Differential Item Functioning Analysis in PISA 2022 Using Rasch Trees: Finland, Turkey, and Singapore

Enes Yavuz[1]

[1]Gebze Technical University, 41400 Gebze, Kocaeli, Turkey
enesyavuz@gtu.edu.tr

Abstract. This study investigates Differential Item Functioning (DIF) in the PISA 2022 mathematics assessment to examine measurement invariance across Finland, Singapore, and Turkey. These countries represent diverse educational contexts: Finland as a reference for European quality, Singapore as both an Asian and a global leader in academic achievement, and Turkey as a cultural and geopolitical bridge between Asia and Europe. Using the official OECD PISA 2022 database with a sample of approximately 24,000 students, the research employs a robust, data-driven psychometric methodology based on Rasch Trees. Unlike traditional DIF detection techniques that rely on predefined group comparisons, this model-based recursive partitioning approach identifies latent subgroups where item parameters vary systematically through an exploratory lens. To rigorously evaluate the stability and robustness of the findings, three distinct analytical scenarios are implemented: (i) a baseline model using a complete set of common items, (ii) a sample augmentation model employing bootstrapping procedures, and (iii) an extended model incorporating multiple imputations to address missing responses across 13 items. Across all scenarios, Country consistently emerges as the primary source of DIF ($p < 0.001$), with Singapore exhibiting a highly distinct item difficulty structure. Additionally, Gender is identified as a statistically significant secondary split within the Finnish sample ($p = 0.018$ in the baseline model). By offering interpretable, tree-based representations of measurement non-invariance, this study underscores the value of advanced psychometric methods for enhancing fairness and cross-cultural validity in international large-scale assessments.

Keywords: Rasch Trees, Differential Item Functioning (DIF), PISA 2022, Measurement Invariance, Recursive Partitioning, Cross-Cultural Analysis.

1 Introduction

International large-scale assessments (ILSAs) such as the Programme for International Student Assessment (PISA) have become cornerstone instruments for global educational benchmarking. These assessments provide policymakers and educators with a comparative lens through which to evaluate the effectiveness of national education systems and guide large-scale reforms. However, the utility of such cross-national comparisons rests entirely on the assumption of measurement invariance; the

© The Author(s) 2026
C. Y. Lye et al. (eds.), *Proceedings of the Pacific Rim Objective Measurement Symposium (PROMS 2025)*, Atlantis Highlights in Social Sciences, Education and Humanities 56,
https://doi.org/10.2991/978-94-6239-834-4_15

principle that the instruments used across diverse cultures, languages and educational contexts measure the same underlying construct in the same way [10].

Differential Item Functioning (DIF) represents a significant threat to this validity [3]. DIF occurs when individuals from different groups, who possess identical levels of the underlying latent trait, have different probabilities of endorsing or answering an item correctly. In the complex landscape of PISA, where test items are translated into multiple languages and administered across diverse curricular traditions, item-level bias can easily emerge from cultural nuances, linguistic variations or differences in instructional emphasis. If DIF remains undetected, observed performance disparities between countries may reflect measurement artifacts rather than genuine differences in student competencies [10].

Traditional approaches to DIF detection such as the Mantel–Haenszel procedure [16] or logistic regression [7] have long served as the industry standard. However, these methods exhibit notable limitations in exploratory contexts. They typically require the *a priori* definition of comparison groups and often struggle to account for complex interactions between multiple covariates. In a globalized educational dataset, bias may not stem from a single grouping variable (e.g., just country) but from the interaction of country, gender and language background. Consequently, there is an increasing need for more flexible and data-driven methodologies capable of uncovering latent subgroup structures without the constraints of predefined group comparisons.

This study employs Rasch Trees [12], a model-based recursive partitioning approach [15] to investigate measurement invariance in the PISA 2022 mathematics assessment. By integrating the theoretical rigor of the Rasch model with the exploratory power of tree-based partitioning, this method allows for the simultaneous evaluation of multiple covariates and the visualization of complex DIF patterns.

The analysis focuses on three strategically selected countries: Finland, Singapore, and Turkey. These nations represent distinct educational archetypes: Finland is renowned for its European excellence, Singapore stands as both an Asian and a global leader in academic achievement and Turkey serves as a unique cultural and geopolitical bridge between Europe and Asia. By applying Rasch Tree models across three analytical scenarios; baseline, sample augmentation and missing data imputation, this research aims to evaluate the stability of measurement invariance and provide a more nuanced understanding of how cultural and contextual factors influence item performance in international assessments.

2 Theoretical Background

Differential Item Functioning (DIF) has long been recognized as a critical threat to the validity of test score interpretations, particularly in large-scale educational assessments. DIF occurs when individuals from different groups who possess equivalent levels of

the underlying latent trait, respond differently to specific test items due to factors unrelated to the construct being measured [3]. In the context of international large-scale assessments, such as the Programme for International Student Assessment (PISA), DIF poses a substantial challenge as it may compromise the comparability of results across countries, cultures and demographic subgroups.

