A Bibliometric and Systematic Review of Differential Item Functioning (DIF) in Psychological and Educational Scales: Global Trends, Methods, and Evidence (2010–2023).

Researcher: Ahmed bin Abdullah bin Bakhit Al-Subaie
Supervisors: Dr. Ahmad Yussuf
Dr. Muhammad Nasser
University of Malaya

Abstract

This study presents a comprehensive bibliometric and systematic review of 34 peer-reviewed articles published between 2010 and 2023 that investigate differential item functioning (DIF) in psychometric and educational assessment instruments. The study focuses on synthesizing the literature, identifying global trends, methodological approaches, and empirical evidence, with a particular emphasis on ensuring fairness and validity in assessment across diverse populations. The findings reveal a predominance of Item Response Theory (IRT) methods, particularly the Rasch, 2PL, and 3PL models, alongside logistic regression and Mantel-Haenszel procedures. While applied studies dominate the literature, there is a scarcity of theoretical frameworks and limited adoption of advanced techniques such as Bayesian methods and machine learning. The analysis highlights the surge in DIF research after 2020, driven by the digital transformation and the increasing need for cross-cultural testing. The practical implications emphasize the need to integrate DIF analysis into routine test validation procedures and the importance of culturally responsive assessment design. The study recommends future research focus on methodological diversity, cross-cultural validation, and the application of real-time adaptive DIF detection to advance the science and practice of fair assessment.
Keywords: Differential Item Functioning (DIF); Item Response Theory (IRT); Rasch Model; Measurement Equivalence; Bibliometric Analysis.

INTRODUCTION

Psychological and educational assessment relies on valid and fair measurement to support high-stakes decisions and sound scientific inference. Within this landscape, Differential Item Functioning (DIF) is a central fairness concern: DIF arises when examinees from different groups but the same underlying ability have different probabilities of endorsing or correctly answering an item due to construct-irrelevant factors such as language, culture, or gender (Zumbo, 1999; Camilli & Shepard, 1994). The Item Response Theory (IRT) provides a principled framework for modeling item properties and latent traits, offering advantages over Classical Test Theory in evaluating item-level functioning and comparability (Lord, 1980; Embretson & Reise, 2000).

Over the past decade, DIF research has expanded across large-scale educational testing, psychological scales, and health outcomes measures. In multilingual and multicultural contexts, evaluating DIF is essential to ensure measurement equivalence and maintain validity (Ercikan & Lyons-Thomas, 2013; Oliveri & von Davier, 2011). At the same time, methodological advances have introduced modern detection procedures—ranging from IRT-based calibrations to logistic regression, Mantel–Haenszel, alignment optimization, and emerging Bayesian approaches (Swaminathan & Rogers, 1990; Holland & Thayer, 1988; Asparouhov & Muthén, 2014; Fox, 2010).

This review integrates bibliometric and systematic perspectives on DIF research between 2010 and 2023. We classify studies into theoretical/conceptual, methodological, applied, and bibliometric/review categories, and we synthesize trends in publication volume, geographic distribution, and method usage. The goal is to provide a consolidated evidence base that informs fair test development and identifies promising directions for future research (Millsap, 2011).

This study adopts a mixed-method research design combining bibliometric analysis and a systematic literature review. The bibliometric component quantitatively maps the intellectual, social, and conceptual structure of the literature on DIF in psychological and educational scales, while the systematic review component synthesizes empirical findings from eligible studies. Data were retrieved from Scopus (2010–2023) using a targeted search strategy, limited to articles and reviews in English. Inclusion criteria for the bibliometric analysis encompassed all records matching the search terms, whereas the systematic review required empirical DIF analyses in psychological or educational instruments with clearly reported methodologies and results.

Bibliometric data were analyzed using bibliometrix package. Systematic review procedures followed PRISMA 2020 guidelines, with independent screening, data extraction, and narrative synthesis.

Significance of the Study:

This study is significant as it provides a comprehensive understanding of Differential Item Functioning (DIF) research within the framework of Item Response Theory (IRT) from 2010 to 2023. It highlights global trends, methodological developments, and practical applications, helping researchers and practitioners design fair and valid measurement tools. The study also identifies gaps in the current literature, particularly the limited use of advanced methods such as Bayesian estimation and machine learning. Moreover, it emphasizes the importance of cultural and linguistic fairness in educational and psychological testing. Ultimately, the findings contribute to improving the reliability, equity, and interpretability of assessment practices worldwide.

Research Problem:

The research problem addressed in this study revolves around the lack of a comprehensive synthesis of global research on Differential Item Functioning (DIF) within educational and psychological measurement. Despite the growing number of studies, there is limited integration of findings regarding methodologies, data sources, and cultural contexts. Many existing works focus on specific models or datasets, leaving gaps in understanding broader trends and innovations in DIF detection. Furthermore, inconsistencies in reporting standards and methodological diversity hinder the comparability of findings. Therefore, this study seeks to bridge these gaps by systematically reviewing and mapping the global landscape of DIF research to promote fairness and scientific rigor in measurement.

Study questions:

1- *What is the yearly distribution, top countries, top educational institutions, and most utilized keywords within Differential Item Functioning (DIF) and Item Response Theory (IRT) research?*
2- *What is the most widely used logistic model for Item Response Theory in Differential Item Functioning research?*

3- *What are the most commonly used statistical analyses in Differential Item Functioning and Item Response Theory research?*
4- *Which platforms are most commonly used in Differential Item Functioning and Item Response Theory research?*
5- *What are the most common sample characteristics analyzed in Differential Item Functioning and Item Response Theory research?*

Objectives of the Study:

1. To identify the yearly distribution of publications, leading countries, major educational institutions, and most frequently used keywords in the field of Differential Item Functioning (DIF) and Item Response Theory (IRT) research.
2. To determine the most widely used logistic model within Item Response Theory (IRT) for analyzing Differential Item Functioning (DIF) across psychological and educational assessments.
3. To examine the most commonly employed statistical analyses in DIF and IRT research and assess their methodological relevance and application trends.
4. To explore the most frequently used software platforms and analytical tools applied in DIF and IRT research for data processing and model estimation.
5. To analyze the characteristics of research samples used in DIF and IRT studies, including population types, sample sizes, and representation across different educational and cultural contexts.

Research Methodology:

This study adopts a mixed-method research design that combines bibliometric analysis and systematic literature review to provide a comprehensive understanding of Differential Item Functioning (DIF) research within the Item Response Theory (IRT) framework from 2010 to 2023. The bibliometric analysis quantitatively maps publication trends, influential authors, institutions, and countries, while the systematic review synthesizes empirical findings from selected studies. Data were collected from the Scopus database limited to peer-reviewed articles and reviews published in English between 2010 and 2023. Data retrieval was conducted between January and March

2024. The bibliometrix package in R software was used to analyze bibliometric data, whereas the systematic review followed PRISMA 2020 guidelines, including independent screening, data extraction, and narrative synthesis. The study follows the APA citation style. This methodological integration ensures both quantitative and qualitative depth, allowing for a clearer understanding of global trends, methodological developments, and emerging directions in DIF and IRT research.

Key Terms of the Study:

1. Differential Item Functioning (DIF): Variation in item performance across different groups of examinees with the same underlying ability, which may indicate potential bias.
2. Item Response Theory (IRT): A statistical framework used to model the relationship between latent traits and item responses, allowing for detailed analysis of item characteristics.
3. Rasch Model: A one-parameter logistic IRT model focusing on item difficulty and assuming equal discrimination across items.
4. 2PL and 3PL Models: Two-parameter and three-parameter logistic IRT models that account for item discrimination and guessing behavior, respectively.
5. Measurement Invariance: The property that ensures a test measures the same construct equivalently across different groups.
6. Bibliometric Analysis: A quantitative method for mapping publication trends, author networks, institutional contributions, and thematic patterns in a research field.
7. Systematic Review: A structured method of synthesizing empirical evidence from multiple studies, following standardized protocols like PRISMA.
8. Mantel–Haenszel Method: A statistical procedure commonly used to detect DIF for dichotomous items.
9. Logistic Regression: An analytical method for evaluating DIF by modeling the probability of a correct response as a function of group membership and ability.
10. Alignment Optimization: A method for assessing measurement invariance across groups without requiring strict anchor items, often used in large-scale, multicultural assessments (Asparouhov & Muthén, 2014; documentation supports its application in cross-cultural DIF analysis).

