Available online at ijci.wcci-international.org

International Journal of Curriculum and Instruction 15(3) (2023) 2079–2093

IJCI
International Journal of Curriculum and Instruction

A Machine Learning Approach to Assess Differential Item Functioning of PISA 2018 ICT Engagement Questionnaire

Gözde Sırgancı [a]

[a] *Southern Methodist University, Simmons School of Education and Human Development, Dallas and 75205, USA*

Abstract

This study aimed to investigate the different item functioning (DIF) of the Program for International Student Assessment (PISA) 2018 information, communication, and technology (ICT) questionnaire items based on country, gender, Economic, social, and cultural status (ESCS) variables. The sample included 29,277 15-year-old students from Eastern Europe and Central Asia (EECA) countries. The study employed the generalized partial credit model with lasso penalization, a machine learning approach, to evaluate DIF. The findings showed that two out of 21 items exhibited DIF based on country, gender, and ESCS overall. There were seven out of 21 items that showed DIF in favor of males and six out of 21 items that showed DIF in favor of females. According to ESCS, only three out of 21 items displayed DIF. All items exhibited DIF based on countries. According to the GPCMlasso coefficient, when reference group Bulgaria and focus group Georgia, Croatia, Kazakhstan, and Turkey are compared, 28%, 71%, 33%, and 33% of all items hold DIF, respectively. In pairwise comparisons, the most DIF-prone items were found between Bulgaria and Croatia, while the fewest were between Bulgaria and Georgia.

Keywords: Differential item functioning, regularization, GPCMlasso, machine learning, PISA 2018

1. Introduction

Measurement is fundamental to science, especially in education psychology, and related disciplines. It is crucial that scores from measures, like achievement tests or symptom inventories, are comparable across individuals. If scores inaccurately represent the underlying trait for certain groups, then comparisons become invalid. For example, if the scores overestimate the latent trait for some people (i.e., females) and underestimate it for others (i.e., Americans), then observed differences will not reflect true differences. Using such distorted scores in further analysis can lead to misleading results, potentially hiding some effects, overstating others, or producing erroneous conclusions due to measurement errors (Bauer et al., 2020; Millsap, 2011). Measurement bias poses a significant challenge in high-stakes evaluations, as it can distort test results, affecting not just the skills under scrutiny but also being swayed by unrelated factors. Administered by the Organization for Economic Co-operation and Development (OECD), the Program for International Student Assessment (PISA) is a triennial global assessment focusing on the literacy of 15-year-old students in math, science, and reading. It has gained prominence as a key metric for comparing educational systems worldwide and shaping educational policies (OECD, 2019a). In its 2015 edition, PISA incorporated a questionnaire on information, communication, and technology (ICT) engagement, which was informed by Self-Determination Theory (SDT, Deci & Ryan, 1985). According to SDT, both intrinsic and extrinsic motivators play roles in human behavior, centering around three fundamental needs: skill, freedom of action, and social connections. Therefore, the ICT questionnaire examines four areas: 1) enthusiasm for ICT as a measure of intrinsic motivation; 2) self-perceived ICT skills; 3) one's sense of independence when using ICT; and 4) the extent to which social media is used for interaction (Zylka et al., 2015). The same questionnaire was used in the 2018 PISA. Given that the ICT engagement questionnaire was first introduced in 2015, it remains a novel tool for assessment. Ensuring it is free from measurement bias is vital, particularly when its purpose is to facilitate cross-group comparisons, a central aim of PISA. This phenomenon is more precisely referred to as differential item functioning (DIF; Millsap, 2011). The majority of measurement bias and DIF research has been historically viewed from a unidimensional, group-specific perspective, such as gender (i.e., males and females), and ethnicity (i.e., white and Hispanic). A possible benefit of this perspective is that it is able to provide a reasonable understanding of bias (Belzak, 2022). In the literature, there are several statistical methods to detect DIF in dichotomous (Holland & Thayer, 1988; Swaminathan & Rogers, 1990; Thissen et al., 1993) and polytomous items (Chang et al., 1996; Choi et al., 2011) across groups of individuals. Meanwhile, most of these methods are limited to a single dichotomous covariate. On the other hand, the unidimensional, group-specific perspective also has several disadvantages. A significant issue with this perspective is its oversight of intricate measurement biases, especially in high-stakes exams, resulting in adverse outcomes. Such biases can arise from various elements, including the examinees' individual and group characteristics, influencing the test outcomes. The second problem is the limitation of statistical analyses such as Mantel-Haenszel, SIBTEST, and IRT-LR-DIF to conditional, group-based analyses. This means they cannot be used to assess continuous background variables (e.g., age, time spent on a test question, etc.) in DIF. It is also important to note that multiple background variables may influence test responses and scores in a variety of linear, nonlinear, and even nonparametric ways, many of which cannot be detected with these methods (Belzak, 2023; Schauberger & Mair, 2020). In order to overcome these drawbacks, a wide variety of regularized or bias-correcting methods have been proposed, known as machine learning (ML) approaches. Recent developments in ML techniques have also led to the rapid development of DIF evaluation. Among these techniques are regularization (Bauer et al., 2020; Belzak & Bauer, 2020; Liang & Jacobucci, 2020; Tutz & Schauberger, 2015), recursive partitioning/decision trees (Strobl et al., 2015; Tutz & Berger, 2016), and boosting (Schauberger & Tutz, 2016). ML methods allow DIF to be evaluated across multiple background characteristics in linear, nonlinear, and nonparametric formats. Therefore, since ML has begun to provide exploratory tools for understanding, interpreting, and justifying complex input, it seems reasonable to assume that ML can also assist in identifying and explaining complex measurement biases (Belzak, 2023). Recent psychometric studies have shown that the most effective ML method for identifying complex DIF effects is to use regularization within a general linear model (GLM) or IRT framework (Bauer et al., 2020; Belzak & Bauer, 2020; Tutz & Schauberger, 2015). Besides, regularization techniques outperform traditional DIF detection methods, especially when the sample size for one or both interest groups is relatively small (Belzak & Bauer, 2020).