The importance of DIF analysis in international assessments extends beyond technical measurement concerns and directly affects educational policy and policy-oriented interpretations. PISA results are frequently used to benchmark national education systems, inform reforms and guide resource allocation. When item-level bias is present but undetected, observed differences in country performance may reflect measurement artifacts rather than genuine disparities in student competencies [11]. As a result, ensuring measurement invariance across groups has become a central methodological requirement in cross-national assessment research.

Previous research has highlighted that cross-national DIF may emerge from a variety of sources, including cultural norms, curricular emphasis, language-related nuances and differences in instructional practices [3, 17]. In assessments such as PISA where items are designed to be culturally neutral yet administered across highly diverse educational systems, even carefully constructed items may function differently across national contexts. These challenges underscore the necessity of methodological approaches capable of detecting latent and multidimensional sources of measurement non-invariance [11].

Traditional approaches to DIF detection have been extensively developed and applied in educational measurement. Classical methods such as the Mantel–Haenszel procedure, logistic regression DIF analysis and analysis of variance-based techniques rely on predefined group comparisons and assume that potential sources of bias are known a priori [7, 16]. Within the item response theory (IRT) framework, multi-group models have been widely used to assess parameter invariance by estimating item parameters separately for different groups and comparing them statistically [17]. While these methods have proven effective in many applied settings, they exhibit notable limitations in complex and heterogeneous data structures.

One major limitation of traditional DIF methods lies in their reliance on single, predefined grouping variables. In international datasets such as PISA, student responses may be influenced by multiple interacting factors, including country, gender, language background and educational context. When DIF is driven by such interactions, conventional methods may fail to detect nuanced patterns of item bias or require an impractical number of separate analyses. Moreover, predefined group comparisons may obscure latent subgroup structures that are not directly observable or theoretically anticipated, limiting the exploratory potential of classical DIF techniques.

The Rasch model has been widely regarded as a foundational framework for addressing issues of measurement invariance. By modeling person ability and item

difficulty on a common latent scale, the Rasch model embodies the principle of specific objectivity, whereby comparisons between individuals are independent of the particular items administered and vice versa [8]. This property has made the Rasch model particularly attractive for cross-cultural and cross-national assessment contexts where invariant measurement is essential for valid comparisons [2].

Despite its theoretical strengths, the standard Rasch model assumes that item parameters are invariant across all respondents, an assumption that may be violated in heterogeneous populations. Empirical studies have demonstrated that even when overall model fit is acceptable, item-level parameter instability may persist across subgroups defined by demographic or contextual characteristics [1]. Consequently, extensions of the Rasch framework have been proposed to relax the assumption of global parameter invariance and allow for the systematic investigation of DIF within a Rasch-based measurement paradigm [14].

In cross-national assessments, the tension between the ideal of specific objectivity and the empirical reality of heterogeneous populations becomes particularly pronounced. Countries participating in PISA differ substantially in curricular organization, pedagogical traditions and sociocultural expectations regarding schooling. For example, Finland has often been characterized by its emphasis on equity and student-centered learning, while Singapore is frequently cited for its highly structured, performance-oriented education system [6]. Turkey, by contrast, occupies a unique position shaped by both European and Asian educational influences. Such contextual diversity increases the likelihood that item parameters may not remain invariant across countries, even when overall ability distributions are controlled [11].

In response to these challenges, model-based recursive partitioning approaches have emerged as flexible tools for detecting parameter heterogeneity. Zeileis, Hothorn, and Hornik [15] introduced a general framework for model-based recursive partitioning, which enables statistical models to be recursively fitted to subgroups defined by covariates that induce significant parameter instability. This framework provides a principled way to combine parametric modeling with tree-based partitioning, allowing researchers to explore how model parameters vary across different segments of the population.

Building on this framework, Strobl, Kopf, and Zeileis [12] introduced Rasch Trees as a novel method for DIF detection. Rasch Trees integrate the Rasch model with recursive partitioning techniques to identify subgroups in which item difficulty parameters differ significantly. Unlike traditional DIF methods, Rasch Trees do not rely exclusively on predefined group comparisons. Instead, they evaluate multiple covariates simultaneously and recursively partition the sample based on statistically significant parameter changes. This data-driven approach allows for the detection of complex DIF structures that may arise from interactions among multiple grouping variables.

A key advantage of Rasch Trees is their interpretability. The resulting tree structure provides a transparent and intuitive visualization of how measurement non-invariance unfolds across subgroups. Each split in the tree corresponds to a covariate that induces item parameter instability, while terminal nodes represent relatively homogeneous subgroups in which the Rasch model holds locally. This combination of statistical rigor and visual clarity makes Rasch Trees particularly well suited for exploratory analyses in complex assessment datasets [12].

The emphasis on visualization and interpretability is especially relevant in policy-oriented assessments such as PISA. Findings from international large-scale assessments are frequently communicated not only to researchers but also to policymakers and stakeholders without specialized training in psychometrics. Tree-based representations of DIF offer an accessible way to illustrate where and how measurement bias emerges, facilitating more informed interpretations of assessment results and reducing the risk of oversimplified cross-country comparisons.

Subsequent research has demonstrated the applicability of Rasch Trees in educational and psychological measurement contexts. Komboz et al. [4] extended the Rasch Tree methodology by proposing conditional inference procedures for DIF detection, further enhancing its robustness in finite samples. Empirical applications have shown that Rasch Trees can uncover meaningful DIF patterns that remain undetected by traditional methods, particularly in datasets characterized by multiple covariates and heterogeneous populations.

