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

Research Report

Thijs Carrière, 5843545

December 19th, 2021

Programme: Methodology and Statistics for the Behavioural, Biomedical and Social Sciences

Supervisors: Remco Feskens, *Cito & University of Twente*
Dylan Molenaar, *University of Amsterdam*

Wordcount: 2497

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, of which the use and administration of Technology Enhanced Items (TEIs) are one (International Test Commission & Association of Test Publishers, in press). 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). Therefore, TEIs and digital assessment give access to a new type of data in addition 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, which holds didactically important information (Wools et al., 2019) and could allow for improved measurements of tests (Molenaar, 2015).

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 a equal level of a latent trait. When for a test-item groups have unequal probabilities on a correct response, despite an equal score on the latent trait intended to be measured, the test-item is subjected to Differential Item Functioning (DIF; Hambleton and Rogers, 1989). DIF can indicate a meaningful difference between groups when it can be explained (Ercikan, 2002; Kalaycioğlu & Berberoğlu, 2011). This can for example be the case when one group is more proficient on a subtopic of a test than another group. In that case, the probability of a respondent answering correct might be higher when the overall proficiency scores are equal. However, when DIF cannot be explained, it is a nuisance factor. Such DIF items would be biased and cannot be used to compare groups. These items should either be deleted from the test or be changed so the bias disappears (Hagquist, 2019). In ILSAs, the groups compared on average proficiency level are often countries and important policies are based on these comparisons. Biased items must therefore be identified and dealt with to ensure valid comparisons.

Current models find occasional use for process data. A popular type of process data are response times (Molenaar et al., 2015). Response times are assumed to contain information about the latent trait measured that is not covered by the response itself. Several models have been developed to include both responses and response times in one model (Entink, 2009; Molenaar et al., 2015; van der Linden, 2007). However, these models are not aiming to detect or explain DIF. A new use for response data might be found in the explanation of DIF.

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.

The current study will address the link between process data and DIF, which to the best of my knowledge has not been studied yet. This link will be studied by two research questions applied on ILSA data. Firstly, it will be investigated *to what extend Technology Enhanced Items are subjected to Differential Item Functioning.* After that, it will be studied *how process data obtained by Technology Enhanced Items can be used in the detection and explanation of Differential Item Functioning.* The first research question is a prerequisite for the second question, since DIF must be detected for it to be explained. DIF has already been detected in ILSA items before (Avşar & Emons, 2021; Choi et al., 2014; Feskens et al., 2019; Özdemir, 2015). Therefore, it is expected that items with DIF can be identified. 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 found DIF to a certain extent.

Models for DIF and Process data

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 (y_i) are loading on a latent variable (θ). A grouping variable (z) is then regressed on the latent variable. Lastly, the item under investigation for DIF is regressed on the grouping variable. This model can be written as:

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

Here, y_i^* is the latent underlying variable for the response on item y_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 where different groups are modeled by the use of dummy variables, β_i is the regression coefficient from item i on grouping variable z and ϵ_i is the random error for item i. The investigated item is subjected to DIF when $\beta_i \neq 0$.

As explained in Wirth and Edwards (2007), the latent underlying variable, y_i^*, is a normal continuous variable and its response category (c) in item y_i is determined based on threshold τ, where

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

This 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 as explanation for DIF (Chun, 2014; Woods, 2009). This latter addition is an advantage for this model over other DIF detection models, since the second research question of this study asks for additional covariates to be added to the model where DIF is investigated.

Another addition to the MIMIC model is its ability to detect both uniform and non-uniform DIF, where the original MIMIC model can only detect uniform DIF (Woods & Grimm, 2011). A test item has uniform DIF when the difference in the probability of giving a correct answer is constant over the latent trait. On the other hand, an item is subjective to non-uniform DIF when there is an interaction between group membership and response probability. For example, when one group has a higher chance of answering an item correctly when having a low score on the latent ability, while the other group has a higher chance 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:

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

Here, ω_i is the regression coefficient from the interaction term between the grouping variable and the latent trait (θz) 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.

