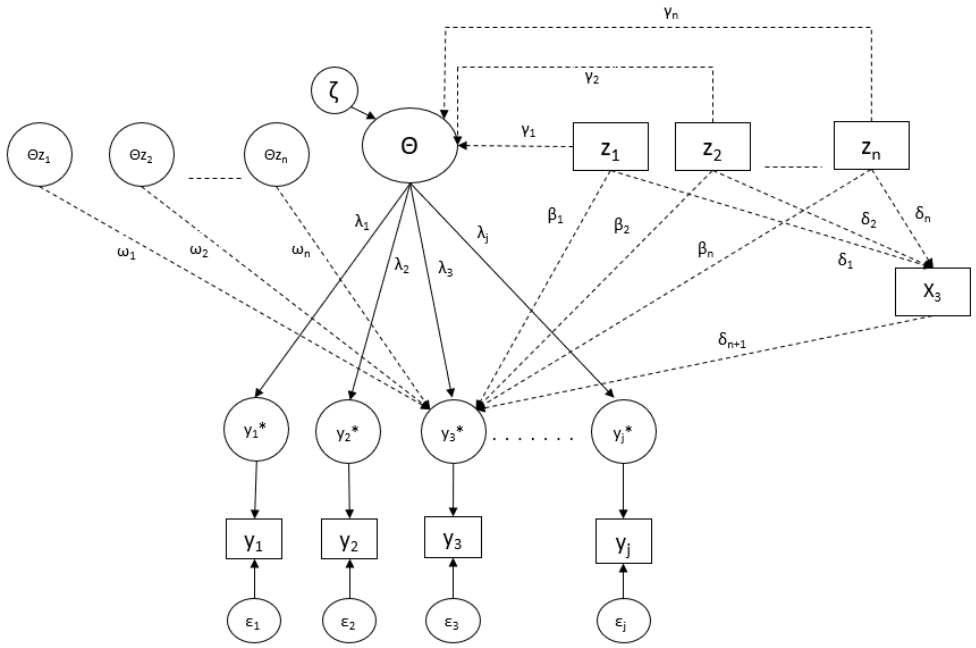

Figure 2.: *Process data of item 3 as a covariate (x_3) added to the MIMIC model, explaining the DIF in item 3. Dashed arrows indicate regression.*

3. Methods

3.1. Simulations

In order to form a better understanding of the functioning of the proposed model and to see how often DIF is correctly explained, the proposed model is tested on simulated data. Following the standard TIMSS analyses, no distinction is made between uniform and non-uniform DIF in the detection, and the factors detecting this difference, θz, are not included in the model. In the official analyses of TIMSS, higher level IRT models are used, but no distinction is made between uniform and non-uniform DIF (von Davier, 2019). Within all simulations within this study, log-response times are used as process data.

The simulated data follows the situation where two groups made a test of 30 test items. The number of items in the test is based on the number of items used in TIMSS booklets. For each item, a response and a response time are generated for

each observation. The responses, 1 for correct and 0 for incorrect, are obtained by a binomial distribution, where the probability parameters are obtained by a 2PL IRT-model (Birnbaum, 1968). The 2PL model can be written as

$$P(X_i = 1|\theta, a_i, b_i) = \frac{e^{a_i(\theta)-b_i}}{1 + e^{a_i(\theta)-b_i}}, \tag{4}$$

where the formula gives the probability for a correct response on item i. θ is the latent ability score of the student, and b_i and a_i are the difficulty and discrimination parameter of item i, respectively. For each simulation the b-parameters are sampled from a standard normal distribution, and the a-parameters are sampled from a uniform distribution ranging from 1 to 3.

The response times are simulated following van der Linden (2007), using a normal linear homoscedastic factor model. This model can be written as

$$ln(T_{pi}) = \lambda_i - \tau_p + \epsilon_{pi}. \tag{5}$$

Here $ln(T_{pi})$ is the log-response time for person p on item i, λ_i is an item specific intercept for item i, and ϵ_{pi} is the random error with VAR$(\epsilon_{pi}) = \sigma_{pi}$. Lastly, τ_p is a person specific latent variable underlying the response times. Having a higher latent response-time variable results in faster response times.

Following Molenaar and de Boeck (2018), the latent ability variable θ and the latent response time variable τ are correlated with $r = .4$, indicating that people with faster responses tend to have a higher latent ability and thus more correct answers. Both the latent variables are obtained with the use of a multivariate normal distribution, $\mathcal{N} \sim (\boldsymbol{\mu_g}, \boldsymbol{\Sigma})$. Here, $\boldsymbol{\Sigma}$ is a covariance matrix that represents the correlation of .4 between the latent variables, and $\boldsymbol{\mu_g}$ represents the group means on both variables. These group means may differ between groups but are the same for both variables within one group.

A number of the items is subject to DIF, introduced by a strategy factor. This strategy factor, δ_p, is drawn from a separate normal distribution for each DIF item, $\delta_p \sim \mathcal{N}(\mu_{stg}, 1)$. Here μ_{stg} differs over the two groups, resulting in an easier item for one of the two groups. With the strategy factor included, equation 6 follows from equation 4 and can be written as

$$P(X_i = 1|\theta, a_i, b_i) = \frac{e^{a_i(\theta)-b_i+\delta_p}}{1 + e^{a_i(\theta)-b_i+\delta_p}}, \tag{6}$$

where δ_p is the strategy factor for person p. For one of these items, the strategy factor also influences the corresponding response times. This means that the DIF for this particular item can be explained by the response times, whereas for the other DIF items this is not the case. With this factor included, equation 7 describes how the

strategy factor is added to the response times. The equation follows from equation 5 with the inclusion of δ_p, and can be written as

$$ln(T_{pi}) = \lambda_i - \tau_p + \epsilon_{pi} + \delta_p. \quad (7)$$

3.1.1. Conditions

The model is evaluated in different conditions. Four variables are changed over the conditions, following a 2x2x2x3 model. This results in 24 conditions, that are summarised in Tables 1 and 2.