In conclusion, DIF is an important validity issue in multi-cultural/national assessments (Xu &Tracey, 2017). Also the DIF of the newly introduced ICT engagement questionnaire in PISA has not been fully examined. There are limited studies that have established the equivalence of the ICT engagement questionnaire (Ma & Qin, 2021; Meng et al., 2019; Odell et al., 2021). On the other hand, no study detects DIF using machine learning algorithms in PISA studies. It is therefore necessary to further validate the questionnaire using different techniques before applying it to make robust national and international comparisons. Also, machine learning approaches enable the test of the effect of several potential variables, which may lead to DIF. Thus, the present study aimed to detect the DIF of the ICT engagement questionnaire items across different countries using the machine learning approaches on PISA 2018 data. Economic, social, and cultural status (ESCS) and gender variables were simultaneously included as potential covariates that may contribute to the development of DIF.

The following research question is necessary in order to accomplish the purpose of this study: "Do the items in the PISA 2018 ICT questionnaire demonstrate differential item functioning according to country, gender, and economic, social, and cultural status index variables?"

2. Method

This section contains information about the research model, participant characteristics, data collection tool, and data analysis process.

2.1. Research Model

Descriptive research focuses on presenting a situation in its authentic state, as highlighted by Karasar (2017). The aim of the study is to investigate the differential item functioning of PISA 2018 ICT questionnaire. Thus, this study adopts a descriptive research approach to depict the current situation.

2.2. Participant characteristics

Analyzes are conducted with participants from Eastern Europe and Central Asia (EECA) countries. The PISA-EECA countries have continuously increased their involvement, and ten countries participated in 2018: Baku (Azerbaijan), Belarus, Bulgaria, Croatia, Georgia, Kazakhstan, Moldova, Romania, Turkey, and Ukraine. However, five out of 10 countries participated in the ICT questionnaire (OECD, 2021). A summary of EECA countries' information is provided in Table 1.

Table 1. Frequencies and percentages related to the study group

Code	**Country**	**Gender**	**f**	**%**	**N**	**ESCS (mean±sd)**
BRG	Bulgaria	Female	1276	49.6	2574	-0.11±0.9
		Male	1298	50.4		-0.10±1.0
GEO	Georgia	Female	1336	51.1	2613	-0.28±0.9
		Male	1277	48.9		-0.29±0.9
HRV	Croatia	Female	2508	51.9	4835	-0.24±0.8
		Male	2327	48.1		-0.19±0.8
KAZ	Kazakhstan	Female	6724	48.8	13775	-0.26±0.8
		Male	7051	51.2		-0.29±0.8
TUR	Turkey	Female	2736	49.9	5480	-1.13±1.2
		Male	2744	50.1		-1.08±1.2

2.3. Data Collection Tool

There are four sub-scales in the PISA 2018 ICT engagement questionnaire: interest in ICT contains six items, perceived competence of ICT contains five items, perceived autonomy of ICT contains five items, and use of social media contains five items (OECD, 2019a). ICT is a four-point Likert scale. Response categories are from strongly disagree to strongly agree. Higher values indicate better ICT engagement.

Table 2. Standard deviation, and reliabilities of the ICT engagement questionnaire

Scale	**Interest in ICT**		**Perceived competence of ICT**		**Perceived autonomy of ICT**		**Social media contains**	
Country	(mean±sd)	α	(mean±sd)	α	(mean±sd)	α	(mean±sd)	α
Bulgaria	4.33±1.1	.90	3.54±0.9	.90	3.67±0.9	.90	3.56±0.9	.89
Georgia	4.48±1.1	.90	3.80±0.9	.90	3.65±0.9	.90	3.85±1.0	.91
Croatia	3.34±0.8	.82	3.04±0.8	.85	3.11±0.8	.85	3.68±0.9	.91
Kazakhstan	3.57±0.9	.85	3.40±0.8	.93	3.26±0.8	.90	3.31±0.8	.90
Turkey	4.29±1.1	.88	3.57±0.9	.89	3.68±0.9	.90	3.77±0.9	.89

2.4. Data Analysis

DIF analysis was performed according to the GPCMlasso regularization method.

2.4.1. DIF model for generalized partial credit models

DIF analysis was performed according to the GPCMlasso regularization method (GPCMlasso; Schauberger & Mair, 2020). The GPCMlasso model for assessing DIF among EECA countries (G1), adjusted by students' gender (G2) and ESCS (G3), shows equation 1.

$$\log\left(\frac{P(Y_{pi}=r)}{P(Y_{pi}=r-1)}\right)=\beta_i=\left[\theta_p+x_p^T\alpha-\delta_{ir}-\left(\gamma_{i1}\times G_1+\gamma_{i2}\times G_2+\gamma_{i3}\times G_3\right)\right] \qquad (1)$$

In this model γ_{i1}, γ_{i2} and γ_{i3}, which represents DIF parameters, are the effects of grouping variables (G_1, G_2, G_3) on item *i,* respectively. If these parameters are not equal to zero after applying lasso penalization, they are considered uniform DIF. Furthermore, the GPCMlasso can control multicollinearity associated with highly correlated variables (i.e., EECA countries, gender, and ESCS) in the model, as well as covariate adjustment. This GPCMlasso model calculates the parameter estimate by solving the following lasso penalized log-likelihood function illustrated in equation 2 (Schauberger & Mair, 2020).

