Drug Design, Development and Therapy

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

DIF: Concepts, Measurement, and Impact in Patient-Focused Drug Development and Regulatory Decision-Making

Kai Cao[1,2], Wei Liu[1], Lei Yang[1], Meimei Luo[1], Yan Hou[1,3,4]

[1]Department of Biostatistics, School of Public Health, Peking University, Beijing, 100191, People's Republic of China; [2]Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital Medical University, Beijing, 100005, People's Republic of China; [3]Beijing Cancer Hospital, Beijing, 100142, People's Republic of China; [4]Peking University Clinical Research Center, Institute of Advanced Clinical Medicine, Peking University, Beijing, 100191, People's Republic of China

Correspondence: Yan Hou, Email houyan@bjmu.edu.cn

Abstract: Differential Item Functioning (DIF) is critical for Patient-Focused Drug Development (PFDD), particularly in validating Patient-Reported Outcome (PRO) tools. This study links DIF analysis to regulatory decision-making. DIF directly impacts the reliability of drug approval evidence, trial result interpretation, and subgroup consistency evaluation. However, it is not a primary driver of regulatory approval decisions, which prioritize substantial evidence of efficacy and safety. As a "regulatory-impacting measurement bias", DIF detection ensures measurement fairness (a prerequisite for valid cross-group comparisons) and supports the pursuit of validity by identifying tool flaws for optimization, reveals group-specific response differences, and safeguards the accuracy/comparability of PRO data, thereby supporting precise drug development and reliable efficacy assessments. Methodologically, DIF analysis relies on Classical Test Theory (CTT), Item Response Theory (IRT), and methods such as Mantel-Haenszel, logistic regression, and hybrid approaches, but faces caveats including strict sample size requirements (eg, large calibration samples for IRT models) and challenges in detecting multidimensional DIF. With the growing application of PROs in clinical trials, DIF detection has become an indispensable step to mitigate biases that could mislead drug development decisions. This article systematically reviews DIF's definition, classification, detection methods, application in drug development, and existing challenges. Integrating the latest global research and regulatory evidence, it discusses DIF's practical implications for clinical trial design (eg, subgroup analysis, cross-cultural trial adaptation) and regulatory decision-making in PFDD. The work aims to provide theoretical reference and practical guidance for relevant research and practice, highlighting the need to address methodological limitations to strengthen DIF's role in PFDD.

Keywords: differential item functioning, patient-focused drug development, patient-reported outcomes, fairness, drug development

Background

In the Patient-Focused Drug Development (PFDD) process, if Patient-Reported Outcome (PRO) tools have Differential Item Functioning (DIF),[1–3] it will lead to measurement bias,[4] which affects the conclusions of drug effectiveness, the evaluation of treatment subgroup consistency, and the judgment of cross-cultural applicability, thus becoming a key blocking point in regulatory review.

DIF in PRO tools is closely intertwined with drug delivery and formulation development, two core pillars of translating therapeutic potential into clinical benefit. Modern formulation strategies (eg, prodrugs, solid dispersions, nano-delivery systems, and gastroretentive floating drug delivery systems) are increasingly designed to optimize bioavailability, target tissue delivery, and improve patient compliance, particularly for challenging molecules such as PROTACs with poor solubility or membrane permeability.[5] However, the success of these formulation innovations ultimately depends on accurate PRO-based efficacy and safety assessments: for instance, a prodrug designed to enhance oral absorption may show improved pharmacokinetic profiles, but DIF in symptom severity items could mask true

https://doi.org/10.2147/DDDT.S583986
Received: 24 November 2025
Accepted: 24 April 2026
Published: 3 June 2026

therapeutic effects in specific patient subgroups (eg, elderly patients with altered gastrointestinal motility or patients with comorbidities affecting formulation activation).[6,7] Similarly, cross-cultural DIF may invalidate comparisons of PRO data from global clinical trials evaluating region-specific formulation adjustments (eg, taste-masked pediatric formulations or dose-adjusted elderly formulations), undermining the regulatory acceptance of these delivery optimizations.[8] Thus, addressing DIF in PRO tools is not only a measurement fairness issue but a prerequisite for validating the clinical value of advanced drug delivery and formulation strategies.

In the process of measurement and evaluation using scales/questionnaires, the examination of their reliability, validity, and discriminant validity has been widely recognized in the academic community and has become a routine practice. However, their fairness is often overlooked, even though ensuring the fairness of measurement tools is crucial. The neglect of DIF may lead to inaccurate reflection of patients' actual conditions by measurement results, misleading the evaluation of drug efficacy and affecting drug development decision-making.

Amid the growing emphasis on the concept of PFDD, the development of PRO tools faces numerous challenges and requirements. PFDD emphasizes fully considering patients' needs, perspectives, and experiences throughout the entire drug development process, while PRO tools serve as the key means to capture patient-reported outcomes.[9,10] Since PRO tools primarily rely on patients' subjective reports,[11–13] different patient groups may have variations in their understanding and responses to the same item. If the issue of DIF is not addressed during the development of PRO tools, the measurement results may fail to truly reflect patients' actual conditions, which in turn could impact decision-making in drug development. For example, DIF in a symptom assessment item among a particular patient group may result in biased scores that fail to reflect actual symptom severity, thereby misleading assessments of therapeutic efficacy.[14] In PFDD, particularly during the development of PRO tools, placing significant emphasis on the DIF issue is an essential requirement to ensure the accuracy and fairness of measurement results, thereby driving drug development toward a direction that better aligns with patients' interests.[15,16]