Firstly, the sample size is set to 500 and 2,000 observations per group. Secondly, the number of items with DIF is altered. There is always only one test item where DIF is present that can be correctly explained by response times. However, several extra items are modeled to have DIF as well. In one condition four items in total are subject to DIF and in the other condition eight items are subject to DIF. Third, the difference between the two test groups on the latent traits (for both the response and the response times) differs over the conditions. This difference is captured with parameter μ_g. In one condition, it is set to 0 for both groups. In the second condition it is set to 0 for the first group, but to 1 for the second group, indicating 1 standard deviation difference on the latent traits. Lastly, the strength of the strategy factor, and therefore the strength of the DIF, differs over the conditions. This parameter, μ_{stg}, is set to 0, .5, and 1 for one group, and set to 0 in all conditions for the other group. This illustrates the cases of no DIF, weaker DIF, and stronger DIF, respectively. The first condition functions as a control condition.

For each condition we run 100 iterations. In each iteration, first a normal MIMIC model is run to see if the correct items are detected as having DIF. Second, the proposed model is run and it is investigated whether the DIF correctly disappears in the one item where response times should explain the DIF. For both steps, an α of .05 is used. To account for general group differences in response times, the response times are person mean centered. The person means are calculated with the exclusion of the item that is under investigation for DIF, in this case the item where the DIF can be explained. This choice and method follows from the subsection 2.4.

As outcome summary statistics, it is indicated in how many of the iterations DIF is detected and in how many of the cases DIF is detected and also explained. Here, only the result of the item that has DIF that can be explained by the reaction times is considered, and the outcome of the other DIF items is not taken into account. Relative explanation rates are also calculated. This is the percentage of models in which DIF is explained when only the models are taken into account where DIF was detected in the first place. The outcomes give an indication of the power of the proposed model. Iterations with non-converged models are excluded in the calculations.

3.2. Empirical Data

3.2.1. TIMSS Data

The data used in this study are data collected in the 2019 iteration of the TIMSS (Fishbein et al., 2021). These data are publicly available[1]. TIMSS is an ILSA that investigates the mathematics and science knowledge of 4th and 8th graders since 1995 in cycles of four years. Both mathematics and science are measured by multiple choice and open ended items, where each test version is consisting of 24 to 36 items per domain (Mullis & Martin, 2017). In addition to mapping proficiency in mathematics and science, a number of surveys are conducted on school, student and teacher level to collect contextual information.

In the TIMSS assessment of 2019, 58 countries participated resulting in over 230,000 students per grade being assessed world wide (Martin et al., 2020). Since 2019, participating countries have the choice to administer TIMSS by computer (eTIMSS) or still opt for the PBA. During the assessment of eTIMSS, process data are gathered for the students.

Within TIMSS several test versions (so called booklets) are used. The main reason for using different test versions is to limit the number of items that each student has to answer, while maintaining a full coverage of the complete mathematics and science domain. Within this study we make use of the mathematics items of the first TIMSS booklet for grade 8. The mathematics items are chosen over the science items because mathematical questions have more diverse answer processes that could result in meaningful DIF. The responses of seven eTIMSS countries are compared to identify test items that are subject to DIF. The countries used are England, France, Hong Kong, Korea, Norway, Qatar, and Taiwan, resulting in 1,951 observations. These countries are selected because they have relatively different educational systems (Hofstede, 1986). These differences can result in meaningful DIF, which could then be explained by process data.

For the analyses to explain DIF with process data, we use the response times collected in eTIMSS. Process data in eTIMSS are available for every student-item interaction. In eTIMSS response times are measured in seconds.

3.2.2. Pre-processing of data

Before the data were in a feasible format for the analyses, several steps were taken. Firstly, the data of the different countries were taken together as one data set. New person identifiers were created by combining the identifier for country and student, since student identifiers were not unique over countries. Then, scores were computed based on the given answers. A score of 0 was given for a false or missing answer and 1 for a correct answer. For questions existing of multiple sub questions, one combined

[1]The TIMSS data can be obtained from https://www.iea.nl/index.php/data-tools/repository/timss

score was given. Response times were then converted to a compatible format to the response data. For the last item of the test, unrealistically high response times were found. They were replaced with a capped value of 400 seconds. Lastly, the data were put in a format fit to be put in a Dexter database to make the data easier accessible and make analyses possible (Maris et al., 2018).

3.2.3. Detection of DIF

The first step of the analyses is the detection of DIF items. This detection will be done with the use of a MIMIC model, where for each item is investigated whether it has DIF over the seven countries. A test item is classified as having DIF when the DIF model fits significantly better than the base model (the model without an item regressed on the group variable). This is checked with a χ^2-difference test, where an α of .05 is used. A Bonferroni correction is applied to the p-values to prevent capitalization on chance, since the analysis is run for each item separately (31 times in total). When DIF is found in an item, it is investigated for what countries this is the case by looking at what β-parameters differ significantly from 0.