$$\ell_p(\theta,\alpha,\delta,\beta,\gamma)=\ell(\theta,\alpha,\delta,\beta,\gamma)-\lambda\sum_{i=1}^{l}\sum_{j=1}^{m}w_{ij}\left|\gamma_{ij}\right| \qquad (2)$$

where $\ell(\theta,\alpha,\delta,\beta,\gamma)$ is the regular version of the log-likelihood function, $\lambda\sum_{i=1}^{l}\sum_{j=1}^{m}w_{ij}\left|\gamma_{ij}\right|$ represents the lasso penalty term and $\lambda\geq 0$ is the tuning parameter that controls the degree of penalization applied to the vector of regression coefficients γ_{ij}. In this model, x_p represents an m dimension covariate vector of person p (G_1, G_2, G_3), and θ represents the underlying latent construct. Besides, DIF parameters , γ_{i1}, γ_{i2} and γ_{i3}, item step parameter δ_{ij}, main effect α and, item discrimination β_i parameters are estimated for item *i*. It is important in penalized likelihood approaches to determine the optimal tuning parameter λ. To determine the optimal value for λ, one can utilize either model selection criteria such as AIC or BIC or employ cross-validation (CV) techniques. In DIF assessment, BIC and CV have technical and theoretical differences. CV focuses on optimal model prediction, while BIC is more consistent regarding variable selection. CV can be time-intensive, as it necessitates repeated iterations with multiple training and testing datasets. While both criteria were employed in this study, the interpretations primarily leaned on BIC, given that optimal variable selection is more crucial than prediction in DIF assessment. The BIC is defined for the GPCMlasso model in equation 3 (Schauberger & Mair, 2020).

$$BIC(\lambda) = -2L\lambda(.) + df(\lambda)\log(n) \tag{3}$$

where $L\lambda(.)$ denotes the likelihood for the parameters estimated with tuning parameter λ and $df(\lambda)$ denotes the total number of parameters estimated (uniquely) unequal to zero. Using the GPCMlasso package, BIC values were calculated for an automatically generated sequence of λ values, which were then arranged from largest to smallest. Ultimately, the package chose the λ value associated with the lowest BIC. (Schauberger & Mair, 2020; Jafari et al., 2022). The GPCMlasso package in R software (R Core Team, 2018) was used to assess DIF.

3. Results

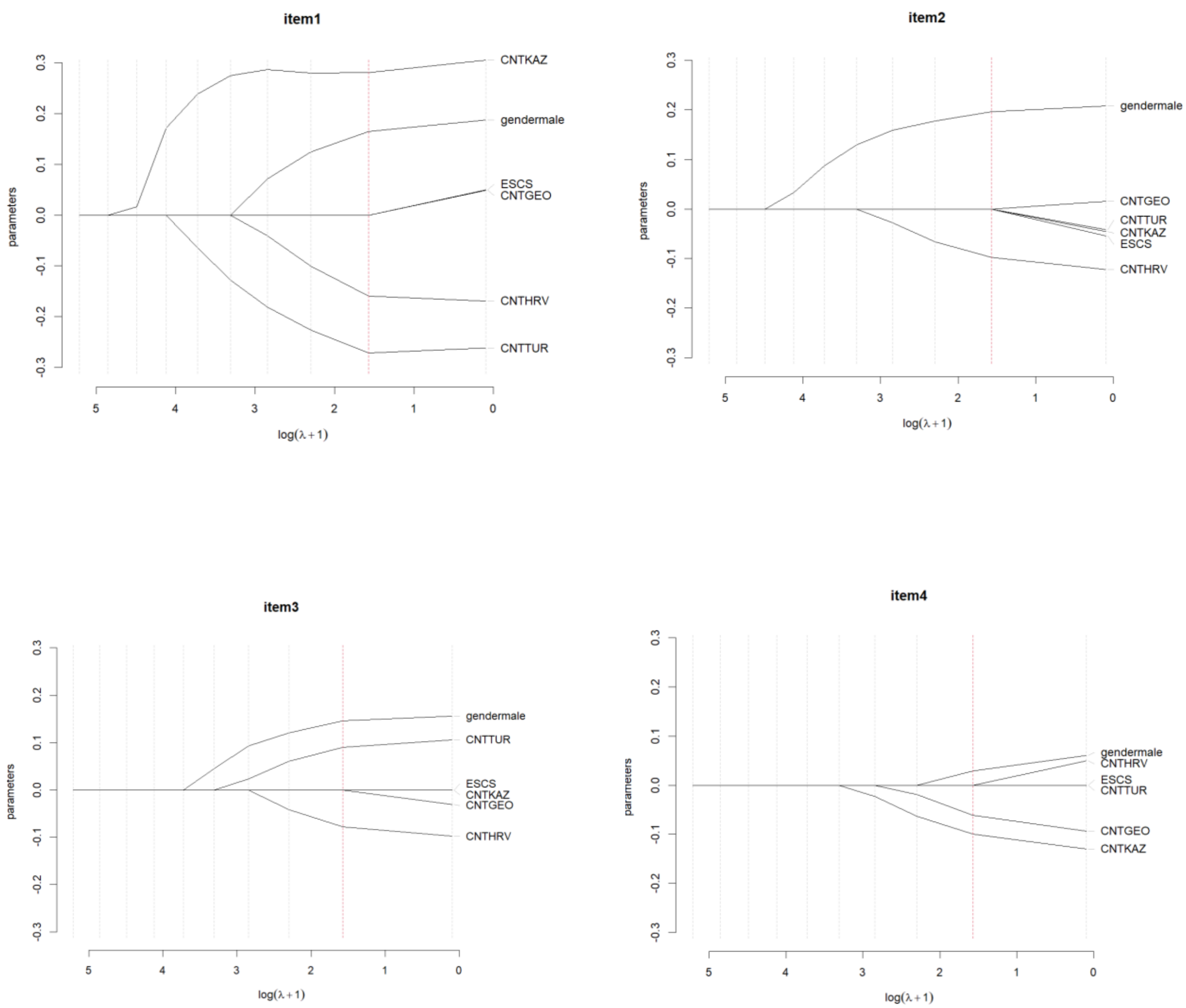
Table 3 represents the results of GPCMlasso model for detecting DIF for country, gender, and ESCS variables based on BIC criteria. The variable country is encoded by dummy coding with Bulgaria as the reference category. Therefore, all the results related to the country variable in Table 3 reflect the results of the pairwise comparison of each country with Bulgaria. Also, female students are the reference group for gender. Uniform DIF occurs whenever lasso coefficients are non-zero for a grouping variable. In the context of lasso regression, the sign of the coefficient represents the direction of the relationship between the predictor and the outcome variable. A positive coefficient indicates that the item is biased in favor of the focal group, while a negative coefficient indicates that the item is biased against the focal group. It is important to remember that the coefficient's magnitude also provides information about the strength or size of the DIF effect. The bigger the absolute value of the coefficient, the stronger the DIF effect is. Thus, both the sign and magnitude of the lasso coefficient are essential for understanding and interpreting DIF in the context of the examined items and groups. The findings in Table 3 are presented in detail only for the items in the dimension of interest in ICT in order to avoid repetition. It is possible to make similar comments about other dimensions as well.