In the context of PISA research, measurement invariance has been extensively studied using both classical and IRT-based approaches. Several studies have highlighted challenges related to cross-country comparability and the potential presence of item bias in PISA assessments [11]. These studies emphasize that violations of measurement invariance may stem not only from item content but also from differences in test administration, student engagement and assessment design [10, 17].

An additional methodological challenge in PISA research concerns the handling of missing data. Due to its rotated booklet design, not all students respond to all cognitive items, resulting in substantial planned missingness at the item level [6]. This design feature, while efficient from a testing perspective, complicates item-level analyses and increases uncertainty in parameter estimation [5]. Researchers have noted that naïve handling of missing data may exacerbate bias or undermine model stability, particularly in exploratory DIF analyses.

Consequently, the literature has increasingly emphasized the importance of principled missing data handling strategies in large-scale assessments. Approaches such as multiple imputation [9] and resampling techniques have been proposed to preserve statistical power and reduce estimation bias in the presence of incomplete data [11, 13]. Integrating such strategies with Rasch-based models allows researchers to

balance methodological rigor with the practical constraints imposed by complex assessment designs.

Taken together, the existing literature underscores the need for flexible, data-driven methods that can accommodate heterogeneity and uncover latent subgroup structures in international assessments. Rasch Trees offer a promising methodological approach by combining the theoretical strengths of the Rasch model with the exploratory power of recursive partitioning. By applying Rasch Tree models to PISA 2022 data, the present study contributes to the growing body of research on measurement invariance and DIF detection in cross-national assessments, while addressing a methodological gap in the existing PISA literature.

3 Methodology

The data for this study was obtained from the OECD PISA 2022 database, specifically the "student questionnaire data" and "cognitive item data"[6]. Approximately 24,000 students are included in this dataset, and the study was conducted specifically focusing on PISA 2022 Mathematics items. Integrating these datasets was essential to examine the interaction between students' personal profiles and their responses to cognitive items. While the PISA 2022 database encompasses a global scale, this study focuses on three specific countries: Finland, Turkey, and Singapore. Finland was selected due to its long-standing reputation for educational excellence. Singapore was included as it represents the highest tier of academic performance in international assessments, serving as both an Asian and a global leader. Turkey was chosen for its unique geopolitical and cultural position as a bridge between Asia and Europe, incorporating diverse educational elements from both contexts. To provide a quantitative context for these selections, Table 1 presents the mean mathematics scores of the three included countries in comparison with the overall Organisation for Economic Co-operation and Development (OECD) average. These performance metrics highlight the diverse achievement levels across the selected educational systems, ranging from Singapore's top-tier results to Finland's stable European performance and Turkey's evolving academic standing.

Table 1. Mean mathematics performance in PISA 2022 for selected countries.

Country	**Mean Score**
Finland	484
Turkey	453
Singapore	575
OECD Average	472

To ensure linguistic homogeneity and mitigate potential language related bias in the Finnish dataset, Swedish-speaking students (representing approximately 4% of the Finnish sample) were excluded from the analysis. To control for developmental differences and potential age related bias, the sample was restricted to students born in 2006. Furthermore, to minimize variance stemming from differing educational levels, only 9th and 10th-grade students were included in the scope of this study.

3.1 Rasch Model

The fundamental Rasch model defines the probability of a correct response as a function of the difference between a person's ability (θ_n) and an item's difficulty (β_i). The probability is expressed through the following logistic function [8]:

$$P(X_{ni} = 1|\theta_n, \beta_i) = \frac{\exp(\theta_n - \beta_i)}{1 + \exp(\theta_n - \beta_i)}$$

In the context of PISA 2022, these challenges are further intensified by the assessment's large scale, multilingual administration and rotated booklet design [6]. Consequently, while the Rasch model provides a theoretically rigorous foundation for measurement invariance, extensions that allow for the identification of subgroup-specific parameter heterogeneity are required [14]. This motivates the use of Rasch Tree models [12], which integrate the Rasch framework with recursive partitioning techniques to detect Differential Item Functioning across covariates such as country and gender.

3.2 Rasch Trees

PISA 2022 datasets present significant challenges regarding missing data; an analysis of items intended for all students revealed a missingness rate of approximately 88%. While a booklet-level analysis was considered, the resulting reduction in sample size would have compromised the stability of the Rasch Tree models. Consequently, this exploratory study implemented three distinct analytical scenarios to evaluate model stability.

Choice of Rasch Trees over traditional Differential Item Functioning (DIF) methods is justified by their flexible and data-driven approach, which allows for the detection of DIF without relying solely on single and predefined comparisons. Rasch Trees utilize predefined grouping variables while combining Rasch modeling with clear, tree-based visualizations to provide a more precise and intuitive illustration of measurement bias. This method supports the systematic comparison of multiple subgroups and ensures high interpretability, effectively illustrating complex DIF patterns in a structured manner [12].