3.2.4. Modeling process data

In order to answer the research question with empirical data, response times are modeled for each of the test items found to be subject to DIF. In accordance with the model and the simulations, the response times were person-mean centered. Again, the person means were calculated with the exclusion of the item under investigation. Therefore, modeling of the process data follows the proposed model used in the simulation study as described in the subsection 2.4. For each item it is investigated if β-parameters that were significant in the detection of DIF have become non-significant. If this is the case, the DIF is (partly) explained by the process data.

3.3. Software and Ethics

All of the analyses will be performed with R (R Core Team, 2022) in RStudio, version 2021.9.0.351 (RStudio Team, 2021). All structural equation models are run with the lavaan package, version 0.6.10 (Rosseel, 2012). The scripts for all analyses can be found on the public GitHub[2] of this project. Ethical approval for this study was obtained through the FETC of Utrecht University.

[2]The public GitHub page for this project can be found at https://github.com/ThijsCarriere/Master-Thesis

4. Results

4.1. Simulation outcomes

The summarized outcomes of the simulations can be found in Table 1 ($n = 500$) and Table 2 ($n = 2,000$). When looking at the strength of DIF, Table 1 and Table 2 show that detection is best with strong DIF. However, in cases of a smaller sample size, the detection of strong DIF drops compared to cases with a high sample size (.54 to .81 and .98 to .99, respectively). For weaker DIF, the detection rates are lower for both cases with a high sample size and cases with a smaller sample size (.59 to .81 and .18 to .27, respectively). The drop in DIF detection because of the sample size is stronger for the conditions with weaker DIF compared to the conditions with stronger DIF, indicating an interaction between sample size and strength of DIF regarding the detection of DIF. For the difference in latent ability, there seems a small effect where there is a higher detection rate in cases with no latent difference between the groups. This effect is not clearly present however and could be further investigated. Lastly, the number of DIF items does not seem to have an effect on the DIF detection rates.

Considering the rates of both correct detection and explanation of the DIF-item, sample size shows an effect, with relative explanation ranging from .831 to .929 and from .700 to .947 for conditions with a high sample size and conditions with a smaller sample size, respectively. There is an interaction between sample size and DIF strength when looking at the explanation rates. For the conditions with stronger DIF, sample size does not influence the explanation rates. For weaker DIF however, relative explanation rates drop when comparing the smaller sample size to the higher sample size. The number of DIF-items does not play a role in the explanation rates. This might however be different when the outcome of more than one item is considered. The rates of correct decisions for all the items would probably be lower in that case. The difference in latent ability does not have an effect on the explanation of DIF either.

All the effects on the explanation of DIF are also visible in the absolute explanation rates. They are clearer in the relative explanation rates however, since they take the DIF-detection rates into account. These relative explanation rates are also more relevant, since DIF detection will always be executed first before DIF will be explained and explanation will only happen for cases where DIF is detected.

The conditions without DIF function as expected. The DIF detection rates range from .01 to .07 and are around the set α of .05. The rates of both detection and explanation are even lower, ranging from .00 to .02. The DIF is thus still explained in some of these cases, and has higher relative explanation rates than expected. This is because in these simulations the strategy still influences both the outcome and the response time. The groups do not differ on their average strategy factor, but faster response times are still related to better outcomes, explaining the significant explanations in the model when DIF is incorrectly identified. None of the run models showed non-convergence.

Table 1.:
Simulation conditions and simulation outcomes in percentages of correct outcomes per condition, sample size = 500.

Condition	Number of DIF-items	Ability difference (SD)	DIF strength (SD)	DIF-detected	DIF-detected and explained	Time and DIF significant	Relative explanation-rate
1	4	0	0	.03	.01	.02	.333
2	4	0	.5	.27	.23	.04	.852
3	4	0	1	.81	.74	.07	.914
4	4	1	0	.07	.02	.05	.286
5	4	1	.5	.20	.14	.06	.700
6	4	1	1	.68	.63	.05	.926
7	8	0	0	.07	.01	.06	.143
8	8	0	.5	.18	.13	.05	.722
9	8	0	1	.76	.72	.04	.947
10	8	1	0	.02	.02	.00	1
11	8	1	.5	.22	.16	.06	.727
12	8	1	1	.54	.49	.05	.907

Table 2.:
Simulation conditions and simulation outcomes in percentages of correct outcomes per condition, sample size = 2,000.

Condition	Number of DIF-items	Ability difference (SD)	DIF strength (SD)	DIF-detected	DIF-detected and explained	Time and DIF significant	Relative explanation-rate
13	4	0	0	.03	.00	.03	.000
14	4	0	.5	.76	.69	.07	.908
15	4	0	1	.98	.82	.16	.837
16	4	1	0	.03	.01	.02	.333
17	4	1	.5	.59	.49	.10	.831
18	4	1	1	.98	.84	.14	.857
19	8	0	0	.01	.00	.01	.000
20	8	0	.5	.81	.75	.06	.923
21	8	0	1	.99	.87	.12	.879
22	8	1	0	.02	.00	.02	.000
23	8	1	.5	.66	.61	.05	.924
24	8	1	1	.98	.91	.07	.929

4.2. Empirical outcomes

General descriptives of the seven used countries are presented in Table 3. The table shows that the countries differ in their mean response times and mean accuracy. It also shows that faster countries have higher accuracy ($r = -.91$). On an individual level, there seems no relation between accuracy and mean response time ($r = .00$). On item level, correlations between correct response and response times ranged from $r = -.22$ to $r = .47$. Summary statistics of the test items are reported in Table 4. Interesting is the high mean response time for the last item (ME72209). When this item is considered on country level, the mean times range from 106 to 216 seconds, which is higher than all the other questions. This might indicate either an interesting difference or a mistake in the measurement of the reaction times.