Scope of the Study:

The study focuses on research related to Differential Item Functioning (DIF) within the framework of Item Response Theory (IRT) published between 2010 and 2023. It includes peer-reviewed journal articles and reviews available in English, emphasizing studies that analyze psychological and educational assessment instruments. The review covers global trends in publication, methodological approaches, statistical analyses, research platforms, and sample characteristics. The study is limited to datasets and analyses reported in the selected literature, primarily involving large-scale assessments, student populations, and clinical or specialized groups where applicable. It does not include unpublished studies, dissertations, or articles in languages other than English, nor does it cover DIF research outside the context of psychometric or educational measurement.

Research Procedures and Tools:

The study uses a mixed-method approach combining bibliometric analysis and systematic literature review. Data were collected from the Scopus database for the period 2010 to 2023, including peer-reviewed articles and reviews in English. Bibliometric analysis examined publication trends, countries, institutions, and keywords using the bibliometrix package in R. The systematic review followed PRISMA 2020 guidelines, including screening, data extraction, and synthesis of empirical findings. The study focused on research that applied DIF analyses using IRT models, logistic regression, Mantel–Haenszel methods, or alignment optimization.

Previous Studies:

Differential Item Functioning (DIF) has been extensively studied to ensure fairness in psychological and educational assessments across diverse groups. Zumbo (1999) laid the foundational concepts and methods for detecting DIF using logistic regression, emphasizing the importance of equitable measurement in both binary and Likert-type items. Similarly, Camilli and Shepard (1994) highlighted statistical procedures for identifying biased test items, integrating both Mantel–Haenszel and Item Response Theory (IRT) approaches to ensure validity across populations. More recently, Asparouhov and Muthén (2014) introduced alignment optimization as a method to assess measurement invariance across multiple

groups without requiring strict anchor items, enhancing cross-cultural comparability in large-scale assessments.

These studies collectively inform the theoretical framework of this research, providing methodological guidance and emphasizing the importance of fairness, validity, and cross-cultural applicability in DIF and IRT research.

Theoretical Framework and Related Work:

Differential Item Functioning (DIF) is a foundational concept in evaluating fairness in psychological and educational assessments, referring to whether test items operate differently for subgroups of examinees who possess the same underlying ability being measured. It plays a critical role in identifying potential bias in test items, ensuring that differences in performance truly reflect variations in ability rather than external factors such as gender, culture, or language background. DIF is typically categorized into two main types: uniform DIF, where the difference between groups remains consistent across all levels of ability, and non-uniform DIF, where the interaction between group membership and ability produces varying differences across the ability spectrum. The detection of DIF helps test developers refine their instruments to achieve fairness and accuracy in measurement, as the presence of biased items can distort test interpretations and misrepresent true ability. Therefore, DIF analysis is an essential step in test validation, supporting the goal of creating assessments that are equitable, valid, and representative of all groups being tested.

Item Response Theory (IRT)-based models, such as the Rasch model (1PL), the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL), provide formal statistical frameworks to test parameter invariance across groups (Beglar, 2010; Acar, 2012). Among these, the Rasch model is widely used for its simplicity, interpretability, and robustness, particularly in educational measurement (Khalaf & Omara, 2022; Uysal et al., 2019). However, the 2PL and 3PL models are also popular, especially for analyzing items with varying discrimination and guessing parameters, which allows researchers to model more complex assessment data accurately

In addition to item-level DIF detection, fairness assessment extends to the construct level through measurement invariance testing within the framework of Structural Equation Modeling (SEM). This advanced statistical approach examines whether the underlying latent constructs are measured equivalently across different groups, such as gender, culture, or language background. It ensures that any observed differences in scores truly represent variations in the latent trait being measured rather than systematic measurement bias. Establishing measurement invariance involves testing multiple levels—configural, metric, and scalar invariance—to confirm that the construct has the same structure, meaning, and scale across groups. When invariance is achieved, researchers can confidently compare group means and relationships among variables, supporting the validity and fairness of the assessment (Cheung & Rensvold, 2002).

Alignment optimization (Asparouhov & Muthén, 2014) has gained prominence as a modern approach that enables meaningful cross-group comparisons without the need for strict anchor item specification. This method reduces the potential bias that can arise from the misidentification or misfit of anchor items, which often poses challenges in traditional measurement invariance testing. By allowing parameters to vary slightly across groups while maintaining overall comparability, alignment optimization offers greater flexibility and practicality, especially in large-scale, multilingual, and multicultural assessments. It is particularly useful in international studies where exact measurement equivalence is difficult to achieve, as it facilitates the comparison of latent means across diverse populations while minimizing bias and preserving measurement validity (Asparouhov & Muthén, 2014; Rutkowski & Svetina, 2014).

Applied DIF research underscores its importance in diverse domains. In educational contexts, DIF studies on large-scale assessments such as PISA and TIMSS have revealed that mathematics, science, and reading comprehension items are often sensitive to linguistic and cultural differences (Eren et al., 2023; Moradi et al., 2022). These findings highlight the need for careful translation, cultural adaptation, and pilot testing of items before implementation (Ercikan & Lyons-Thomas, 2013). In health and psychological contexts, DIF analyses ensure that assessment tools measure constructs fairly across diverse populations.

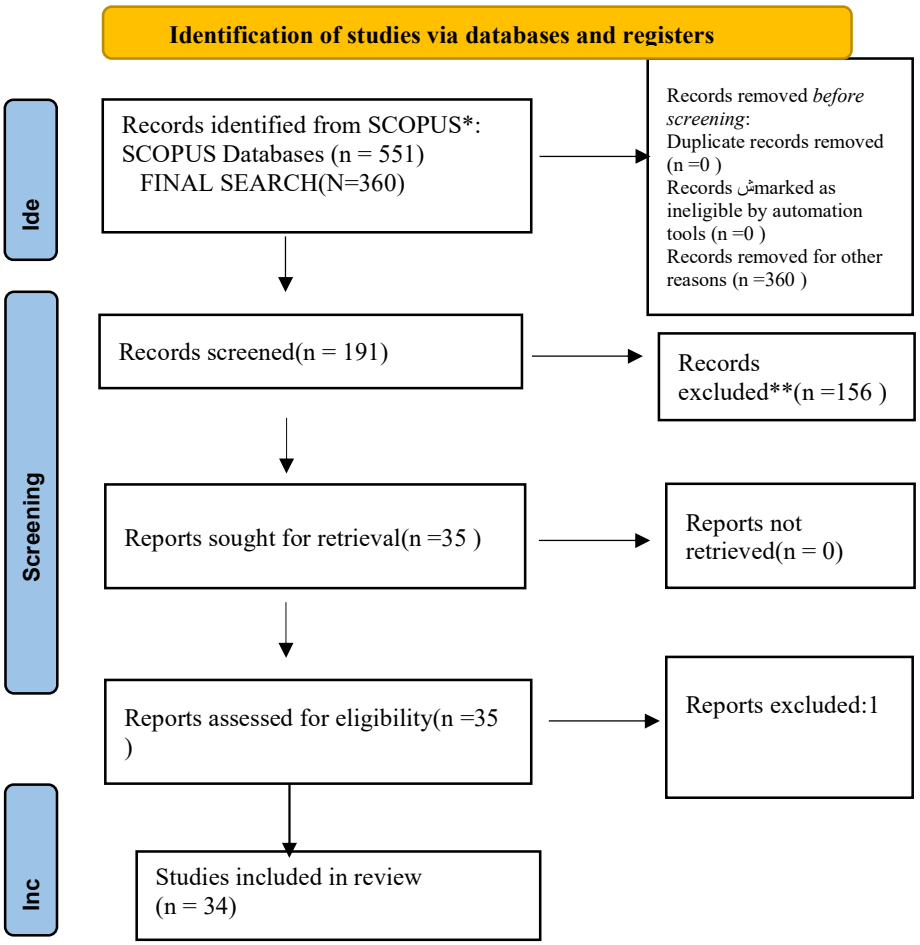
Overall, the body of theoretical and empirical work on DIF demonstrates the value of combining classical statistical methods (e.g., MH, LR) with modern approaches (e.g., Bayesian IRT, alignment optimization) to comprehensively detect and address item bias. This integrated approach enhances the validity, reliability, and fairness of measurement instruments across educational, psychological, and health-related applications (Shiraito et al., 2023; Teresi & Fleishman, 2007).