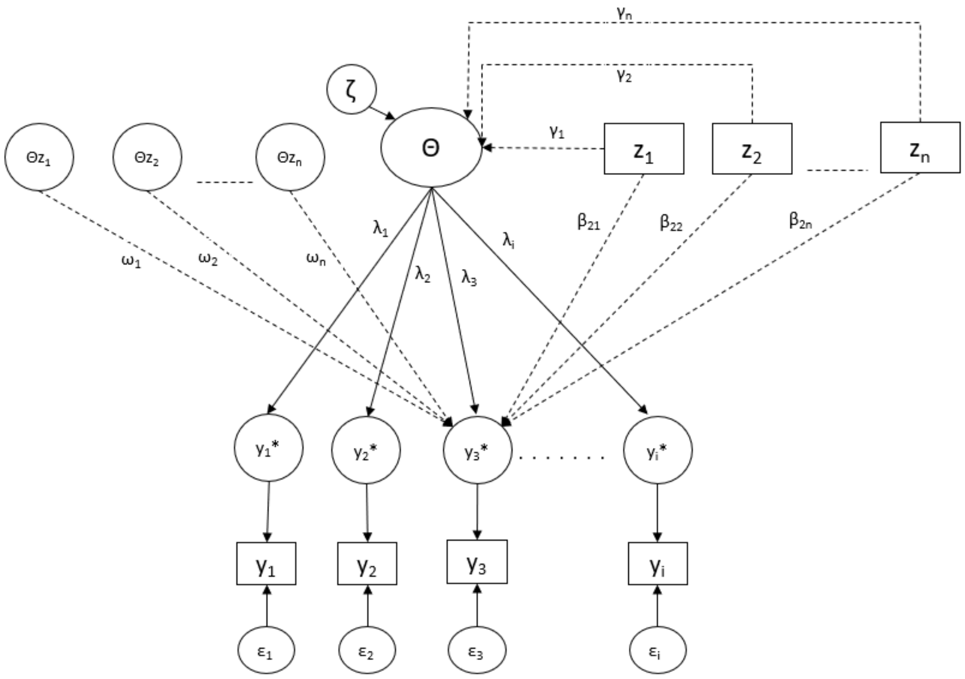

Figure 1: *Mimic model to investigate uniform and non-uniform DIF. In this model, DIF is checked for item 3. Dashed lines indicate regression.*

Models on response and response time

As mentioned earlier, numerous 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 give 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 response times. In this framework, for test taker j taking a test both a vector of responses ($\mathbf{U}_j$) and a vector of 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 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 can be 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. This frequentist framework has several advantages (Molenaar et al., 2015). It allows for common used models to be used in this framework, and multilevel components can be added easily, which educational data can benefit from. 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 for this study.

A Model for DIF and Process data

Process data can be modeled in different ways to explain DIF and multiple ways of modeling are used in this study. In the first proposed model the process data related to the specific DIF item is added as a covariate in the MIMIC model. This covariate trait is then regressed on the grouping variable and the item flagged with DIF is regressed on this covariate. To see whether the found DIF can be (partly) explained by the covariate, it can be checked if the β-parameters and ω-parameters from the MIMIC model becomes non-significant. If this is the case, the DIF is (partly) explained. This first proposed model is shown in Figure 2.

The second way process data are modeled to explain DIF, is by loading all of one type of process data on another latent variable. This second latent variable is comparable to the latent factor, tau, in the model by Molenaar et al. (2015) and van der Linden (2007) discussed above. This latent variable will then be regressed on the grouping variable, and the item with DIF will be regressed on this latent variable. The latent variable therefore takes the place of the covariate in the model in Figure 2. DIF is again explained when the β-parameters and ω-parameters become non-significant.

Methods

TIMSS Data

The data used in this study are data collected within the *Trends in International Mathematics and Science Study* (TIMSS; Fishbein et al., 2021). 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 mathematics and science proficiency, 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.