To have an idea what items might give relevant outcomes in the DIF explanation and DIF detection, Figures 3 and 4 show the score and time per county on each item respectively. In Figure 3 item ME72002 differs a lot from the general trend. The non-Asian countries are expected to score higher since it is a 2-point question, but score even lower than average, indicating possible DIF. Especially England scores different than expected in this question. In Figure 4 item ME72209 is remarkable. The order is almost the opposite of the other items and of what could be expected given the general trend. This item was already marked as standing out considering the summary statistics as well. However, the accuracy of this item shows no sign for strong DIF. Another notable item in Figure 4 is item ME72017, where the speed order deviates a lot from the other items with Taiwan being the slowest country while for the other items Taiwan is with the faster countries. In Figure 3, ME72017 does not have a remarkable order. However, the ratios between the countries is different for this item than the overall ratios.

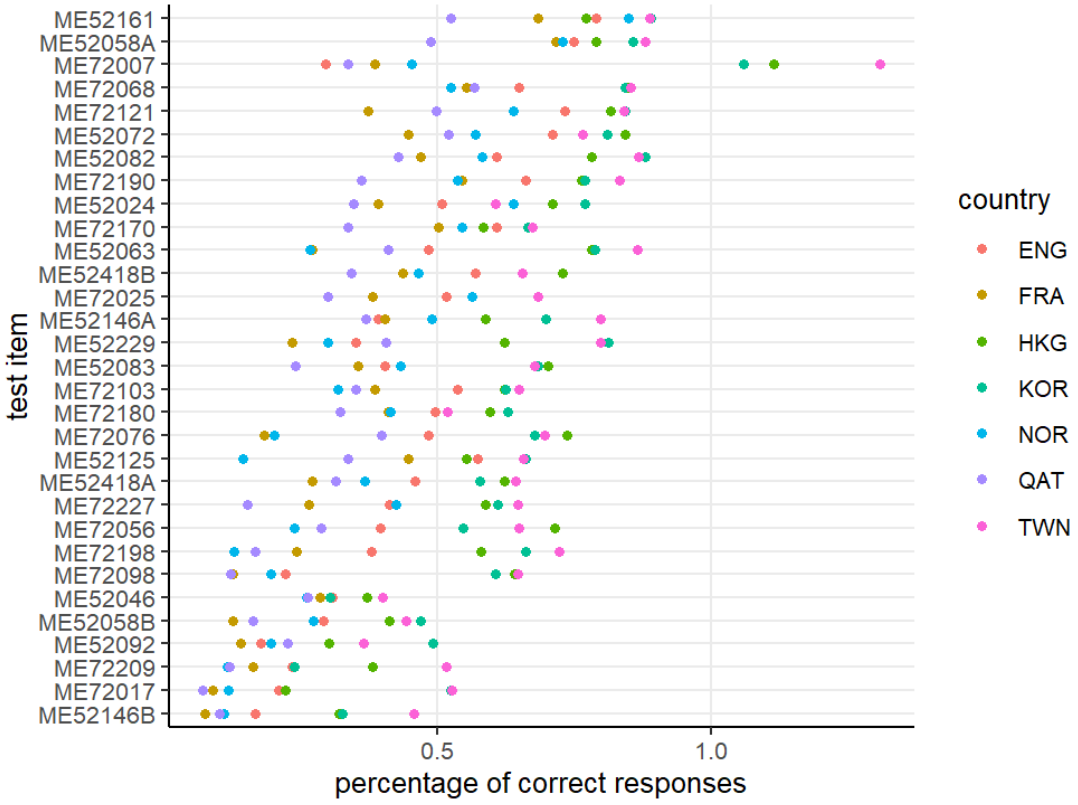

Figure 3.: *Accuracy per country per test item.*

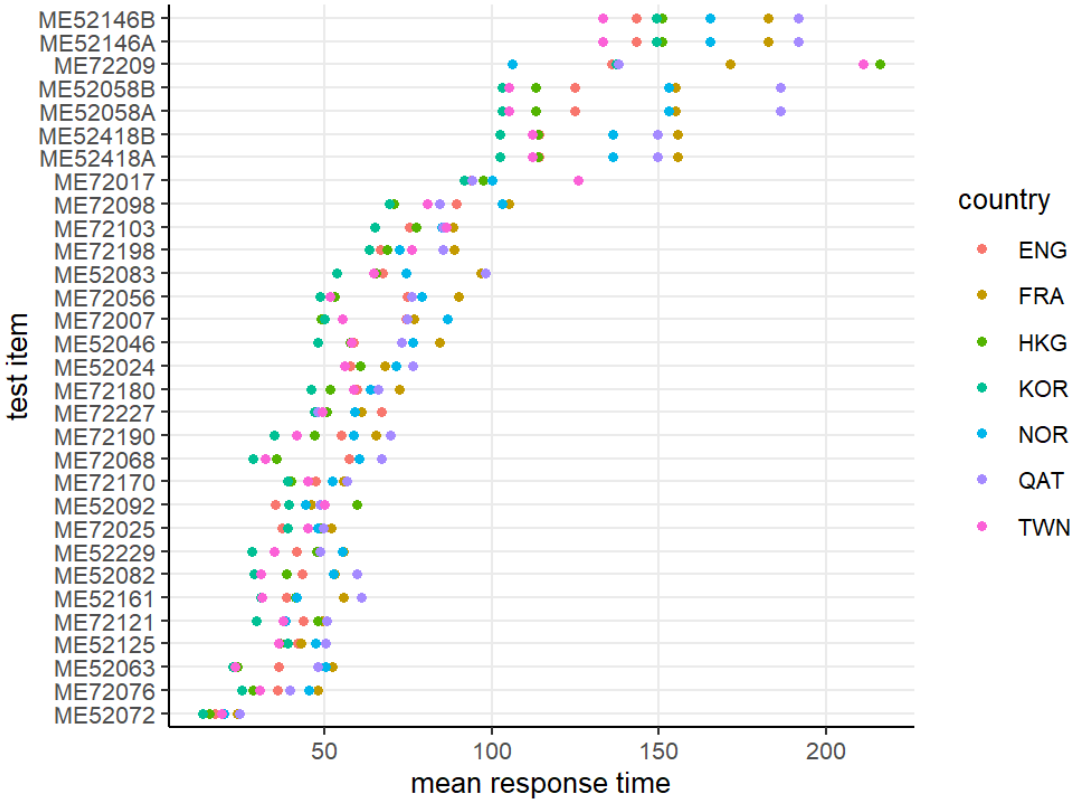

Figure 4.: *Response time (in seconds) per country per test item.*

Table 3.:
Summary statistics of the e-TIMSS countries used in the analyses.

	P-value		Response-time		Correct answers	
Country	Mean	SD	Mean	SD	Mean	SD
England	.464	.503	71.5	61.2	14.4	7.41
France	.343	.482	88.6	73.4	10.6	5.89
Hong Kong	.638	.510	70.5	70.2	19.8	7.69
Norway	.391	.495	80.7	71.6	12.1	6.25
Qatar	.319	.472	88.1	80.4	9.9	6.82
South Korea	.666	.500	60.8	58.2	20.7	8.04
Taiwan	.694	.499	68.8	64.2	21.5	8.21

Table 4.:
TIMSS test item summary statistics. Items are in order of appearance in the test.

Item	Mean p-value	Mean response-time	time-response-correlation
ME52024	.568	63.8	.10
ME52058A	.747	134.3	-.06
ME52058B	.314	134.3	-.14
ME52125	.478	42.2	.01
ME52229	.509	44.5	-.11
ME52063	.553	36.8	-.22
ME52072	.663	19.4	-.10
ME52146A	.544	159.1	.11
ME52146B	.228	159.1	-.05
ME52092	.274	46.2	.13
ME52046	.315	65.5	.08
ME52083	.502	74.4	-.07
ME52082	.663	43.8	.05
ME52161	.776	42.7	-.10
ME52418A	.466	126.5	.13
ME52418B	.548	126.5	.10
ME72007	.721	67.1	.06
ME72025	.547	45.8	.04