A significant methodological hurdle in analyzing PISA datasets is the structural missingness inherent in its design. Due to the rotated booklet design, each student

responds to only a small subset of the total item pool, resulting in a planned missingness rate of approximately 88% at the item level . Such an extensive missing data structure poses a severe challenge for standard Rasch-based analyses, as it can compromise the stability of item parameter estimation and model convergence. Recognizing this limitation, this study intentionally avoids a single-model approach and instead implements three distinct analytical scenarios, Baseline, Sample Augmentation and Multiple Imputation to rigorously evaluate the robustness of Differential Item Functioning (DIF) patterns under varying levels of data density. Each scenario was evaluated on independently constructed datasets, allowing the assessment of DIF stability across different sample realizations.

3.3 Baseline Scenario

An analysis of a core set of 8 items to which students from all three countries provided common responses. The primary objective of the baseline model is to investigate measurement invariance in its most "unfiltered" state, free from the potential influence of sample augmentation or missing data imputation. By utilizing only observed responses, this scenario provides a direct assessment of whether the Rasch Tree algorithm identifies "Country" and "Gender" as stable split variables under standard conditions. Furthermore, the results from this baseline analysis are critical for validating the robustness of the more complex scenarios. Any Differential Item Functioning (DIF) patterns identified here serve as a diagnostic tool to determine if the item difficulty profiles remain consistent when the data density is later increased through bootstrapping or MICE procedures [13]. This hierarchical validation strategy supports the interpretation that the identified measurement non-invariance reflects a structural characteristic of the assessment items rather than a byproduct of specific missing data treatments. The items utilized in this specific scenario include CM948Q01S, CM948Q02S, CM948Q03S, CMA113Q01S, CMA113Q02S, CMA113Q03S, DMA124Q03C and CMA124Q02S.

3.4 Sample Augmentation Scenario

An analysis of eight items was conducted using a bootstrap-based resampling procedure to increase data density for robust parameter estimation. Bootstrap resampling with replacement was applied to generate an augmented dataset based on repeated realizations of observed response patterns. This simulation-based strategy was employed to assess the stability and feasibility of Rasch Tree estimation under enhanced data density, rather than to introduce new information or to imply an effective increase in independent sample size. The items included in this scenario were CMA143Q01S, CMA143Q02S, CMA143Q03S, CMA143Q04S, CMA147Q01S, CMA147Q02S, CMA147Q03S and CMA147Q04S.

3.5 Missing Data Imputation Scenario

This scenario involved an expanded analysis of 13 items where missing responses were estimated using multiple imputation techniques to maintain statistical power. Missing responses were addressed using Multiple Imputation by Chained Equations (MICE) in order to minimize information loss and reduce potential bias associated with listwise deletion [13]. Given the dichotomous (0/1) nature of the item responses and the requirements of Rasch-based modeling, a logistic regression imputation model (logreg) was employed for all test items. This approach is particularly appropriate for binary variables, as it preserves the categorical structure of the data and ensures that imputed values remain within the original response scale. The number of imputations was set to $m = 20$, consistent with methodological recommendations for settings with substantial missing data, where a sufficiently large number of imputations is required to ensure stable parameter estimates [9]. The imputation process was iterated for 50 cycles ($maxit = 50$) to ensure convergence and stability of parameter estimates. A fixed random seed ($seed = 123$) was specified to guarantee the reproducibility of the imputation procedure.

Importantly, only item responses were subjected to imputation, while grouping variables used for Rasch Tree splits (e.g., Country and Gender) were retained as observed predictors and not imputed. Following imputation, the completed datasets were transformed back into numeric binary format (0/1) to meet the assumptions of the Rasch model and subsequent Rasch Tree analyses. This strategy allows for a principled treatment of missing data that is consistent with the probabilistic framework of Rasch modeling and enables a more robust examination of differential item functioning across subgroups. Certain DM and DMA-coded components related to mathematics tasks were retained for exploratory robustness checks. The items utilized in this specific scenario include CM905Q01S, CMA109Q01S, CMA109Q02S, CMA109Q03S, CMA128Q01S, CMA128Q02S, CMA128Q03S, CMA161Q01S, CMA161Q02S, CMA161Q03S, DMA161Q04C, DM00KQ02C and DM905Q02C.

3.6 Software Implementation

Rasch Tree models were executed in the R environment using the `psychotree` package. The variables "Country" and "Gender" were utilized as split variables in the recursive partitioning process. A minimum group size (`minbucket`) was set to 30 to ensure reliable and stable item parameter estimates across all terminal nodes. The following R code illustrates the specific implementation for the first Rasch Tree (Baseline Scenario) only, serving as a template for the subsequent analyses:

```
library(psychotree)
library(haven)
# Defining variables as factors
rasch$Country <- as.factor(rasch$Country)
rasch$Gender <- as.factor(rasch$Gender)
# Rasch Tree model specification
```

```
rasch_tree <- raschtree(
  cbind(CM948Q01S, CM948Q02S, CM948Q03S,
        CMA113Q01S, CMA113Q02S, CMA113Q03S,
        DMA124Q03C, CMA124Q02S) ~
  Country + Gender,
  data = rasch,
  minbucket = 30
)
# Visualizing and summarizing results
plot(rasch_tree)
summary(rasch_tree)
```

All analyses were conducted using open source R software and the analytical workflow is fully reproducible given access to the PISA 2022 public-use datasets [6] and the `psychotree` package.