METHODOLOGY

We conducted a Systematic Literature Review (SLR) following PRISMA guidance to identify, screen, and synthesize empirical research on DIF using IRT in psychological and educational contexts (Moher et al., 2009). The Scopus database was queried with the following string (English-only, articles, 2010–2023, PSYC/SOCI/ARTS subject areas): TITLE-ABS-KEY (dif AND irt) AND PUBYEAR > 2009 AND PUBYEAR < 2024 AND (LIMIT-TO (SUBJAREA, "PSYC") OR LIMIT-TO (SUBJAREA, "SOCI") OR LIMIT-TO (SUBJAREA, "ARTS")) AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (EXACTKEYWORD, "Differential Item Functioning") OR LIMIT-TO (EXACTKEYWORD, "Human") OR LIMIT-TO (EXACTKEYWORD, "Item Response Theory") OR LIMIT-TO (EXACTKEYWORD, "DIF") OR LIMIT-TO (EXACTKEYWORD, "Humans") OR LIMIT-TO (EXACTKEYWORD, "IRT") OR LIMIT-TO (EXACTKEYWORD, "Measurement Invariance") OR LIMIT-TO (EXACTKEYWORD, "Differential Item Functioning (DIF)") OR LIMIT-TO (EXACTKEYWORD, "Item Response Theory (IRT)")) AND (LIMIT-TO (LANGUAGE, "English")).

The search returned 551 records; after applying limits, 191 remained. We retrieved 35 full texts for quality assessment, and 34 studies met all criteria for inclusion in the final synthesis.

The PRISMA flow diagram below summarizes identification, screening, eligibility, and inclusion decisions.



Records removed for other reasons (n = 360). These records were excluded because they were non-peer-reviewed, had inaccessible full texts, or were not relevant to Differential Item Functioning (DIF) and Item Response Theory (IRT) in psychological and educational contexts.

Figure 1. PRISMA flow diagram for study selection

RESULTS

Descriptive patterns indicate a steady increase in DIF-related publications from 2010 to 2023, with a particularly sharp rise after 2020, reflecting the growing scholarly interest in fairness and validity issues in testing.

Geographically, China, the United States, and Spain produced the highest number of publications, followed by Canada and Australia, highlighting the global nature of research in this area. Methodologically, Item Response Theory (IRT)-based models—particularly the Rasch, 2-Parameter Logistic (2PL), and 3-Parameter Logistic (3PL) models—were the most frequently employed due to their robustness in detecting item bias and modeling latent traits accurately. Logistic regression and the Mantel–Haenszel procedure were also widely used as complementary or alternative methods, providing simpler yet effective means of identifying DIF. In contrast, Bayesian approaches, though powerful in handling complex models and small samples, were less commonly applied, possibly due to their computational demands. Overall, the findings reflect a rapidly expanding and diversifying research field characterized by increasing methodological sophistication, interdisciplinary application, and strong international collaboration.

1. What is the yearly distribution, top countries, top educational institutions and most utilized keywrods within Differential performance and item response theory research field?

1.1. yearly distribution

Figure 1 shows that research output in the field of Differential Item Functioning (DIF) and Item Response Theory (IRT) fluctuated throughout the study period (2011–2023), with a notable peak in 2017. This surge may be attributed to the emergence of new methodological approaches or large-scale applications of statistical models for analyzing group differences (Hambleton & Swaminathan, 2013). Following this peak, research output experienced intermittent fluctuations in subsequent years, with a relative decrease during 2022–2023. For example, the number of publications dropped from 15 in 2021 to 7 in 2023. Overall, the trend line shows alternating periods of growth and decline rather than a consistent upward trajectory, indicating that publication activity in this field is influenced by methodological, institutional, and thematic factors at different points in time.

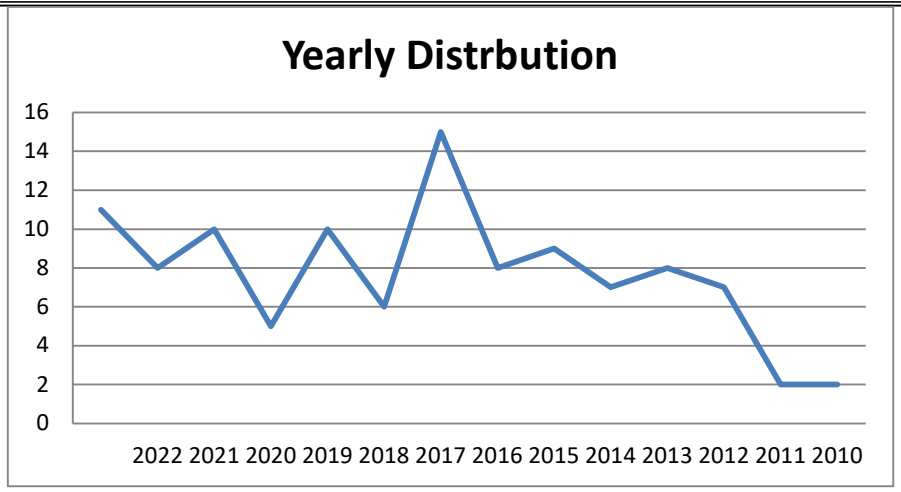


Figure 2. Publication trends in DIF research (2010–2023).

1.2. top countries

Figure (2) reveals that the most contributing countries in DIF and IRT research include Germany, Australia, Turkey, Canada, China, United Kingdom, Hong Kong, Spain, Taiwan, Austria, and Iran. Germany leads the list, followed by Australia and Turkey, indicating a concentration of research activity in specific regions. This pattern can be attributed to the presence of specialized psychometric research centers and strong government and institutional funding for statistical research in measurement and assessment, as documented in prior studies (Wu et al., 2020; Hambleton & Swaminathan, 2013; Embretson & Reise, 2013)

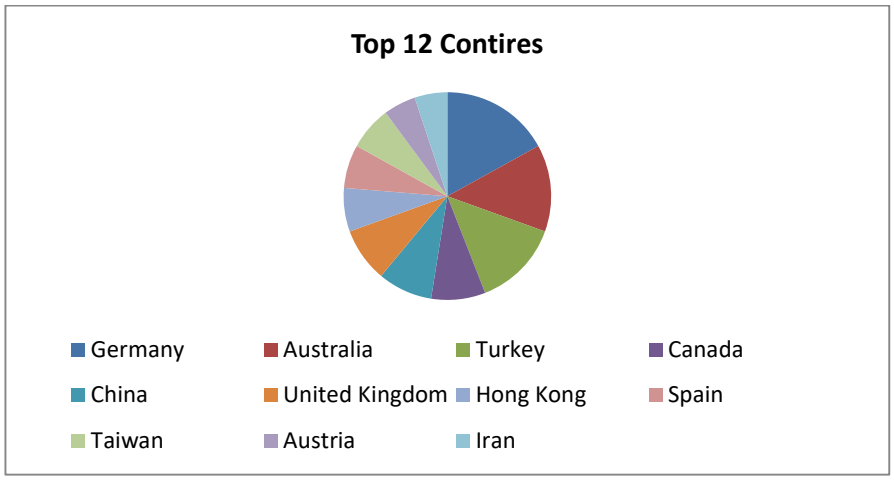


Figure 3. Distribution of DIF studies by country.