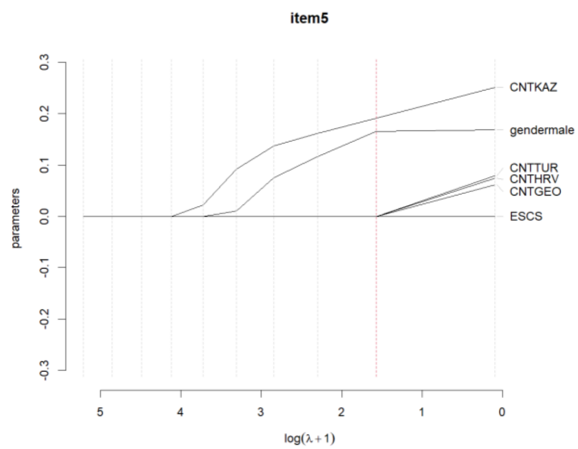
Table 3. DIF Results for Country, Gender, and ESCS Variables Based on GPCMlasso Coefficients

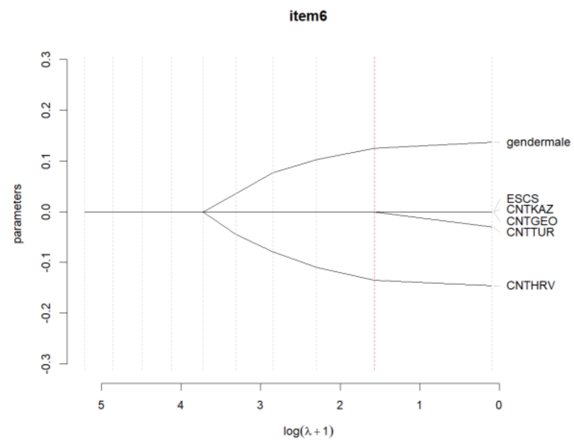
	Code	CNT_GEO	CNT_HRV	CNT_KAZ	CNT_TUR	ESCS	gendermale
Interest in ICT	13Q01 *(item1)*	0	-0.159	0.28	-0.271	0	0.164
	13Q04 *(item2)*	0	-0.098	0	0	0	0.196
	13Q05 *(item3)*	0	-0.078	0	0.09	0	0.146
	13Q11 *(item4)*	-0.061	0	-0.1	0	0	0.029
	13Q12 *(item5)*	0	0	0.191	0	0	0.165
	13Q13 *(item6)*	0	-0.135	0	0	0	0.125
Perceived competence of ICT	14Q03 *(item7)*	0.132	-0.087	-0.069	0.074	0	0
	14Q04 *(item8)*	0.013	0	0	0	-0.035	0
	14Q06 *(item9)*	0.052	-0.151	0	0	-0.056	0.066
	14Q08 *(item10)*	0	0	0	0	0	0
	14Q09 *(item11)*	0	0	0	0	0	0
Perceived autonomy of ICT	15Q02 *(item12)*	-0.045	0.082	-0.066	0	0	-0.113
	15Q03 *(item13)*	0	0.1	-0.017	0	0	-0.054
	15Q05 *(item14)*	0	-0.041	0	-0.011	-0.027	0
	15Q07 *(item15)*	-0.02	-0.067	0	0	0	-0.072
	15Q09 *(item16)*	0	-0.087	0	0.131	0	0
Social media contains	16Q01 *(item17)*	0	0.058	0	-0.037	0	0
	16Q02 *(item18)*	0	0.113	0	0	0	0
	16Q04 *(item19)*	0	0.112	0.093	0	0	-0.298
	16Q05 *(item20)*	0	0.105	0	0	0	-0.078
	16Q07 *(item21)*	0	0.092	0	0.06	0	-0.05