4 Findings

This section presents the empirical results of the Rasch Tree analyses conducted across three distinct analytical scenarios: Baseline, Sample Size Augmentation and Missing Data Imputation. The findings consistently reveal that the assumption of global measurement invariance is challenged by substantial conditional parameter instability across educational contexts, with "Country" emerging as the primary driver of item parameter instability in every model. Statistical fit indices across all scenarios, ranging from the 8-item core set to the expanded 13-item set, generally support the partitioned tree structures over the global Rasch models, confirming the presence of complex Differential Item Functioning (DIF) patterns that vary by both country and gender. These results underscore the stability of the latent subgroup structures identified through the model-based recursive partitioning approach.

4.1 Baseline Scenario

The first stage of the analysis involved the baseline scenario, which focused on a core set of 8 common items with no missing values or sample size augmentation. The model fit statistics for this scenario are presented in Table 2. The recursive partitioning yielded a significant improvement in model fit; the log-likelihood increased from -637.158 in the null model to -588.426 in the split model. While the BIC was slightly higher for the split model (1352.931 vs. 1315.402), the marked reduction in the AIC (1236.852 vs. 1288.317) provides strong statistical support for the presence of parameter heterogeneity across the identified subgroups.

Table 2. Model fit statistics for the baseline scenario.

Model	Log-Likelihood	AIC	BIC
Split	-588.426	1236.852	1352.931
No Split	-637.158	1288.317	1315.402

As illustrated in Figure 1, the Rasch Tree analysis identified "Country" as the primary source of DIF ($p < 0.001$), creating an initial split that separated Finland from Singapore and Turkey. For the Finnish sample, a further significant split occurred based on "Gender" ($p = 0.018$), resulting in two distinct terminal nodes: Node 3 for female students ($n = 54$) and Node 4 for male students ($n = 59$).

The right branch of the tree, comprising Singapore and Turkey, was further partitioned by "Country" ($p < 0.001$) at Node 5. This split highlights substantial differences in item difficulty profiles between the two nations, resulting in Node 6 for Singapore ($n = 39$) and Node 7 for Turkey ($n = 202$).

The item difficulty profiles in the terminal nodes reveal specific patterns of measurement non-invariance. For instance, in Node 4, a significant decrease in the difficulty of Item 7 is observed, suggesting a potential advantage for this subgroup on that specific item compared to others. In contrast, the profile for Turkey (Node 7) shows a relatively stable difficulty across items 5 through 8, following a peak at item 3.

A notable observation in Node 4 (Finnish males) is the absence of a difficulty parameter for the 8th item (CMA124Q02S). In Rasch modeling, item parameters cannot be estimated when a subgroup exhibits an extreme response pattern, meaning all respondents in that node either answered correctly or incorrectly. Given that Item 7 already showed a significant drop in difficulty (-4.35 logit) for this group, the non-estimable status of Item 8 likely indicates a ceiling effect, where the item was too easy for Finnish male students to provide meaningful measurement variation.

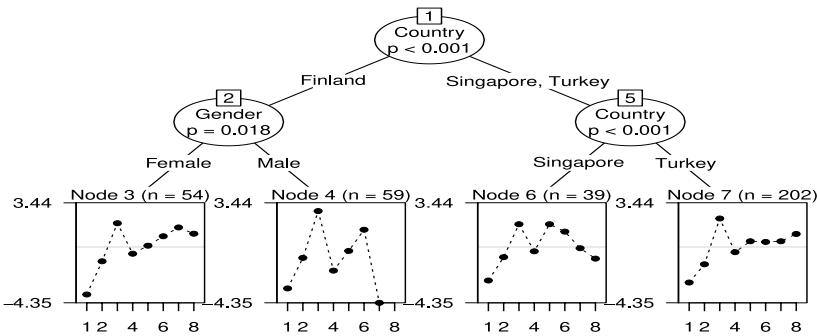


Fig. 1. Rasch tree analysis for the baseline scenario using eight common mathematics items, illustrating a primary partition by Country and a statistically significant secondary split by Gender within the Finnish cohort.

4.2 Sample Size Augmentation

The second analytical scenario was conducted to evaluate the stability and robustness of the Differential Item Functioning (DIF) patterns by artificially increasing the sample size through bootstrapping. This approach aims to ensure that the parameter estimates remain reliable, particularly in subgroups where the original sample size might be considered marginal. The model fit statistics for this scenario are presented in Table 3.

Table 3. Model fit statistics for the sample augmentation scenario.

Model	**Log-Likelihood**	**AIC**	**BIC**
Split	-620.998	1299.995	1410.864
No Split	-668.303	1350.607	1377.368

Consistent with the baseline results, the partitioned model demonstrated a superior fit compared to the global model. The log-likelihood improved from -668.303 to -620.998 and the AIC showed a clear reduction from 1350.607 to 1299.995. While the BIC for the split model (1410.864) was higher than the no-split model (1377.368), this is interpreted because of the stricter penalty BIC applies for the additional parameters introduced by the tree structure.