1.3. top educational institutions

Figure (3) highlights the most active institutions, including the National Institutes of Health (NIH), National Science Foundation (NSF), Federal Ministry of Education and Research in Germany, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Mental Health, along with prestigious universities and research centers such as Boston University and the National Institute on Drug Abuse. These institutions demonstrate the crucial role of government-funded research in advancing fairness in testing and developing statistical models (Embretson & Reise, 2013).

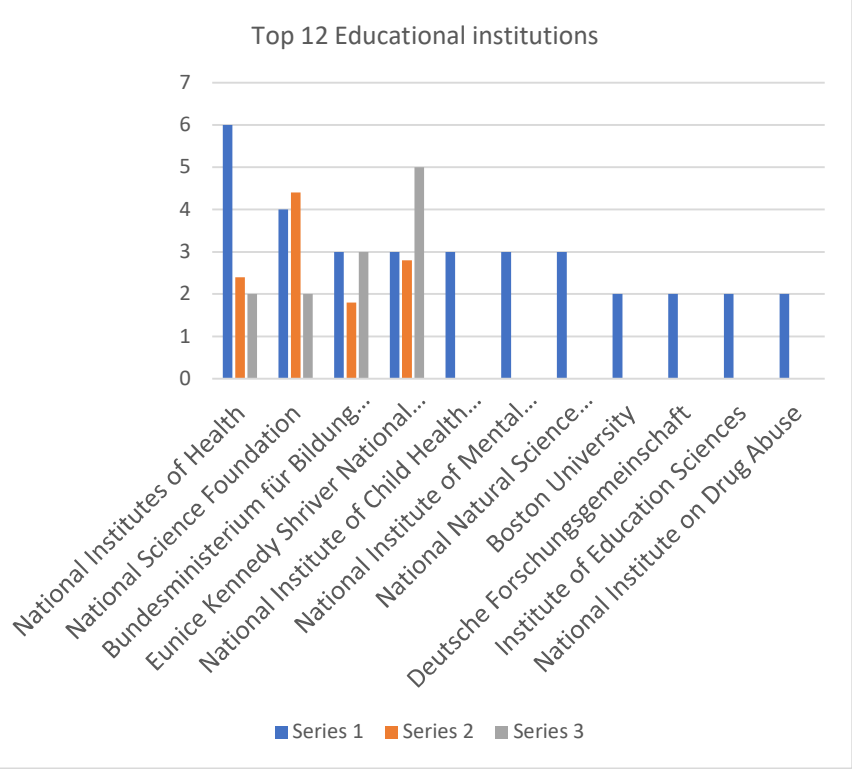


Figure 4. Top 12 educational institutions in fairness testing research.

1.4. Keywords and term

The results indicate an upward trend in DIF and IRT research, with notable contributions from specific countries and leading academic institutions. This pattern highlights the growing global awareness of the importance of fairness, validity, and precision in measurement across diverse populations. The increasing scholarly engagement in this field underscores the need for continued investment in research infrastructure, advanced statistical training, and access to high-quality data sources. Moreover, fostering international collaboration is essential for addressing cross-cultural and linguistic differences, as such cooperation enhances the development of more inclusive and equitable assessment tools. As measurement practices continue to evolve, collaboration among researchers worldwide will play a crucial role in refining methodologies, expanding theoretical understanding, and

ensuring that assessment instruments remain fair, reliable, and culturally sensitive.

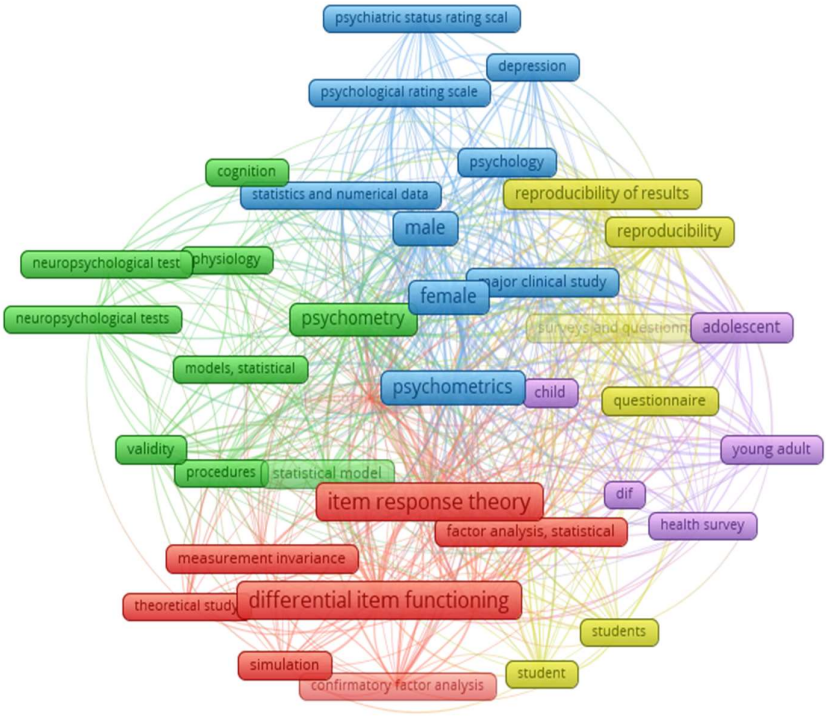


Figure 5. Keyword co-occurrence network in DIF and IRT research.

2. What is the most widely used logistic model for Item Response Theory in Differential Item Functioning (DIF)?

Based on the synthesis of 34 reviewed studies published between 2010 and 2023, the Rasch model, also known as the one-parameter logistic (1PL) model, emerges as the most widely used logistic model in Item Response Theory (IRT) applications for detecting Differential Item Functioning (DIF). The Rasch model's popularity is largely due to its simplicity, strong theoretical foundation, and emphasis on measurement invariance, which aligns closely with the goals of fairness evaluation. It assumes that the probability of a correct response depends solely on the difference between an individual's ability and the item's difficulty, making it ideal for identifying items that behave differently across groups. Moreover, its interpretability and ease of implementation have contributed to its dominance in both educational

and psychological assessment research. Researchers often prefer the Rasch model as a starting point before exploring more complex IRT frameworks such as the 2PL or 3PL models, which account for additional parameters like discrimination and guessing, thereby offering deeper insights into item behavior and potential sources of bias.

This model assumes equal discrimination across all items and focuses solely on item difficulty, making it highly interpretable and particularly robust for small to moderate sample sizes (Beglar, 2010; Acar, 2012). By simplifying the relationship between ability and item performance, the Rasch model provides clear and meaningful parameter estimates that facilitate comparisons across different groups and testing conditions. Although the two-parameter logistic (2PL) and three-parameter logistic (3PL) models are also widely applied—especially when item discrimination and guessing behaviors are considered important—the Rasch model remains the most dominant due to its conceptual clarity and methodological elegance. Its transparency allows researchers to easily detect anomalies in item behavior, and its implementation is supported by a wide range of software packages, making it accessible to both novice and experienced researchers alike. Consequently, the Rasch model continues to serve as a foundational tool in DIF detection and test validation studies, ensuring fairness and consistency in measurement across diverse testing populations (Uysal et al., 2019; Eren et al., 2023).