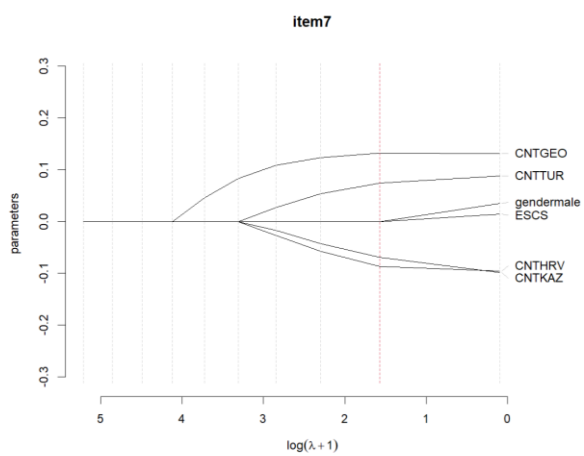
All items in *Interest in ICT* subscale indicate DIF based on a country variable. One item (item 4) between Bulgaria and Georgia and four items between Bulgaria and Croatia (item 1, item 2, item 3, item 6) show DIF in favor of Bulgaria. Between Bulgaria and Kazakhstan, item 1,5 includes DIF in favor of Kazakhstan and item 4 includes DIF in favor of Bulgaria. Between Bulgaria and Turkey, item 1 includes Bulgaria, and item 3 includes DIF in favor of Turkey. On the other hand, ESCS is not a significant predictor of any item in this subscale. In other words, no item in this subscale shows DIF, according to ESCS. Based on gender, all items indicate DIF in favor of males. In the *Perceived competence of ICT* subscale, item 7, item 8, and item 9 have DIF among counties. Also, item 8 and item 9 show DIF according to ESCS. Item 9 has DIF based on gender. Item 10 and item 11 do not show DIF for any variables. All items in the *Perceived autonomy of ICT* subscale display DIF among countries. Only item 14 has a DIF, according to ESCS. Item 12, item 13 and item 14 show DIF against males. In *Social media contains* subscale, no item shows DIF between Bulgaria and Georgia. On the other hand, between Bulgaria and Croatia, all items indicate DIF in favor of Croatia. Between Bulgaria and Kazakhstan, only item 19 has DIF in favor of Kazakhstan. Between Bulgaria and Turkey, item 17 favors Bulgaria, and item 21 favors Turkey. In this subscale, no item indicates DIF according to ESCS. Meanwhile, items 19, 20, and 21 display differential item functioning against males.

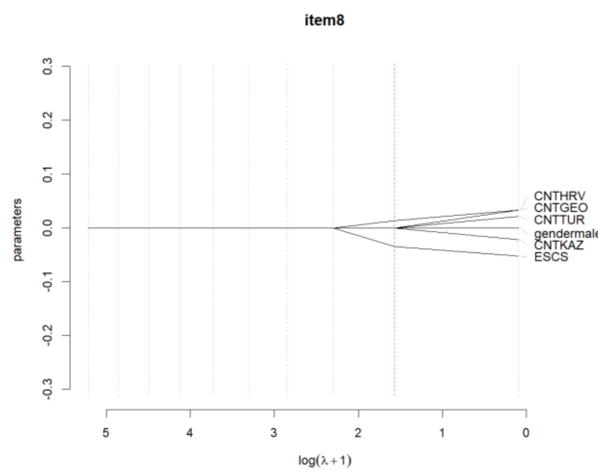
Figure 1 illustrates the coefficient paths for all grouping variables based on the tuning parameter λ, for all items. There were three DIF parameters associated with each variable in this study. The paths are plotted separately for each item.