ME72017	.261	100.4	.27
ME72190	.640	53.0	-.08
ME72068	.690	49.7	-.03
ME72076	.478	36.4	.00
ME72056	.437	67.6	.04
ME72098	.369	86.6	.14
ME72103	.497	80.9	.01
ME72121	.678	42.1	-.05
ME72180	.481	60.0	-.07
ME72198	.414	74.7	-.02
ME72227	.447	54.4	.16
ME72170	.561	48.1	-.01
ME72209	.257	159.6	.47

4.2.1. DIF detection

The outcome of the DIF detection analysis is shown in Table 5. Notable is that every item is flagged as having some form of DIF for one or more countries. This might indicate that the MIMIC model is too sensitive in DIF detection. This could be supported by the fact that even items with only one country deviating from the general trend, such as ME72180 where the DIF is only due to Taiwan scoring different than expected, are flagged with DIF. There is no pattern in the DIF, so none of the countries has consistently more DIF or only DIF in one direction.

4.2.2. DIF explanation

The results of the DIF explanation are summarized in Table 6. During the DIF detection all items where found to have DIF for one or more countries. Therefore, the explanation analysis was run for every test item. The last column of Table 6 shows the relation between response time and a correct response. To see whether DIF is (partly) explained by the response times, only items where response times have a significant relation with answering correctly are relevant. Therefore, the outcomes of items that had no significant relation with response time are not shown in the table.

For most items, there are small changes, but these changes do not alter the detection of DIF and for these items response times do not explain the DIF. For items 52146A and 52092 countries that previously had no DIF are detected as having DIF when response times are added to the model. Here again, response times do not explain the DIF. Item 52082 has the countries with DIF and without DIF switched when response times are added. For some items, the DIF partly disappears for some countries (for example 52072, 52418A, and 72170). However, since there is no clear pattern, and the effect can be inconsistent over different countries in the same item, these changes might not be explanation of the DIF.

Table 5.:
Estimates for the β-parameters and overall χ^2-difference test as outcomes of the DIF detection in TIMSS.