As illustrated in Figure 2, the hierarchical structure of the tree remained largely consistent with the baseline scenario. The primary split again occurred based on "Country" ($p < 0.001$), separating Finland from Singapore and Turkey. Within the Finnish sample, "Gender" remained a significant split variable ($p = 0.031$), resulting in terminal nodes for females (Node 3, $n = 72$) and males (Node 4, $n = 52$). For the combined group of Singapore and Turkey, the secondary split by "Country" was also maintained ($p = 0.027$), yielding distinct profiles for Singapore (Node 6, $n = 106$) and Turkey (Node 7, $n = 108$).

Interestingly, the item difficulty profile for Singapore (Node 6) reveals high logit values for specific items (Item 5 and Item 6) indicating that these tasks are relatively more demanding for Singaporean students compared to their Turkish counterparts (Node 7). This divergence emphasizes that a high national mean score does not necessarily equate to lower perceived difficulty across all individual tasks; rather, it reflects a unique parameter heterogeneity where certain item properties function differently within the Singaporean educational context. Such findings suggest that the observed measurement non-invariance may be driven by country-specific factors, such as curricular emphasis or instructional practices, which can make certain items relatively more difficult even for high-performing cohorts.

The item difficulty profile for Finnish males (Node 4) is characterized by an extreme polarization between Item 5 and Item 6. While Item 5 exhibits a significant drop in difficulty, becoming the easiest task for this subgroup with a near-minimum logit value, Item 6 reaches the upper threshold of the difficulty scale at +2.95. Furthermore, Item 7 also has a high difficulty level for this group, distinguishing their response pattern from Finnish females (Node 3), for whom Item 7 appears relatively easier.

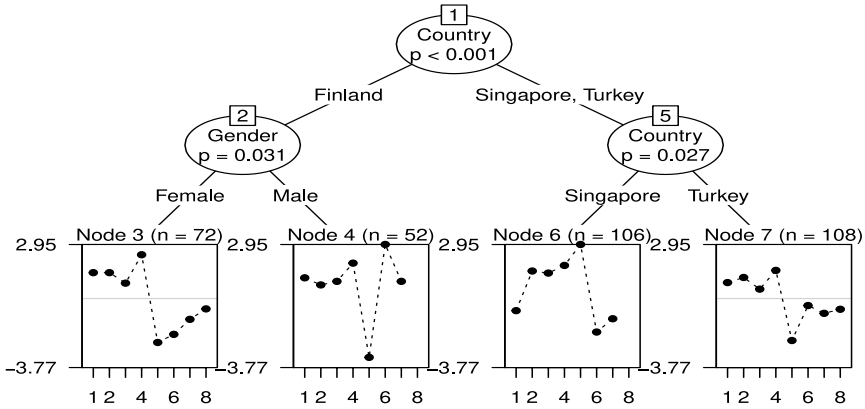


Fig. 2. Rasch tree analysis for the sample size augmentation scenario (eight items) using bootstrap resampling, highlighting the structural stability of identified DIF patterns across Country and Gender covariates.

4.3 Missing Data Imputation

The third stage of the analysis focused on an expanded set of 13 mathematics items, addressing the "planned missingness" inherent in the PISA 2022 rotated booklet design. To preserve statistical power and reduce estimation bias, missing responses were estimated using MICE techniques. The comparison of model fit statistics between the partitioned Rasch Tree and the global (no split) model is presented in Table 4 .

Table 4. Model fit statistics for the missing data imputation scenario.

Model	Log-Likelihood	AIC	BIC
Split	-6650.008	13376.02	13577.99
No Split	-6779.046	13582.09	13645.87

As shown in Table 4, the partitioned model demonstrated a superior fit compared to the global model across all indicators. The log-likelihood improved significantly from -6779.046 to -6650.008. Notably, in this scenario, both the AIC (decreasing from 13582.09 to 13376.02) and the BIC (decreasing from 13645.87 to 13577.99) clearly favored the Rasch Tree structure. The substantial reduction in BIC, despite the penalty

for model complexity, provides robust evidence for the existence of country-level parameter heterogeneity within the expanded item set.

The resulting tree structure, illustrated in Figure 3, reveals a distinct pattern of Differential Item Functioning (DIF). Primary Split (Node 1): The initial partitioning was driven by "Country" ($p < 0.001$), isolating Singapore from the European/Eurasian group (Finland and Turkey). Secondary Split (Node 3): Within the combined branch of Finland and Turkey, a further significant split based on "Country" was identified ($p < 0.001$), confirming that item difficulty parameters vary systematically even between these two contexts. Terminal Nodes: The final model resulted in three distinct terminal nodes representing each participating nation: Singapore (Node 2, $n = 559$), Finland (Node 4, $n = 455$) and Turkey (Node 5, $n = 489$). The item difficulty profiles for these nodes exhibit unique trajectories across the 13 items. For example, Singapore's profile (Node 2) shows specific peaks and troughs in difficulty that differ markedly from the patterns observed in Finland (Node 4) and Turkey (Node 5). These variations underscore the importance of accounting for cross-national differences when interpreting international assessment results.