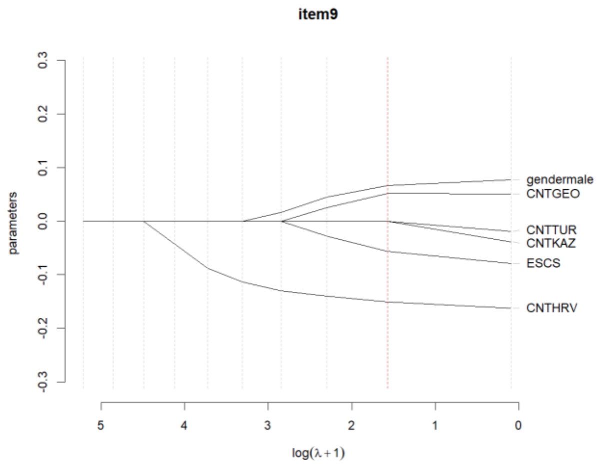


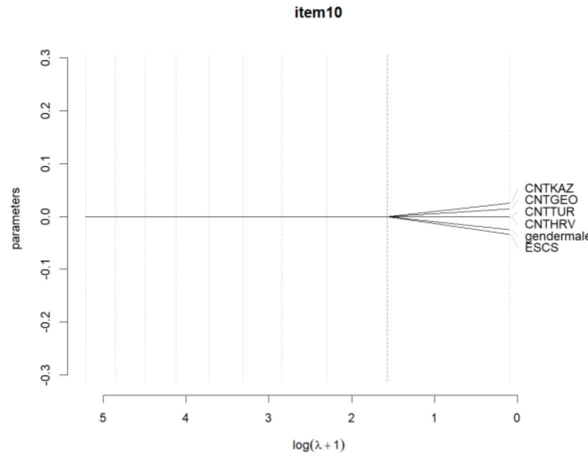

item5
parameters
log(λ+1)
CNTKAZ
gendermale
CNTTUR
CNTHRV
CNTGEO
ESCS

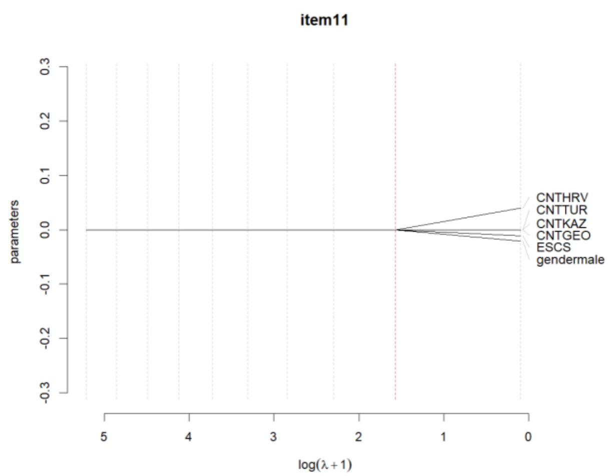

item6
parameters
log(λ+1)
gendermale
ESCS
CNTKAZ
CNTGEO
CNTTUR
CNTHRV

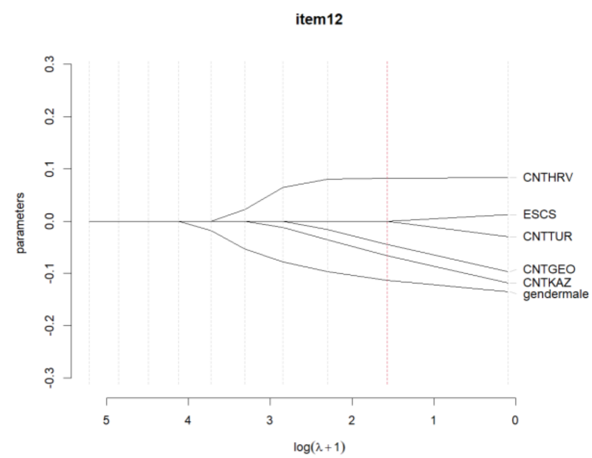

item7
parameters
log(λ+1)
CNTGEO
CNTTUR
gendermale
ESCS
CNTHRV
CNTKAZ

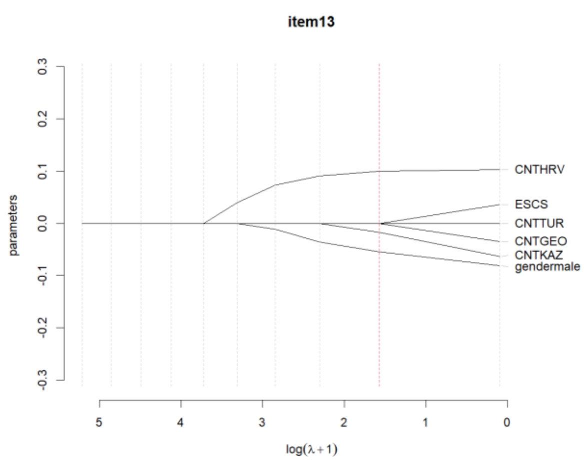

item8
parameters
log(λ+1)
CNTHRV
CNTGEO
CNTTUR
gendermale
CNTKAZ
ESCS

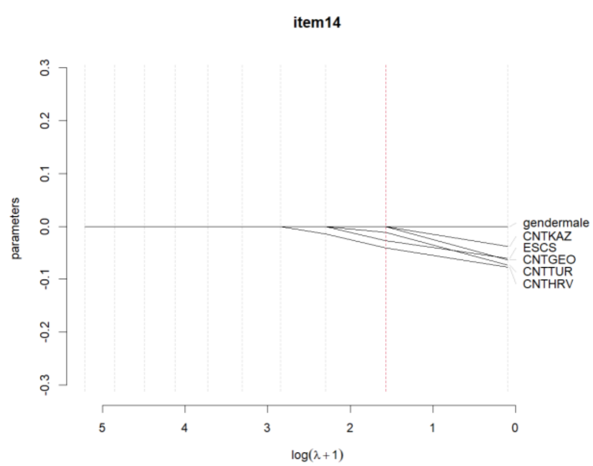

item9
parameters
log(λ+1)
gendermale
CNTGEO
CNTTUR
CNTKAZ
ESCS
CNTHRV

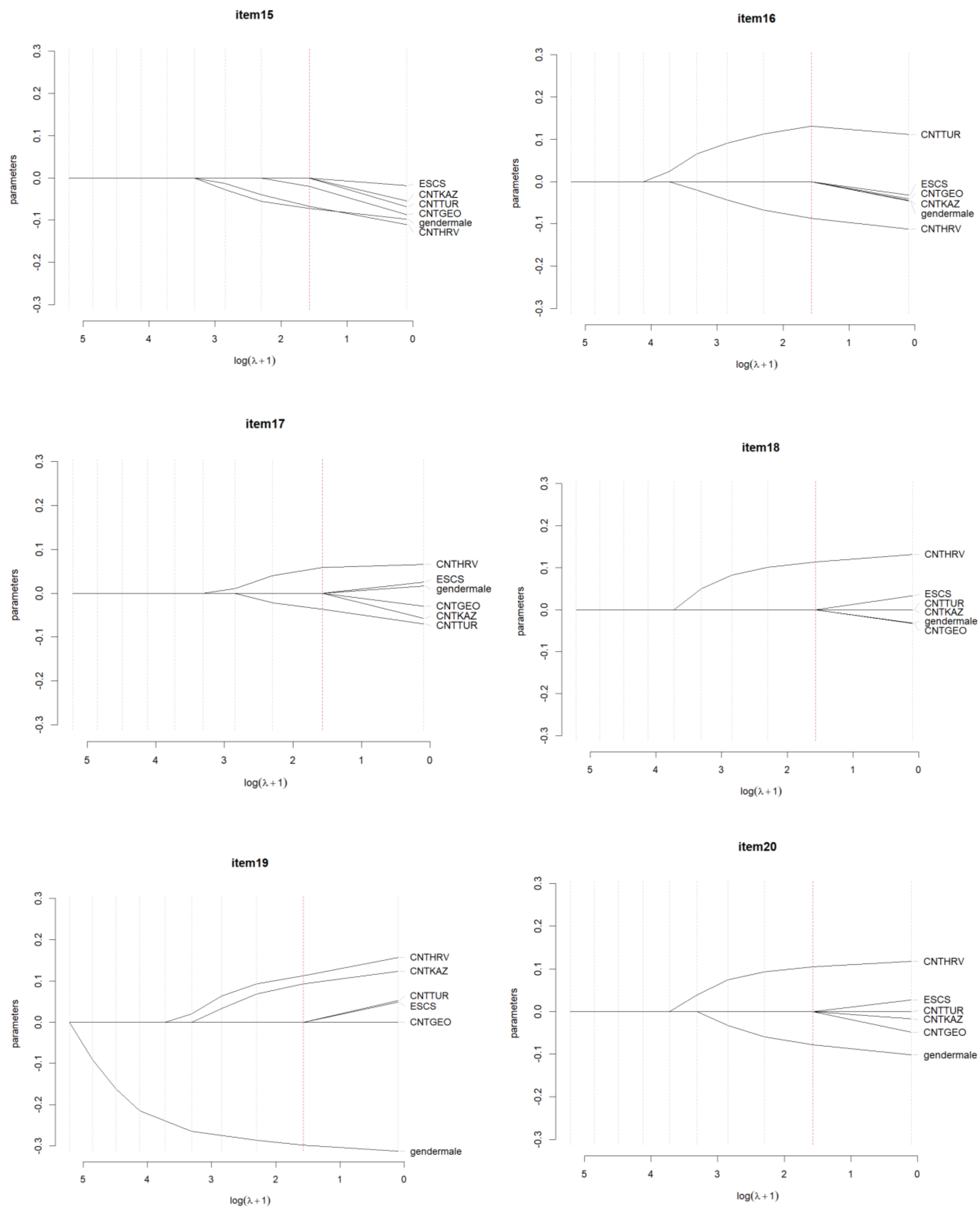

item10
parameters
log(λ+1)
CNTKAZ
CNTGEO
CNTTUR
CNTHRV
gendermale
ESCS


item11
parameters
log(λ+1)
CNTHRV
CNTTUR
CNTKAZ
CNTGEO
ESCS
gendermale


item12
parameters
log(λ+1)
CNTHRV
ESCS
CNTTUR
CNTGEO
CNTKAZ
gendermale


item13
parameters
log(λ+1)
CNTHRV
ESCS
CNTTUR
CNTGEO
CNTKAZ
gendermale


item14
parameters
log(λ+1)
gendermale
CNTKAZ
ESCS
CNTGEO
CNTTUR
CNTHRV


item15
ESCS
CNTKAZ
CNTTUR
CNTGEO
gendermale
CNTHRV
item16
CNTTUR
ESCS
CNTGEO
CNTKAZ
gendermale
CNTHRV
item17
CNTHRV
ESCS
gendermale
CNTGEO
CNTKAZ
CNTTUR
item18
CNTHRV
ESCS
CNTTUR
CNTKAZ
gendermale
CNTGEO