Item	χ^2-difference *(df = 7)*	England	France	Hong Kong	Norway	Qatar	South Korea	Taiwan
ME52024	102.56***	-.042	-.013	.087	.502***	-.083	.202**	-.352***
ME52058A	74.14***	.175*	.435***	-.140	.326***	-.134*	.054	.077
ME52058B	62.92***	-.007	-.195*	-.222**	.196*	.042	-.157**	-.328***
ME52125	148.65***	.338***	.342***	-.143*	-.738***	.097	.082	-.005
ME52229	121.01***	-.299***	-.250***	-.084	.-.208**	.315***	.429***	.281***
ME52063	151.72***	-.062	-.308***	.351***	-.454***	.134	.303***	.535***
ME52072	57.58***	.317***	-.017	.296***	.150*	.233***	.085	-.156**
ME52146A	51.32***	-.278***	.047	-.150*	.154*	.003	.095	.345***
ME52146B	77.78***	-.233**	-.272**	-.303***	-.239**	-.034	-.377***	-.121*
ME52092	43.39***	-.229***	-.216*	-.185*	-.091	.148	.273***	-.112
ME52046	41.93***	.003	.131*	-.133	.008	.147	-.372***	-.167**
ME52083	38.85***	-.162*	-.008	.263***	.083	-.303***	.152*	.072
ME52082	43.63***	-.035	-.077	.074	.088	-.133	.410***	.279***
ME52161	100.55***	.210**	.200**	-.260**	.631***	-.179*	.163	.075
ME52418A	28.06***	.103	-.005	.009	.105	.189**	-.201***	-.115*
ME52418B	27.57***	.162*	.145*	.185*	.087	-.052	-.099	-.172**
ME72007	118.71***	-.692***	-.116	.069	-.137*	-.148*	-.093	.178***
ME72025	47.99***	.047	.054	.040	.386***	-.120	-.036	-.117*
ME72017	111.95***	-.172*	-.333***	-.619***	-.322***	-.391***	.154*	.071
ME72190	36.79***	.199**	.268***	.025	.094	-.158*	-.035	.114
ME72068	46.64***	.079	.172*	.290**	-.041	.266***	.198*	.164*
ME72076	125.34***	-.095	-.434***	.344***	-.511***	.281***	.090	.066
ME72056	61.52***	-.007	-.032	.299***	-.208**	.197**	-.263***	-.087
ME72098	66.58***	-.371***	-.278***	.189**	-.178**	-.216**	-.007	-.008
ME72103	36.89***	.215**	.192**	-.038	-.144*	.158*	-.110	-.129*
ME72121	59.95***	.300***	-.293***	.109	.257***	.095	.137	.046
ME72180	24.11**	.100	.121	.052	.034	-.081	.088	-.252***
ME72198	86.66***	.001	.027	-.023	-.594***	-.182*	.116*	.195***
ME72227	53.96***	.052	.056	-.035	.340***	-.286***	-.064	-.062
ME72170	54.46***	.236**	.262***	-.211**	.251***	-.127	-.051	-.100
ME72209	78.82***	-.058	.018	-.064	-.334***	-.124	-.555***	.138*

Note: $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 6.:
Estimates of the β-parameters as outcome of the DIF explanation with response times in TIMSS.

Item	England	France	Hong Kong	Norway	Qatar	South Korea	Taiwan	Response Time
ME52058B	.050	-.193*	-.174*	.249***	.100	-.099	-.287***	-.002***
ME52229	-.291***	-.320***	-.056	.-.210**	.277***	.440***	.290***	-.004***
ME52063	.006	-.239**	.453***	-.385***	.139	.387***	.638***	-.007***
ME52072	.298***	-.095	.296**	.086	.141*	.063	-.187*	-.009***
ME52146A	.248***	.506***	356***	.619***	.472***	.604***	.876***	.002***
ME52092	-.369***	-.268**	-.326***	-.196*	.112	.168*	-.213**	.004***
ME52082	-.282***	-.291***	-.139	-.150*	-.389***	.183*	.075	.004***
ME52161	.206*	.194*	-.241**	.666***	-.158*	.178	.088	-.006***
ME52418A	.123	-.087	.037	.094	.155*	-.173**	-.094	.003***
ME52418B	.178**	.127	.199**	.054	-.066	-.092	-.158*	.002***
ME72007	-.692***	-.098	.170**	-.146*	-.132	-.017	.269***	.006***
ME72017	-.237**	-.378***	-.693***	-.463***	-.477***	.136*	.027	.004***
ME72190	.220**	.310***	.051	.133*	-.164*	-.010	.126	-.004***
ME72098	-.468***	-.424***	.163*	-.296***	-.290***	-.060	-.067	.004***
ME72121	.309***	-.250***	.115	.313***	.155*	.136	.047	-.003***
ME72180	.078	.176*	.010	.021	-.094	.051	-.272***	-.004***
ME72227	-.020	.122	-.050	.362***	-.193**	-.104	-.062	.003***
ME72170	.251***	.393***	-.195**	.341***	-.034	-.054	-.095	-.003***
ME72209	-.011	.044	-.226**	-.165	-.052	-.514***	.063	.004***

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

5. Discussion

This study focused on the use of process data in the explanation of DIF with the goal to improve distinction between meaningful DIF and DIF as a nuisance factor. A proposed model, based on the MIMIC model was tested on simulated data and afterwards tested with a real data example. The simulations show that there is high power for the MIMIC model to detect DIF in cases with stronger DIF and higher sample size. In cases with either weaker DIF or a smaller sample size, the power of the model dropped and was sometimes rather low. For the explanation of DIF with the simulated response times, the power levels follow a similar trend. All relative explanation rates are high, indicating that DIF is explained in most cases using this model, considering it is detected first. For the absolute explanation rates, the model only shows high power for cases that have both a high sample size and stronger DIF. These results mean that the model functions correctly in specific cases and can indeed be used to explain DIF with process data.