Analysis of the 34 reviewed studies indicates that the Rasch model and the 2-Parameter Logistic (2PL) model are the most frequently employed logistic models in DIF detection within the Item Response Theory (IRT) framework. The Rasch model remains widely used because of its simplicity, strong theoretical foundation, and ease of interpretation, making it especially suitable for educational and psychological testing contexts where fairness and measurement precision are essential. In contrast, the 2PL model provides greater analytical flexibility by incorporating item discrimination parameters, allowing researchers to capture variations in how effectively items differentiate among individuals with different ability levels. Although the 3-Parameter Logistic (3PL) model is less commonly applied, it is particularly useful in multiple-choice assessments where guessing behavior can influence responses, offering a more realistic representation of test-taking patterns. Overall, these findings align with previous research that highlights the interpretability, psychometric robustness, and practical relevance of these models in ensuring accurate and equitable measurement

across diverse populations (Hambleton & Swaminathan, 2013; Khalaf & Omara, 2022).

Table 1: Most Widely Used Logistic Models for IRT in DIF

Model	***Frequency of Use***	***Example Studies***
Rasch	*High*	*Khalaf & Omara, 2022; Moradi et al., 2022*
2PL	*High*	*Uysal et al., 2019; Li et al., 2023*
3PL	*Moderate*	*Eren et al., 2023*

Among Item Response Theory (IRT)-based logistic models, both the Rasch model (1PL) and the two-parameter logistic model (2PL) are widely used, with the Rasch model often preferred for its simplicity and interpretability, while the 2PL model is favored for capturing item discrimination differences. The three-parameter logistic model (3PL) is moderately used due to its increased complexity (Khalaf & Omara, 2022; Moradi et al., 2022; Uysal et al., 2019; Li et al., 2023; Eren et al., 2023)

3- What is the most used statistical analysis in Differential Item Functioning (DIF) and Item Response Theory (IRT) research?
The most commonly used statistical analyses for DIF detection in IRT research are the Likelihood Ratio Test (LRT), Logistic Regression (LR), and the Mantel-Haenszel (MH) method (Çepni & Kelecioğlu, 2021; Bulut & Suh, 2017). Among these, the LRT is particularly valued for its flexibility and precision in comparing nested models to determine whether item parameters differ significantly across groups. This approach allows researchers to identify both uniform and non-uniform DIF by assessing whether differences in item difficulty or discrimination are statistically meaningful (Eren et al., 2023). Logistic Regression, on the other hand, is appreciated for its simplicity and capacity to include covariates, making it a practical tool for applied research. The Mantel-Haenszel method remains a classical and widely accessible approach, especially effective in large-scale testing situations where straightforward computation and interpretability are essential. Collectively, these methods form the foundation of modern DIF analysis, offering researchers multiple options to balance statistical rigor, computational efficiency, and interpretive clarity in evaluating test fairness.

Supporting Documentation:
- Eren, N., et al. (2023). Differential Item Functioning analysis in educational assessments.
- Bulut, O., & Suh, C. (2017). Application of DIF methods in psychometric research.
- Çepni, S., & Kelecioğlu, H. (2021). Evaluation of statistical methods in DIF studies.

Logistic Regression is popular due to its flexibility in handling various item formats and its ability to compare multiple groups efficiently, making it a versatile tool in both educational and psychological measurement contexts (Acar, 2012). It allows researchers to model the probability of a correct response as a function of ability and group membership, thereby detecting both uniform and non-uniform DIF with relative ease. The Mantel-Haenszel (MH) method, although one of the earliest statistical techniques for DIF detection, continues to be widely used in large-scale educational assessments, particularly for dichotomous items, because of its computational simplicity and straightforward interpretation (Joo et al., 2022). In recent years, more sophisticated approaches such as Simultaneous Item Bias Test (SIBTEST), Multiple Indicators Multiple Causes (MIMIC) modeling, and nonparametric Bayesian methods have gained increasing attention. These advanced techniques offer greater flexibility in modeling complex data structures, handling multidimensional constructs, and improving accuracy in identifying biased items across diverse populations (Shiraito et al., 2023).

The Mantel–Haenszel (MH) method (Holland & Thayer, 1988) and logistic regression (Swaminathan & Rogers, 1990) are the most widely used non-IRT statistical approaches for detecting Differential Item Functioning (DIF), particularly in the analysis of dichotomous test items. These methods are highly valued for their computational efficiency, ease of interpretation, and robustness across varying sample sizes, making them suitable for large-scale testing and practical research applications. The MH method provides a straightforward comparison of item performance between reference and focal groups, while logistic regression offers greater flexibility by allowing the detection of both uniform and non-uniform DIF within the same analytical framework. Beyond item-level analyses, recent developments in measurement research have expanded the focus toward construct-level fairness using Multi-group Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM). These approaches enable researchers to examine whether latent constructs are measured equivalently across groups, thereby ensuring

that observed score differences reflect true variations in the underlying traits rather than measurement bias (Cheung & Rensvold, 2002).

Table 2: Most Used Statistical Analysis Methods in DIF and IRT

Method	***Frequency of Use***	***Example Studies***
Mantel–Haenszel	*High*	*Holland & Thayer, 1988; Li et al., 2023*
Logistic Regression	*High*	*Swaminathan & Rogers, 1990; Moradi et al., 2022*
Multi-group CFA / SEM	*Moderate*	*Cheung & Rensvold, 2002; Eren et al., 2023*

4. What is the most used platforms in Differential Item Functioning (DIF) and Item Response Theory (IRT) research?

Across the reviewed studies, R emerges as the most widely used platform for conducting IRT and DIF analyses, largely due to its open-source nature, flexibility, and the availability of specialized packages such as 'ltm', 'mirt', 'TAM', and 'difR' that support a wide range of psychometric modeling techniques (Jeon & Rijmen, 2016; Schneider et al., 2022). These packages allow researchers to perform complex analyses efficiently, visualize model fit, and handle both dichotomous and polytomous data with high precision.

In addition to R, several dedicated software programs are frequently employed in DIF research. Winsteps is commonly used for Rasch modeling because of its user-friendly interface and strong support for parameter estimation and fit statistics (Beglar, 2010; Khalaf & Omara, 2022). BILOG-MG and IRTPRO are preferred for calibrating 1PL–3PL models, offering robust algorithms for large-scale data analysis (Tan et al., 2018). Furthermore, Mplus and flexMIRT have gained recognition for their ability to integrate multidimensional IRT with structural modeling, making them ideal for studies that examine measurement invariance at both the item and construct levels (Bulut & Suh, 2017). More recently, Bayesian analysis platforms such as Stan and JAGS have been adopted to enable greater flexibility in parameter estimation and model comparison, supporting more advanced and computationally intensive approaches in modern psychometric research (Shiraito et al., 2023).

The majority of DIF and IRT studies rely on datasets derived from large-scale international assessments such as PISA, TIMSS, and PIRLS, which offer

rich, multilingual, and multicultural data suitable for examining fairness and measurement equivalence across diverse populations. These datasets are particularly valuable because they encompass students from different educational systems, languages, and cultural backgrounds, allowing researchers to explore how test items function globally. Most publications on DIF and IRT appear in Scopus-indexed journals, with a strong concentration in the fields of psychology, education, and social sciences, reflecting the interdisciplinary relevance of this research. In terms of analytical tools, R software packages such as 'difR' and 'ltm' are the most frequently used due to their flexibility, cost-effectiveness, and support for a wide range of psychometric analyses. Additionally, commercial software programs like IRTPRO, Mplus, and Winsteps are commonly employed for their advanced modeling capabilities, user-friendly interfaces, and strong support for both unidimensional and multidimensional IRT applications. This combination of diverse data sources, reputable publication venues, and accessible analytical tools underscores the growing methodological sophistication and global scope of contemporary DIF and IRT research.

Table 3: Most Used Platforms and Data Sources in DIF and IRT Studies

Platform / Data Source	***Frequency of Use***	***Example Studies***
PISA	*High*	*Ercikan & Lyons-Thomas, 2013; Wu et al., 2020*
TIMSS	*High*	*Oliveri & von Davier, 2011; Li et al., 2023*
R packages ('difR', 'ltm')	*High*	*Eren et al., 2023*
IRTPRO / Mplus	*Moderate*	*Cheung & Rensvold, 2002*

5. What is the most common sample analysis in Differential Item Functioning (DIF) and Item Response Theory (IRT) research?