item19
CNTHRV
CNTKAZ
CNTTUR
ESCS
CNTGEO
gendermale
item20
CNTHRV
ESCS
CNTTUR
CNTKAZ
CNTGEO
gendermale
parameters
log(λ+1)

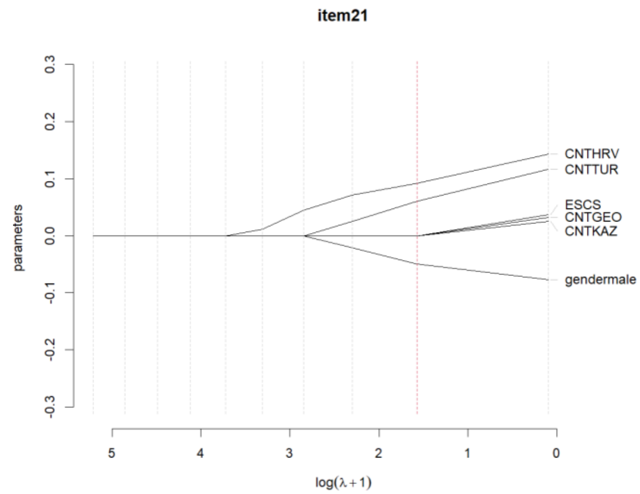


Figure 1. Coefficient paths of all DIF parameters for GPCMlasso applied to ICT scale, separately for each item. *Dashed vertical lines represent the optimal model according to BIC.*

The red dashed lines represent the optimal model according to BIC. If the parameter path for a specific variable crosses the vertical red dashed line an item shows DIF. For example, in the Interest in ICT subscale, item 2 shows DIF in favor of males and against Croatia (HRV) because the parameter path for the group variable crosses the vertical red dashed line. Overall, in ICT questionnaire, only items 10,11 are diagnosed to be completely DIF-free. The GPCMlasso coefficients presented in Table 3 are different from zero, but when examining the paths related to the coefficients to which the penalty parameter is applied, an item containing DIF according to the GPCMlasso coefficients may not show DIF according to the optimal value. This may be due to the fact that the substance contains negligible DIF or that the λ adjustment parameter cannot be determined properly.