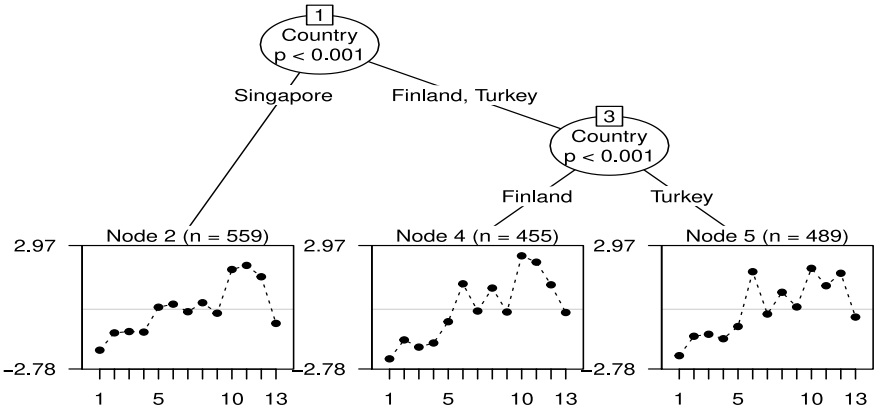


Fig. 3. Rasch tree analysis for the missing data imputation scenario (13 items), demonstrating the hierarchical country-level splits and the distinct item difficulty profiles for Singapore, Finland, and Turkey after MICE treatment.

5 Discussion

The findings of this study provide a comprehensive psychometric evaluation of measurement invariance in the PISA 2022 mathematics assessment. By utilizing Rasch Trees across three distinct analytical scenarios, the research successfully navigated the complexities of large-scale assessment data, specifically addressing structural missingness and sample size limitations. The consistent identification of "Country" as the primary determinant of item parameter instability underscores the pervasive nature of Differential Item Functioning (DIF) in cross-national comparisons.

It should be emphasized that Rasch Tree models are not intended to replace traditional global Rasch analyses but to function as a diagnostic extension for identifying conditional parameter instability within otherwise well-fitting measurement models.

5.1 Stability of Country-Level Splits Across Varying Item Subsets

A foundational contribution of this research lies in the stability of the primary splitting variable. Across the Baseline, Sample Augmentation and Missing Data Imputation scenarios, "Country" consistently emerged as the most significant source of parameter instability ($p < 0.001$). It is essential to emphasize that each scenario utilized different subsets of items and different sample configurations. The Baseline scenario relied on a core set of 8 items with only observed responses to assess invariance in its most unfiltered state, while the Missing Data Imputation scenario expanded the analysis to an expanded set of 13 items.

The fact that Singapore was isolated as a distinct terminal node across these varying conditions (even as the item pool grew from 8 to 13 items) suggests that measurement non-invariance is not an artifact of a specific item subset but a structural feature of the cross-national comparison. This reinforces the principle of specific objectivity in the Rasch model, where comparisons between groups should ideally remain invariant across different item selections. However, the empirical results indicate a persistent violation of this principle at the international level, necessitating caution when interpreting aggregate mean scores.

5.2 Gender-Based Heterogeneity and Subgroup Nuances

The exploratory power of Rasch Trees revealed a secondary layer of heterogeneity that might remain undetected in conventional IRT models. The identification of "Gender" as a significant split variable within Finland ($p = 0.018$ in baseline scenario) indicates that measurement instruments may function differently for subgroups even within educational systems renowned for their equity. Psychometrically, the ceiling effect observed in Node 4 (Finnish males) for Item 8 (CMA124Q02S) highlights a critical limitation of standardized items in heterogeneous populations. When an item becomes too easy for a specific subgroup, the item difficulty parameter cannot be reliably estimated due to extreme response patterns. The recurrence of similar item difficulty patterns of the ease of Item 7 for this subgroup across both the Baseline and Augmentation scenarios demonstrates that Rasch Trees can identify localized psychometric advantages or biases that traditional methods would likely obscure.

5.3 Methodological Resilience to Structural Missingness

The handling of "planned missingness" through Multiple Imputation by Chained Equations (MICE) provided a significant test for the stability of recursive partitioning.

In the Imputation scenario, which focused on the 13-item set, the substantial reduction in both AIC and BIC for the split model confirms that the tree-based structure provides a superior fit compared to the global model. While the BIC often penalizes the additional parameters introduced by terminal nodes, its decrease in the imputation scenario (from 13645.87 to 13577.99) suggests that the evidence for parameter heterogeneity becomes even more compelling as data density and the item pool increase. This methodological resilience is particularly relevant for large-scale assessments like PISA, where the rotated booklet design inherently complicates item-level diagnostics.

6 Conclusion

This study investigated the measurement invariance of the PISA 2022 mathematics assessment across three distinct educational and cultural contexts: Finland, Singapore, and Turkey by using a model-based recursive partitioning approach. By employing Rasch Trees, the research moved beyond traditional, pre-defined group comparisons to a data-driven exploration of item bias. The findings provide critical insights into how cultural and gender-related factors influence student performance in high-stakes international assessments. The results consistently demonstrated that "Country" serves as the most significant source of Differential Item Functioning (DIF) across all analytical scenarios. Singapore, as a global and Asian leader in academic performance, exhibited a highly distinct item difficulty profile compared to Finland and Turkey. This divergence suggests that while Singaporean students achieve superior scores, the functioning of specific items may be influenced by pedagogical or curricular nuances unique to the East Asian educational model. Furthermore, the identification of "Gender" as a secondary split variable within the Finnish sample highlights that even in systems renowned for educational equity, subtle item-level biases can persist, potentially favoring one gender over another in specific mathematical domains.