When the model was used on the empirical data, the detection of DIF was high, with indication of DIF in all test-items. The process data failed to explain most of the DIF. In addition, the cases where DIF disappeared showed no coherent pattern, and cases where more DIF was introduced by adding process data were present as well. The finding that response times do not explain DIF in the TIMSS data can have multiple causes. Firstly, the response times deviated quite from how they were modeled in the simulations. No relation was found between response time and correct response, which was modeled and assumed as .4 in the simulations. A possible explanation for this relation being different, might be that TIMSS is a low-stake test for the test takers in several countries (Fishbein et al., 2021). When a test is a low-stake test, the motivation but also the response times might differ from what is expected in a high-stake test (Finn, 2015; Wise et al., 2009). The simulations might follow the case of a high-stake test, and the empirical data might be the case of a low-stake test. This might possibly explain the difference in findings.

A second reason for the discrepancy in simulation results and empirical results can be that the response times within TIMSS showed some unexpected features. For example, some participant had unrealistically high response times, and for some questions combined times are provided, which cannot be reduced to the times of the individually scored items. These flaws might indicate that the measurement of the response times in TIMSS is not without mistakes, and potential explanation of DIF by the response times could be masked by this. Lastly, within TIMSS the DIF might correctly not be explained by the response times. However, to get more insight in whether this would be the case, the items should be investigated on the substantive content of the questions.

Based on the findings of the simulations it can be concluded that the proposed model can be used to explain DIF with person-mean centered process data as the explaining covariate. Although in this study only response times are used as process

data, the model can be adjusted for the use of other types of process data. There is no reason to assume the model to function differently for these other types of process data. Therefore, The answer to the central research question of this study is that the proposed model can be used to explain DIF with process data. Additionally, it can be stated that a new use for process data can be found in the explanation of DIF. However, an important consideration for the use of the proposed model and for explanation of DIF with process data in practice is that substantive theory must be used to decide what process data could be used to explain DIF. Since DIF can be seen as the flaw of measuring a second construct besides the intended one (Ackerman, 1992), a substantive theory is needed to explain why a specific additional construct is captured by the process data of interest. In order to form such theory, it is needed to consider the content of the concerned items.

In addition to a theoretical framework to use process data in explanation of DIF, some other practical issues of the current state of the area of this study should be considered, to place usability of the studied methods in a context. Firstly, the collection of process data in general is not flawless yet. As process data is collected in the current formats, finding the needed process data and to preprocess them in a compatible format with responses can be a difficult and mostly time consuming job. Since this data wrangling to correct formats is also prone to error, reproducibility and utility of using process data in DIF entanglement might be at question in this current state of data collection. Furthermore, modeling of the combination of response data and process data is still a relative young research area. As mentioned earlier, for the modeling of response times with response data quite some models and studies exist (Entink, 2009; Molenaar et al., 2015; van der Linden, 2007). However, for other types of process data studies and theories on models that combine process data with responses are more scarce. Good entanglement of DIF with process data would benefit if general modeling of process data improves. The current proposed model could be adapted so that it reflects more complicated modeling of process data, possibly resulting in a better explanation of DIF. Lastly, the simulations in this study showed that rather high sample sizes are needed for the model to be able to explain the DIF. This would mean that practical use would be mainly in the comparison of bigger groups. Examples here could be continents, grouped countries or genders. For smaller samples such as one booklet per country as in the empirical example, the current model might be of lesser use.

Although the practical limitations, this model and study are only a first step in using process data in the explanation of DIF. No other studies that relate these topics seem to exist so far. The findings of this study provide therefore several points for further research. Firstly, it is interesting to see how the proposed model can be used with other types of process data. Based on more substantive theory, empirical examples with different types of process data could be studied to find more support for the proposed model. In addition, an empirical example that fits the proposed model better in general

could be found to study the model in practice with more nuance. Furthermore, in further studies the model could be extended to include non-uniform DIF as well, as this was already discussed in this study, but not applied because of theoretical choices related to the empirical example. Another point for further research might be the assumptions of the MIMIC model. As stated earlier, the MIMIC model assumes equal variance for the latent trait over groups. Further research might study if these assumptions hold or if the model needs to be adjusted to account for difference in variances. Lastly, more simulation studies on the proposed model can be done to see how the model functions in other scenarios, such as with more groups, other sample sizes or with less test items.

All taken together, distinction between meaningful DIF and DIF as bias is important and process data can theoretically play a role in this disentanglement. However, given the current state of collection and modeling of process data, wide use of process data in the explanation of DIF is not feasible yet. Even with practical improvements around process data, strong theory is still needed for good explanation of DIF. Regardless of these objections, the current study forms a first step in the use of process data in the explanation of DIF, which with further research can contribute to improved measurements.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Avşar, A. Ş., & Emons, W. H. (2021). A cross-cultural comparison of non-cognitive outputs towards science between turkish and dutch students taking into account detected person misfit. *Studies in Educational Evaluation*, *70*, 101053. https://doi.org/10.1016/j.stueduc.2021.101053

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340. https://doi.org/10.1007/s11336-014-9408-y

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In M. R. Lord F. M. & Novick (Ed.), *Statistical theories of mental test scores* (pp. 397–424). Addison-Wesley.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, *22*(1), 1. https://doi.org/10.7275/70yb-dj34

Bulut, O., & Suh, Y. (2017). Detecting multidimensional differential item functioning with the multiple indicators multiple causes model, the item response theory likelihood ratio test, and logistic regression. *Frontiers in Education*, *2*, 51. https://doi.org/10.3389/feduc.2017.00051

Burkhardt, H., & Pead, D. (2003). Computer-based assessment: A platform for better tests. *Whither Assessment*, 133–148.