4. Discussion and Conclusions

This study aimed to investigate the differential item functioning of the PISA 2018 ICT questionnaire by applying the GPCMlasso method, a machine learning algorithm. Overall, while in the ICT questionnaire, only items 10,11 are diagnosed as completely DIF-free, the rest contain DIF based on country, gender, or ESCS. From the plots in Figure 1, detailed insights were gathered regarding which items exhibited DIF for each variable, as well as the direction of the DIF. According to the GPCMlasso coefficient, when reference group Bulgaria and focus group Georgia, Croatia, Kazakhstan and Turkey are compared, 28%, 71%, 33% and 33% of all items hold DIF, respectively. 14% of items for SES and 62% of items for gender show DIF. In the literature, Odell et al. (2021) employed the alignment technique to examine the measurement invariance of PISA 2015

ICT scales over 47 countries, revealing inconsistencies in the ICT Familiarity Questionnaire across countries. Conversely, using multi-group confirmatory factor analysis, Ma & Qin (2021) found residual (strict) measurement invariance for PISA 2018 ICT scales across 16 countries. Using exploratory structural equivalent modeling (ESEM), Meng et al. (2019) assessed the PISA 2015 ICT scales' measurement invariance between China and Germany, demonstrating strong structural invariance and model fit at the scalar level. As seen in these studies, examinations at the item level have not been conducted GPCMlasso surpassed conventional DIF detection methods, enabling simultaneous assessment of multiple variables and testing of multidimensional frameworks. In this research, the ICT questionnaire composed of four subscales was concurrently evaluated for DIF based on country, gender, and ESCS factors. Country and gender are categorical variables, whereas ESCS is continuous. The GPCM lasso stands out for uncovering the source of DIF by analyzing various types of variables concurrently, unlike traditional methods.

5. Recommendations

There are similarities and inconsistencies between the outcomes of this research and previous literature findings. Thus, to discern which method can most precisely identify DIF under specific conditions, a simulation study examining methods rooted in varying approaches might be beneficial.

References

Belzak, W., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods, 25*(6), 673. https://doi.org/10.1037/met0000253.

Bauer DJ, Belzak WCM, Cole V. (2020) Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning. Struct Equ Modeling.27(1):43-55. https://doi.org/10.1080/10705511.2019.1642754

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods, 25*(6), 673–690. https://doi.org/10.1037/met0000253

Belzak, W. C. (2023). The Multidimensionality of Measurement Bias in High‐Stakes Testing: Using Machine Learning to Evaluate Complex Sources of Differential Item Functioning. *Educational Measurement: Issues and Practice*, *42*(1), 24-33. **https://doi.org/10.1111/emip.12486**

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items an adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33*(3), 333–353.

Choi, S., Gibbons, L., & Crane, P. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(1), 1–30. https://doi.org/10.18637/jss.v039.i08

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of research in personality*, *19*(2), 109-134. https://doi.org/10.1016/0092-6566(85)90023-6

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.

Jafari, P., Mehrabani-Zeinabad, K., Javadi, S. et al. (2022). A Machine Learning Approach to Assess Differential Item Functioning of the KINDL Quality of Life Questionnaire Across Children with and Without ADHD. *Child Psychiatry Hum Dev 53*, 980–991. https://doi.org/10.1007/s10578-021-01179-6

Karasar, N. (2017). Bilimsel araştırma yöntemi (2. yazım, 32. Basım). Nobel Yayın Dağıtım.

Liang, X., & Jacobucci, R. (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 722-734. https://doi.org/10.1080/10705511.2019.1693273

Ma, Y., & Qin, X. (2021). Measurement invariance of information, communication and technology (ICT) engagement and its relationship with student academic literacy: Evidence from PISA 2018. *Studies in Educational Evaluation, 68*, 100982. https://doi.org/10.1016/j.stueduc.2021.100982

Millsap, R.E. (2011). *Statistical approaches to measurement invariance.* Taylor & Francis.

Odell, B., Gierl, M., & Cutumisu, M. (2021). Testing measurement invariance of PISA 2015 mathematics, science, and ICT scales using the alignment method. *Studies in Educational Evaluation*, *68*, 100965. https://doi.org/10.1016/j.stueduc.2020.100965

OECD. (2019). *OECD skills outlook 2019: Thriving in a digital world*. Paris, France: OECD.

OECD. (2019a). PISA 2018 assessment and analytical framework. Paris: PISA, OECD Publishing.

OECD. (2019b). PISA 2018 technical report. Paris: PISA, OECD Publishing.

OECD, U. (2021). Education in Eastern Europe and Central Asia: Findings from PISA.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/

Schauberger, G., Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. *Behav Res 52,* 279–294 (2020). https://doi.org/10.3758/s13428-019-01224-2

Schauberger, G., & Tutz, G. (2016). Detection of differential item functioning in Rasch models by boosting techniques. *British Journal of Mathematical and Statistical Psychology*, *69*(1), 80–103. **https://doi.org/10.1111/bmsp.12060**

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika, 80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

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

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In Holland, P. W., & Wainer, H. (Eds.) Differential item functioning, (pp. 67-113).

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. Psychometrika, 80(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727–750. https://doi.org/10.48550/arXiv.1502.04329

Xu, H., & Tracey, T. J. G. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. The European Journal of Counselling Psychology, 6(1), 75–82. https://doi.org/10.5964/ejcop.v6i1.120

Zylka, J., Christoph, G., Kroehne, U., Hartig, J., & Goldhammer, F. (2015). Moving beyond cognitive elements of ICT literacy: First evidence on the structure of ICT engagement. *Computers in Human Behavior*, *53*, 149-160. https://doi.org/10.1016/j.chb.2015.07.008