A major contribution of this research is the methodological rigor provided by the three-scenario approach (baseline, sample augmentation and missing data imputation). The stability of the DIF patterns across these scenarios reinforces the reliability of the Rasch Tree method in handling complex and large-scale datasets. Specifically, the imputation model's ability to maintain the primary country split while improving model fit indices (AIC and BIC) demonstrates the importance of sophisticated missing data treatments in psychometric research. Unlike traditional DIF detection methods, Rasch Trees offered a clear, interpretable visualization of these inequities, allowing for a more nuanced understanding of where and for whom the measurement instrument might be functioning differently.

From a policy perspective, these findings underscore the necessity for international assessment bodies to move beyond aggregate mean scores. To ensure fairness and validity in cross-cultural comparisons, it is essential to account for the latent heterogeneity identified in this study. The presence of significant DIF across countries

indicates that some performance disparities may reflect measurement artifacts rather than genuine differences in latent ability. Therefore, future PISA cycles should continue to integrate advanced psychometric tools like Rasch Trees during the item calibration and validation phases to minimize cultural and gender bias.

One of the most critical methodological takeaways from this research is the resilience of the Rasch Tree approach in the face of the approximately 88% structural missingness characteristic of PISA assessments. Despite the high sparsity of the data, the consistent emergence of 'Country' as the primary split variable across all three scenarios reinforces the conclusion that the identified measurement non-invariance is a structural feature of the international assessment rather than an artifact of missing data patterns. This stability underscores the reliability of advanced recursive partitioning for future large scale psychometric evaluations.

Despite its contributions, this study is not without limitations. First, the analysis was restricted to three countries which, while strategically selected to represent distinct educational contexts, do not capture the full heterogeneity of the PISA participant pool. Future research may therefore extend this framework by incorporating a broader range of countries and additional student-level covariates, such as socio-economic status or language background, to further examine the robustness of the identified DIF patterns. Second, the exploratory nature of Rasch Tree models implies that the detected splits are data-driven and should be subjected to confirmatory validation using independent samples or alternative modeling approaches. Furthermore, longitudinal comparisons across multiple PISA cycles (e.g., 2018 and 2022) could provide valuable insights into the temporal stability of DIF patterns and the evolution of measurement invariance over time.

Acknowledgments. The author acknowledges the use of artificial intelligence–based language tools for proofreading and linguistic refinement during the preparation of this manuscript. These tools were not used for data analysis, interpretation of results or substantive content generation. All statistical analyses and psychometric modeling were conducted using the R programming language within the RStudio integrated development environment.

Disclosure of Interests. The author declares that there are no competing interests relevant to the content of this article.

References

1. Bond, T.G., Fox, C.M.: Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd edn. Routledge, New York (2013)
2. Engelhard Jr., G., Wang, J.: Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences. 2nd edn. Routledge, New York (2024)
3. Holland, P.W., Wainer, H.: Differential Item Functioning. Routledge, New York (2012)
4. Komboz, B., Strobl, C., Zeileis, A.: Tree-based global model tests for polytomous Rasch models. Educational and Psychological Measurement **78**(1), 128–166 (2018)

5. Mislevy, R.J., Wu, P.K.: Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. ETS Res. Rep. Ser. **1996**(2), i–36 (1996)
6. OECD: PISA 2022 Database. https://www.oecd.org/en/data/datasets/pisa-2022-database.html, last accessed 2026/01/17
7. Osterlind, S.J., Everson, H.T.: Differential Item Functioning. 2nd edn. Sage, Thousand Oaks (2009)
8. Rasch, G.: Probabilistic Models for Some Intelligence and Attainment Tests. Danish Institute for Educational Research, Copenhagen (1960)
9. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. Wiley, New York (1987)
10. Rutkowski, L., Rutkowski, D.: A call for a more measured approach to reporting and interpreting PISA results. Educational Researcher **45**(4), 252–257 (2016)
11. Rutkowski, L., Svetina, D.: Assessing the hypothesis of measurement invariance in the Programme for International Student Assessment. Educational and Psychological Measurement **74**(5), 814–835 (2014)
12. Strobl, C., Kopf, J., Zeileis, A.: Rasch trees: A new method for detecting differential item functioning. Psychometrika **80**(2), 289–316 (2015)
13. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in R. Journal of Statistical Software **45**(3), 1–67 (2011)
14. Verhelst, N.D., Glas, C.A.W.: The one parameter logistic model. In: Fischer, G.H., Molenaar, I.W. (eds.) Rasch Models: Foundations, Recent Developments, and Applications, pp. 215–237. Springer, New York (1995)
15. Zeileis, A., Hothorn, T., Hornik, K.: Model-based recursive partitioning. Journal of Computational and Graphical Statistics **17**(2), 492–514 (2008)
16. Zumbo, B.D.: A Handbook on the Theory and Methods of Differential Item Functioning (DIF). National Defense Headquarters, Ottawa (1999)
17. Zumbo, B.D.: 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 (2007)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