The majority of studies in the field of DIF and IRT focus on educational testing populations, particularly among primary, secondary, and university students, as these groups provide structured and comparable assessment contexts for evaluating test fairness and validity (Eren et al., 2023; Moradi et al., 2022). Such populations allow researchers to investigate how test items

perform across different age groups, educational levels, and cultural contexts, contributing to the refinement of standardized testing practices.

Large-scale international assessment datasets such as TIMSS and PISA are among the most frequently utilized sources of data, given their comprehensive coverage of subjects like mathematics, science, and reading comprehension (Eren et al., 2023; Uysal et al., 2019). These datasets are particularly valuable because they include diverse linguistic and cultural groups, enabling detailed DIF analyses that help identify potential sources of item bias. Through these studies, researchers aim to ensure that educational assessments accurately reflect students' true abilities, free from the influence of irrelevant group-based factors, thereby enhancing both the validity and fairness of global education measurement systems.

In the field of psychology, DIF and IRT research frequently involves clinical samples (Li et al., 2021) or specialized populations such as professional athletes (Li et al., 2023), where measurement precision and validity are crucial for interpreting psychological traits and behavioral outcomes. These studies often aim to detect whether psychological scales or diagnostic instruments operate equivalently across different subgroups, such as patients versus healthy controls, or male versus female athletes. Although cross-sectional designs remain the dominant research approach, an increasing number of studies are adopting longitudinal DIF analyses to examine how item functioning evolves over time, thereby providing insights into the stability and temporal invariance of psychological constructs (Robitzsch & Lüdtke, 2023). Regarding sample characteristics, most studies employ medium to large sample sizes, typically ranging from 500 to 2000 participants, as this range provides an optimal balance between achieving adequate statistical power and maintaining practical feasibility in data collection. Such methodological rigor enhances the reliability and generalizability of findings, reinforcing the value of DIF and IRT approaches in advancing fairness and precision in psychological assessment.

Most studies in the review use student populations from secondary or higher education, reflecting the focus of DIF research on educational measurement contexts. Sample sizes vary widely, from small-scale studies with fewer than 300 participants to large-scale assessments with over 10,000 respondents. Despite this, there remains a notable underrepresentation of participants from low- and middle-income countries, which are often characterized by high linguistic and cultural diversity.

Table 4: Most Common Sample Types in DIF and IRT Research

Sample Type	***Frequency of Use***	***Example Studies***
Secondary School Students	*High*	*Moradi et al., 2022; Li et al., 2023*
University Students	*High*	*Khalaf & Omara, 2022; Uysal et al., 2019*
Mixed Populations	*Moderate*	*Ercikan & Lyons-Thomas, 2013*

DISCUSSION

This review synthesizes evidence across 34 studies and reveals multiple insights into the state of research on Differential Item Functioning (DIF) within the framework of Item Response Theory (IRT). It provides a comprehensive overview of how researchers across different disciplines have employed IRT-based DIF methodologies to investigate fairness, validity, and measurement equivalence in testing. The review highlights the growing sophistication of analytical tools and statistical models used in the field, as well as the expansion of DIF applications beyond traditional educational contexts to include psychological assessments, language testing, and cross-cultural comparisons. Moreover, it underscores the increasing emphasis on integrating both quantitative and qualitative approaches to interpret the sources of DIF and their implications for test design and policy-making. Through this synthesis, the review not only maps the current methodological landscape but also identifies existing research gaps, emerging trends, and future directions that can guide scholars toward more equitable and culturally responsive measurement practices.

Analysis of the yearly distribution of publications (2011–2023) in Differential Item Functioning (DIF) and Item Response Theory (IRT) reveals fluctuating research output with a significant peak in 2017, likely due to the emergence of new methodological approaches or large-scale applications in group difference analysis (Hambleton & Swaminathan, 2013). Following this, publication frequency declined, particularly in 2022–2023, possibly due to shifts in research priorities, funding constraints, or broader global disruptions such as the COVID-19 pandemic, which temporarily shifted research agendas toward urgent health-related priorities and away from large-scale psychometric studies (OECD, 2021; Zhan et al., 2020).

Geographically, leading contributors include Germany, Australia, Turkey, Canada, China, the United Kingdom, Hong Kong, Spain, Taiwan, Austria, and Iran, with Germany at the forefront (Wu et al., 2020). This dominance reflects the presence of advanced psychometric research centers and robust governmental support for statistical and measurement research.

Institutionally, major contributors encompass the National Institutes of Health (NIH), National Science Foundation (NSF), Federal Ministry of Education and Research in Germany, Eunice Kennedy Shriver National Institute of Child Health and Human Development, and the National Institute of Mental Health, along with prestigious academic centers like Boston University and the National Institute on Drug Abuse (Embretson & Reise, 2013). Such entities underscore the role of government funding in advancing fairness in testing and refining statistical models. Their continuous investment reflects a global recognition of the need for psychometric innovation to ensure that assessment tools are culturally unbiased, psychometrically sound, and adaptable across populations. These institutions not only provide financial resources but also foster interdisciplinary collaborations between statisticians, psychologists, and educators, creating an environment that promotes methodological rigor and innovation. Furthermore, their involvement helps in developing open-access data platforms and standardized frameworks for DIF detection, which have become essential for cross-national research and for ensuring equity in educational and psychological measurement practices worldwide.

Institutionally, major contributors encompass the National Institutes of Health (NIH), National Science Foundation (NSF), Federal Ministry of Education and Research in Germany, Eunice Kennedy Shriver National Institute of Child Health and Human Development, and the National Institute of Mental Health, along with prestigious academic centers like Boston University and the National Institute on Drug Abuse (Embretson & Reise, 2013). Such entities underscore the role of government funding in advancing fairness in testing and refining statistical models. Their continuous investment reflects a global recognition of the need for psychometric innovation to ensure that assessment tools are culturally unbiased, psychometrically sound, and adaptable across populations. These institutions not only provide financial resources but also foster interdisciplinary collaborations between statisticians, psychologists, and educators, creating an environment that promotes methodological rigor and innovation. Furthermore, their involvement helps in developing open-access data

platforms and standardized frameworks for DIF detection, which have become essential for cross-national research and for ensuring equity in educational and psychological measurement practices worldwide.

Keyword analysis shows recurring emphasis on terms like "Differential Item Functioning," "Item Response Theory," "Measurement Invariance," and "Fairness in Testing," highlighting the thematic priorities of the field. These patterns reinforce the need for sustained investment in research infrastructure and international collaboration to enhance measurement tools' cultural and linguistic fairness. The frequent recurrence of these terms also indicates that the research community continues to prioritize issues of validity and equity in assessment across diverse populations. Moreover, it reflects the increasing integration of psychometric theory with applied educational and psychological measurement practices. The emphasis on fairness and measurement invariance demonstrates a collective effort to ensure that assessments yield comparable results across genders, languages, and cultural groups. This trend suggests a growing global commitment to ethical testing standards, encouraging transparency, reproducibility, and inclusivity in measurement development and validation processes.

Second, the predominance of IRT-based approaches reflects their conceptual alignment with item-level fairness questions and their ability to separate person and item parameters. The continued reliance on Rasch and 2PL/3PL models mirrors their interpretability, psychometric robustness, and the availability of mature analytical tools. The Rasch model is often preferred in educational contexts due to its strict measurement requirements and simplicity, while the 2PL model offers greater flexibility by allowing item discrimination parameters to vary. Although the 3PL model is widely recognized for addressing guessing behavior, its adoption is less common in applied DIF studies, often limited to large-scale standardized assessments. Bayesian IRT models, while powerful for handling small or imbalanced samples and incorporating informative priors (Fox, 2010), remain underutilized, signaling an important opportunity for methodological development.