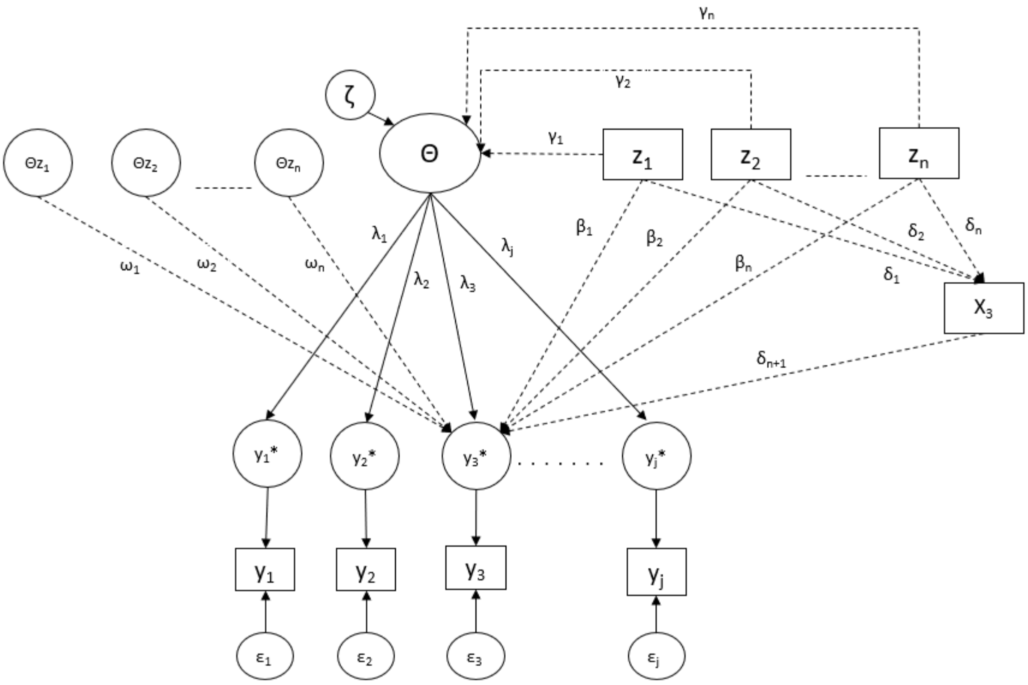

Figure 2: *Process data of item 3 as a covariate (x_3) added to the MIMIC model, explaining the DIF in item 3. Dashed lines indicate regression.*

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 subjective to DIF. The countries used are England, France, Japan, 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.

Process data

The process data used in the analyses to explain DIF are the process data gathered by eTIMSS. Process data are available for every student-item interaction. Two different sources of process data are used in the analyses. Firstly, response times are used because models that can include both responses and response times have been developed before (Molenaar et al., 2015; van der Linden, 2007). Response times are a commonly collected and used process data and are seen as informative for the measured responses (Molenaar, 2015). In eTIMSS response times are measured in seconds. Secondly, number of screen visits of an item is used. These process data can be informative on which items are found difficult. This feature is measured on a ratio scale.

At the moment of writing, it is not yet clear whether additional process data will be available. If this is to be the case, a third or even fourth type of process data will be used in the analyses.

Pre-processing of data

Before the data were in a correct 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 score was given. Process data were then converted to a compatible format to the response data. 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). All steps of the pre-processing will be available on GitHub.

Detection of DIF

The first step of the analyses and the answer to the first research question is the detection of DIF items, since the second research question can only be answered on test items subjected to DIF. 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 if the item has uniform or non-uniform DIF by investigating what β-parameters and ω-parameters differ significantly from 0. An output table could look like Table 1.

Modeling process data

To answer the second research question, each of the selected process data is modeled for each of the test items found to be subjective to DIF. This modeling of the process data is tried in different ways, following the models as proposed in the subsection 'Models for DIF and Process data'. For each of these models, it is investigated if β-parameters and ω-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.

Table 1.

Possible outcomes of DIF detection.

Item	χ^2-difference *(df)*	β_1	β_2	β_3	β_4	ω_1	ω_2	ω_3	ω_4
1	3.242 *(4)*	.029	-.182	.027	-.051	-.004	.087	.102	-.006
2	7.123 *(4)*	-.283	-.243	-.234	-.269	.002	-.011	.008	.004
3	18.938*** *(4)*	.267*	.378***	.229*	-.068	.005	.002	-.010	.022**
...	...	...	...	...	...	...	...	...	...
6	11.527* *(4)*	-.087	.243*	-.289**	.116	.048	-.033*	-.002	.015

Note: Synthetic data was used for this table. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Software

All of the analyses will be performed in RStudio (RStudio Team, 2020), where all structural equation models are run with the lavaan package (Rosseel, 2012).

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.

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.

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*(2), 317–340.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research, and Evaluation*, *22*(1), 1.

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.

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.

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.

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.

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.

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.

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of intercultural relations*, *10*(3), 301–320.

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.

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.

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. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127–143.

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

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.

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

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.

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.

RStudio Team. (2020). *Rstudio: Integrated development environment for r.* RStudio, PBC. Boston, MA. http://www.rstudio.com/

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

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological methods*, *12*(1), 58.

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.

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.

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.

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.