Choi, Y., Alexeev, N., & Cohen, A. (2014). Dif analysis using a mixture 3pl model with a covariate on the timss 2007 mathematics test. *KAERA Research Forum*, *1*(1), 4–14.

Chun, S. (2014). *Using mimic methods to detect and identify sources of dif among multiple groups*. University of South Florida.

Dell-Ross, T. L. (2021). Investigating the performance of (multiple-factor) multiple-group methods for the detection of differential item functioning.

Entink, R. H. K. (2009). *Statistical models for responses and response times*. Thesis.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, *2*(3-4), 199–215. https://doi.org/10.1080/15305058.2002.9669493

Feskens, R., Fox, J.-P., & Zwitser, R. (2019). Differential item functioning in pisa due to mode effects. *Theoretical and practical advances in computer-based educational measurement* (pp. 231–247). Springer, Cham.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, *2015*(2), 1–17. https://doi.org/10.1002/ets2.12067

Fishbein, B., Foy, P., & Yin, L. (2021). *Timss 2019 user guide for the international database*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/international-database/

Gao, X. (2019). *A comparison of six dif detection methods* (Master's thesis).

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform dif. *Journal of Educational Measurement*, *46*(3), 314–329. https://doi.org/10.1111/j.1745-3984.2009.00083.x

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information–an example from the measurement of adolescent mental health. *BMC Medical Research Methodology*, *19*(1), 1–9. https://doi.org/10.1186/s12874-019-0828-3

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care*, *44*(11), S182–S188. https://doi.org/10.1097/01.mlr.0000245443.86671.c4

Hambleton, R. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of irt area and mantel-haenszel methods. *Applied Measurement in Education*, *2*(4), 313–334. https://doi.org/10.1207/s15324818ame0204_4

Han, Y., Liu, H., & Ji, F. (2021). A sequential response model for analyzing process data on technology-based problem-solving tasks. *Multivariate Behavioral Research*, 1–18. https://doi.org/10.1080/00273171.2021.1932403

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations*, *10*(3), 301–320. https : / / doi . org / Culturaldifferencesinteachingandlearning

International Test Commission & Association of Test Publishers. (in press). *Guidelines for technology-based assessment.*

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485

Kalaycioğlu, D. B., & Berberoğlu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in turkey. *Journal of Psychoeducational Assessment*, *29*(5), 467–478. https://doi.org/10.1177/0734282910391623

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review/Revue Internationale de Statistique*, 237–254. https://doi.org/10.2307/1402373

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. *Basic problems in cross-cultural psychology*, 19–29.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Maris, G., Bechger, T., Koops, J., & Partchev, I. (2018). Dexter: Data management and analysis of tests [computer software manual].

Martin, M. O., von Davier, M., & Mullis, I. V. (2020). Methods and procedures: Timss 2019 technical report. *International Association for the Evaluation of Educational Achievement*.

Mellenbergh, G. J. (1983). Conditional item bias methods. *Human assessment and cultural factors* (pp. 293–302). Springer.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3-4), 177–181. https://doi.org/10.1080/15366367.2015.1105073

Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, *83*(2), 279–297. https://doi.org/10.1007/s11336-017-9602-9

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A generalized linear factor model approach to the hierarchical framework for responses and response times. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 197–219. https://doi.org/10.1111/bmsp.12042

Mullis, I. V., & Martin, M. O. (2017). *Timss 2019 assessment frameworks*. ERIC.

Muthén, B., & Asparouhov, T. (2013). Bsem measurement invariance analysis. *Mplus web notes*, *17*, 1–48.

Özdemir, B. (2015). A comparison of irt-based methods for examining differential item functioning in timss 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, *174*, 2075–2083. https://doi.org/10.1016/j.sbspro.2015.02.004

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three mantel-haenszel procedures. *Applied Measurement in Education*, *14*(3), 235–259. https://doi.org/10.1207/S15324818AME1403_3

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, 2231. https://doi.org/10.3389/fpsyg.2018.02231

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rosseel, Y. (2012). Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta). *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

RStudio Team. (2021). *Rstudio: Integrated development environment for r*. RStudio, PBC. Boston, MA. http://www.rstudio.com/

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170–187. https://doi.org/10.1080/13803611.2013.767621

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, *74*(1), 1–33. https://doi.org/10.1111/bmsp.12203

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with scylla and charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in psychology*, *4*, 770. https://doi.org/10.3389/fpsyg.2013.00770

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

von Davier, M. (2019). Timss 2019 scaling methodology: Item response theory, population models, and linking across modes. *Methods and Procedures: TIMSS*, 11–1.

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58. https://doi.org/10.1037/1082-989X.12.1.58

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Woods, C. M. (2009). Evaluation of mimic-model methods for dif testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. https://doi.org/10.1080/00273170802620121

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. https://doi.org/10.1177/0146621611405984

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—threats and opportunities. *Theoretical and practical advances in computer-based educational measurement* (pp. 3–19). Springer, Cham.

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2021). Accurate assessment via process data. *arXiv preprint arXiv:2103.15034*. https://doi.org/10.48550/arXiv.2103.15034

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (dif). *Ottawa: National Defense Headquarters*, *160*.

Zumbo, B. D. (2007). Three generations of dif analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832