Third, in terms of statistical methods used in DIF detection, Mantel–Haenszel (Holland & Thayer, 1988) and logistic regression (Swaminathan & Rogers, 1990) remain the most common non-IRT approaches, especially for dichotomous items, due to their computational efficiency and interpretability.

These methods are often used alongside IRT-based analyses to cross-validate results and ensure robustness. In more recent years, multi-group confirmatory factor analysis (CFA) and measurement invariance testing within structural equation modeling (Cheung & Rensvold, 2002) have become increasingly prevalent for evaluating fairness at the construct level.

Fourth, regarding research platforms and data sources, most DIF and IRT studies are published in Scopus-indexed journals within psychology, education, and social sciences, reflecting the interdisciplinary nature of measurement research. These journals often prioritize empirical rigor, methodological transparency, and cross-cultural validity, which align with the principles underlying DIF and IRT analyses. The most common datasets originate from large-scale international educational assessments, such as PISA, TIMSS, and PIRLS, which offer extensive multilingual and multicultural data. These datasets provide ideal conditions for investigating item bias, response patterns, and construct equivalence across diverse populations. Moreover, the open-access availability of such data encourages replication and comparative studies across nations and educational systems. Publicly available software tools, including R packages ('difR', 'ltm', 'mirt'), along with specialized programs like IRTPRO, Mplus, and Winsteps, have further democratized access to advanced psychometric modeling. This accessibility has supported both experienced researchers and emerging scholars in applying complex IRT and DIF methodologies, contributing to the methodological innovation and expansion of the field worldwide.

Fifth, the majority of samples analyzed in these studies are drawn from student populations in secondary and higher education, reflecting the dominance of educational assessment contexts in DIF and IRT research. Sample sizes vary considerably, ranging from a few hundred participants in small-scale validation studies to over 10,000 respondents in international large-scale assessments. This variation depends largely on research design, data accessibility, and the statistical methods employed. Despite these advances, the review identifies a persistent underrepresentation of participants from low- and middle-income countries, regions often characterized by rich cultural and linguistic diversity. This gap limits the generalizability of findings and underscores the need for more inclusive global research collaborations. Expanding DIF studies to encompass these regions could yield valuable insights into how cultural and linguistic factors influence item functioning and test fairness.

The noticeable surge in DIF-related publications after 2020 appears to be linked to the global shift toward digital learning and assessment environments, accelerated by the COVID-19 pandemic. The rapid growth of online testing platforms has generated vast datasets that facilitate cross-national comparisons and longitudinal analyses. Simultaneously, the international focus on equity and fairness in education and health assessments has intensified, driving greater interest in ensuring that measurement instruments perform consistently across different demographic and cultural groups. These developments have also encouraged the creation of scalable and reproducible DIF analysis pipelines, supported by open-source tools and shared data repositories, marking a significant step forward in the democratization and transparency of psychometric research.

Methodologically, the synthesis supports a multi-method perspective. Combining IRT-based calibration with logistic regression or Mantel–Haenszel enables triangulation of evidence, reducing method-specific biases. Hybrid strategies, such as using alignment optimization to diagnose non-invariance followed by targeted item analyses, streamline workflows in large-group comparisons. Emerging approaches from machine learning offer scalable solutions for detecting potential DIF signals, but careful validation and interpretability are critical before integration into high-stakes testing (Zhan et al., 2020).

Practically, routine incorporation of DIF analyses into test development and validation is recommended.

Establishing transparent and standardized reporting practices in DIF research is essential. Test developers and researchers should clearly document the rationale for anchor item selection or alignment procedures, as this transparency allows others to evaluate the validity of the measurement process. In addition, reporting both statistical significance and effect-size measures provides a more comprehensive understanding of DIF magnitude and practical implications, ensuring that identified differences are not only statistically detectable but also meaningful in context.

Moreover, including item-level diagnostics—such as item characteristic curves, differential functioning plots, and parameter estimates—enables stakeholders, including educators, policymakers, and test users, to interpret results more accurately. Such transparency promotes

accountability and helps prevent misinterpretation or misuse of test results. By adhering to consistent reporting standards, researchers can enhance the comparability of studies, facilitate meta-analyses, and build a more robust cumulative body of knowledge. Ultimately, this commitment to openness strengthens fairness in testing and supports more equitable decision-making processes across educational, psychological, and professional assessment contexts.

Finally, future research should prioritize: (a) advancing Bayesian and semi-parametric IRT models for complex designs; (b) extending cross-cultural validation to underrepresented populations; and (c) operationalizing real-time DIF monitoring in computer-adaptive and digital testing environments. Advancing these areas will enhance both the precision and fairness of measurement in diverse global contexts.

Conclusion

The integrated bibliometric and systematic synthesis clarifies contemporary DIF research patterns and underscores the centrality of item-level fairness for valid inference. By highlighting methodological trends, geographic distribution, and application areas, this review provides a roadmap for researchers and practitioners to implement robust, transparent DIF analyses that support equitable assessment across populations.

Results:

1. The study showed a significant increase in research on Differential Item Functioning (DIF) and Item Response Theory (IRT) since 2010, with a peak in 2017, reflecting growing global interest in fairness and equity in educational and psychological testing.

2. The Rasch Model (1PL) was found to be the most commonly used in DIF-related studies within the IRT framework due to its simplicity, theoretical strength, and ability to ensure fair and stable measurement across different groups, despite the use of more complex models such as 2PL and 3PL in some studies.

3. The most commonly used statistical methods for detecting differential functioning were the Likelihood Ratio Test (LRT), Logistic Regression (LR), and Mantel–Haenszel (MH), due to their computational efficiency, interpretability, and accuracy in identifying biased items.

4. The R platform was identified as the most widely used due to its open-source nature and specialized analytical packages such as difR and ltm, followed by programs like Winsteps, Mplus, and IRTPRO, which are used for more advanced models in psychological and educational analysis.

5. The study concluded that most research focuses on student samples from secondary and higher education, particularly using international databases such as PISA and TIMSS, while low- and middle-income countries remain underrepresented, indicating the need to enhance cultural and linguistic diversity in future studies.

Recommendations

Adopt multi-method DIF detection strategies; integrate DIF analyses into routine validation; strengthen theoretical frameworks that incorporate cultural and linguistic considerations; expand cross-cultural validation in underrepresented contexts; leverage digital platforms for scalable, transparent workflows; and encourage interdisciplinary collaboration among psychometricians, linguists, and domain experts.

Recommendations:

1. Future studies should expand to include samples from countries with greater cultural and linguistic diversity, particularly low- and middle-income nations, to enhance fairness in educational and psychological measurement tools.

2- Researchers should be encouraged to use advanced Item Response Theory (IRT) models alongside the Rasch model to achieve more accurate and comprehensive analyses of differential item functioning.

3- Collaboration between universities and research institutions should be strengthened to develop open-source software specialized in test analysis, improving research quality and reducing reliance on costly commercial programs.

REFERENCES

Acar, T. (2012). Determination of a differential item functioning procedure using the hierarchical generalized linear model: A comparison study with

logistic regression and likelihood ratio procedure. SAGE Open, 2(1). https://doi.org/10.1177/2158244012436760

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. Structural Equation Modeling, 21(4), 495–508.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. Language Testing, 27(1), 101–118. https://doi.org/10.1177/0265532209340194

Binding, G., Koedam, J., & Steenbergen, M. R. (2024). The comparative meaning of political space: A comprehensive modeling approach. Political Science Research and Methods, 12(3), 643–651. https://doi.org/10.1017/psrm.2023.16

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

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Sage.

Çepni, Z., & Kelecioğlu, H. (2021). Detecting differential item functioning using SIBTEST, MH, LR and IRT methods. Journal of Measurement and Evaluation in Education and Psychology, 12(3), 267–285. https://doi.org/10.21031/epod.988879

Chen, H., & Ye, Y. D. (2021). Validation of the Weight Bias Internalization Scale for Mainland Chinese children and adolescents. Frontiers in Psychology, 11, 594949. https://doi.org/10.3389/fpsyg.2020.594949

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233–255.

D'Urso, E. D., De Roover, K., Vermunt, J. K., & Tijmstra, J. (2022). Scale length does matter: Recommendations for measurement invariance testing with categorical factor analysis and item response theory

approaches. Behavior Research Methods, 54(5), 2114–2145. https://doi.org/10.3758/s13428-021-01690-7

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in multiple languages and cultures. Language Testing, 30(2), 201–215.

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. Journal of Measurement and Evaluation in Education and Psychology, 14(1), 76–94. https://doi.org/10.21031/epod.1218144

Fox, J.-P. (2010). Bayesian Item Response Modeling: Theory and Applications. Springer.

Fung, S.-F., & Jin, J. (2023). Gender-based differential item function for the positive and negative semantic dimensions of the Relationship Satisfaction Scale with item response theory. Behavioral Sciences, 13(825). https://doi.org/10.3390/bs13090825

Gershon, S. K., Ruipérez-Valiente, J. A., & Alexandron, G. (2021). Defining and measuring completion and assessment biases with respect to English language and development status: Not all MOOCs are equal. International Journal of Educational Technology in Higher Education, 18(41). https://doi.org/10.1186/s41239-021-00275-w

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory. Sage.

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.

Jafaripour, S., Tabatabaei, O., Salehi, H., & Dastjerdi, H. V. (2024). Applying IRT model to determine gender and discipline-based DIF and DDF: A study of the IAU English proficiency test. International Journal of Language Testing, 14(1), 56–74. https://doi.org/10.22034/IJLT.2023.407117.1268

Jeon, M., & Rijmen, F. (2016). A modular approach for item response theory modeling with the R package flirt. Behavior Research Methods, 48(2), 742–755. https://doi.org/10.3758/s13428-015-0606-z

Joo, S., Ali, U., & Robin, F. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. Large-scale Assessments in Education, 10(18). https://doi.org/10.1186/s40536-022-00135-7

Khalaf, M. A., & Omara, E. M. N. (2022). Rasch analysis and differential item functioning of English language anxiety scale (ELAS) across sex in Egyptian context. BMC Psychology, 10(242). https://doi.org/10.1186/s40359-022-00955-w

Li, B., Ding, C., Shi, H., Fan, F., & Guo, L. (2023). Assessment of psychological zone of optimal performance among professional athletes: EGA and item response theory analysis. Sustainability, 15(7904). https://doi.org/10.3390/su15107904

Li, Y., She, M., Tu, D., & Cai, Y. (2021). Computerized adaptive testing for schizotypal personality disorder: Detecting individuals at risk. Frontiers in Psychology, 11, 574760. https://doi.org/10.3389/fpsyg.2020.574760

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in General English Proficiency Test-Kids listening. Frontiers in Psychology, 12, 767244. https://doi.org/10.3389/fpsyg.2021.767244

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Lawrence Erlbaum Associates.

McManus, I., Chis, L., & Fox, R. (2014). Implementing statistical equating for MRCP(UK) parts 1 and 2. BMC Medical Education, 14(204). https://doi.org/10.1186/1472-6920-14-204

Millsap, R. E. (2011). Statistical Approaches to Measurement Invariance. Routledge.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Medicine, 6(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moradi, E., Ghabanchi, Z., & Pishghadam, R. (2022). Reading comprehension test fairness across gender and mode of learning: Insights from IRT-based differential item functioning analysis. Language Testing in Asia, 12(39). https://doi.org/10.1186/s40468-022-00192-3

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and DIF in international assessments. International Journal of Testing, 11(4), 272–293.

Ozdemir, B., & Alshamrani, A. (2020). Examining the fairness of language test across gender with IRT-based differential item and test functioning methods. International Journal of Learning, Teaching and Educational Research, 19(6), 27–45. https://doi.org/10.26803/ijlter.19.6.2

Reich, H., Rief, W., & Brähler, E. (2018). Cross-cultural validation of the German and Turkish versions of the PHQ-9: An IRT approach. BMC Psychology, 6(26). https://doi.org/10.1186/s40359-018-0238-z

Robitzsch, A., & Lüdtke, O. (2023). Comparing different trend estimation approaches in international large-scale assessment studies. Large-scale Assessments in Education, 11(7). https://doi.org/10.1186/s40536-023-00176-6

Schneider, L., Strobl, C., & Zeileis, A. (2022). An R toolbox for score-based measurement invariance tests in IRT models. Behavior Research Methods, 54, 2101–2113. https://doi.org/10.3758/s13428-021-01689-0

Schwartz, C. E., Stucky, B. D., & Stark, R. B. (2021). Expanding the purview of wellness indicators: Validating a new measure that includes attitudes, behaviors, and perspectives. Health Psychology and Behavioral Medicine, 9(1), 1031–1052. https://doi.org/10.1080/21642850.2021.2008940

Scott, N. W., Fayers, P. M., Aaronson, N. K., et al. (2010). Differential item functioning analyses of health-related quality of life instruments using logistic regression. Health and Quality of Life Outcomes, 8(1), 81. https://doi.org/10.1186/1477-7525-8-81

Shiraito, Y., Lo, J., & Olivella, S. (2023). A nonparametric Bayesian model for detecting differential item functioning: An application to political representation in the US. Political Analysis, 31(3), 430–447. https://doi.org/10.1017/pan.2023.1

Svicher, A., Gori, A., & Di Fabio, A. (2022). The Big Three Perfectionism Scale-Short Form: An item response theory analysis of Italian workers. Frontiers in Psychology, 13, 971226. https://doi.org/10.3389/fpsyg.2022.971226

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27(4), 361–370.

Tabatabaee-Yazdi, M. (2020). Hierarchical diagnostic classification modeling of reading comprehension. SAGE Open, 10(2). https://doi.org/10.1177/2158244020931068

Tan, Q., Cai, Y., Li, Q., Zhang, Y., & Tu, D. (2018). Development and validation of an item bank for depression screening in the Chinese population using computer adaptive testing: A simulation study. Frontiers in Psychology, 9, 1225. https://doi.org/10.3389/fpsyg.2018.01225

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. Quality of Life Research, 16(1), 33–42.

Tian, X., & Dai, B. (2020). Developing a computerized adaptive test to assess stress in Chinese college students. Frontiers in Psychology, 11, 7. https://doi.org/10.3389/fpsyg.2020.00007

Uysal, İ., Ertuna, L., Ertaş, F. G., & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. Journal of Measurement and Evaluation in Education and Psychology, 10(2), 133–148. https://doi.org/10.21031/epod.534312

Zhan, P., et al. (2020). Machine learning approaches for DIF detection. Educational and Psychological Measurement, 80(5), 864–889.

Zhang, C., Dai, B., & Lin, L. (2023). Validation of a Chinese version of the digital stress scale and development of a short form based on item

response theory among Chinese college students. Psychology Research and Behavior Management, 16, 2897–2911. https://doi.org/10.2147/PRBM.S413162

Zhang, Y., Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a computerized adaptive testing for internet addiction. Frontiers in Psychology, 10, 1010. https://doi.org/10.3389/fpsyg.2019.01010

Zhong, S., Zhou, Y., Zhumajiang, W., Feng, L., Gu, J., Lin, X., & Hao, Y. (2023). A psychometric evaluation of Chinese chronic hepatitis B virus infection-related stigma scale using classical test theory and item response theory. Frontiers in Psychology, 14, 1035071.https://doi.org/10.3389/fpsyg.2023.1035071]

Zumbo, B. D. (1999). A Handbook on the Theory and Methods of Differential Item Functioning. Department of National Defense.