A review of key global regulatory documents relevant to PRO/Clinical Outcome Assessment (COA) qualification reveals a critical gap in DIF-specific guidance. The U.S. Food and Drug Administration (FDA)'s PFDD Guidance 4 (Incorporating Clinical Outcome Assessments into Endpoints for Regulatory Decision-Making, draft 2023)[17] mandates robust psychometric validation of PRO tools (including reliability and validity) but does not explicitly require DIF testing or provide methodological standards for evaluating measurement fairness across subgroups. FDA's PFDD Guidance 3 (Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments, finalized October 2025)[18] represents progress by explicitly mentioning measurement invariance and DIF as quantitative methods for evaluating cross-cultural validity; however, it presents these as optional considerations ("one could also present evidence") rather than mandatory requirements, and only in the context of linguistic/cultural adaptation, not for broader subgroup analyses (eg, age, comorbidity). The FDA Drug Development Tools (DDT) Qualification Program, established under the 21st Century Cures Act, provides a framework for PRO tool qualification but does not include DIF evaluation as a mandatory component of the validation package.[19] Similarly, the European Medicines Agency (EMA)'s Reflection Paper on Patient Experience Data (2025)[20] emphasizes patient-centric data collection but lacks specific provisions on DIF assessment for cross-cultural or subgroup analyses. The ICH E9 guideline (Statistical Principles for Clinical Trials, 1998)[21] references "subgroup consistency" but does not identify DIF as a potential source of bias in PRO/COA data. This regulatory gap creates uncertainty for industry: while formulation and delivery innovations are rigorously regulated, the PRO tools used to assess their outcomes lack standardized DIF testing requirements, risking misalignment between technical formulation success and patient-reported clinical benefit.

This review aims to conduct an in-depth exploration of the theories and applications related to DIF, providing references for research and practice in relevant fields.

Introduction

In the landscape of psychometric measurement and applied evaluation, DIF stands as a critical phenomenon that directly impacts the fairness (ie., measurement invariance across groups) and interpretability of assessment tools across diverse populations and contexts, while indirectly supporting validity by mitigating group-specific measurement biases.[3,22] Definition: DIF refers to the systematic variation in an item's statistical performance across distinct groups (eg, gender,

cultural background, disease subgroups) after controlling for the true latent trait or ability level of test-takers.[2] Such differences may arise from biases inherent in the item itself or from variations in culture, language, experience, and other factors across groups. Unlike true differences in the measured construct, DIF introduces measurement bias that distorts score interpretations, leading to inaccurate evaluations.[22,23] Clinical Significance: A symptom assessment item exhibiting DIF across patient subgroups may fail to reflect true symptom severity, thereby misleading judgments of therapeutic efficacy in clinical trials.[24] Given these implications, the detection, classification, and mitigation of DIF have become indispensable in fields where equitable and precise measurement is paramount, particularly in PFDD and the validation of PRO/COA tools.

DIF is categorized along multiple dimensions to guide targeted analysis and interpretation. By the nature and degree of group differences, it is divided into uniform and non-uniform DIF: uniform DIF denotes consistent disparities in item difficulty across all ability levels (eg, a Western history question being systematically harder for Asian students regardless of their historical knowledge[23]), while non-uniform DIF involves inconsistent variations in both difficulty and discrimination, often driven by interactions between items and latent traits.[23] From a dimensionality perspective, unidimensional DIF relates to a single latent trait (detectable via traditional methods) and multidimensional DIF involves multiple traits (requiring advanced analytical frameworks).[25] Additionally, explicit DIF manifests as directly observable score differences, whereas implicit DIF requires statistical modeling to uncover subtle biases.[26] This classification system provides a foundational framework for selecting appropriate measurement methods, as different DIF types demand distinct analytical approaches to ensure accurate detection.

The measurement of DIF relies on two primary theoretical paradigms: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT, focused on observed total and item scores, detects DIF through group comparisons of item performance,[27] while IRT offers a more nuanced approach by modeling item responses as a function of latent traits via Item Characteristic Curves (ICCs), enabling precise quantification of DIF through parameter comparisons.[28–30] Traditional DIF detection methods include the Mantel–Haenszel (MH) method, a nonparametric technique suitable for dichotomous items and small samples, though less sensitive to non-uniform DIF,[25,31,32] and logistic regression, a parametric approach that simultaneously assesses uniform and non-uniform DIF with moderate sample size requirements.[24,33,34] Modern methods, such as IRT models (1PL, 2PL, 3PL) and the Rasch model, enhance flexibility by accommodating diverse item types and enforcing rigorous measurement axioms (eg, fixed discrimination in the Rasch model for interval-scale construction).[35–37] Hybrid methods, integrating parametric and nonparametric techniques (eg, IRT-logistic regression fusion, Rasch-tree hybrids) or machine learning algorithms, further address limitations of single methods, such as handling multidimensional data or latent interactions.[38–40] Collectively, these methods form a comprehensive toolkit for DIF analysis, adaptable to varying research objectives and data characteristics.

In the context of PFDD, DIF analysis assumes heightened significance due to its direct impact on the validity of PRO/COA tools,[41,42] which are central to evaluating patient health status, quality of life, and treatment outcomes in clinical trials.[43] Regulatory authorities worldwide have recognized this importance: the EMA mandates DIF analysis in multinational trials to ensure cross-cultural applicability,[43] while the U.S. National Institutes of Health (NIH) provides PROMIS guidelines recommending DIF assessment.[44] Although explicit DIF regulations for COA tools remain limited at the U.S. FDA and China's Center for Drug Evaluation (CDE), empirical evidence from regulatory reviews demonstrates its pivotal role as supporting evidence in decision-making. A paradigmatic case is the FDA's qualification of COA Tool No.000079 (PROMIS® Oncology Physical Function Scale), where DIF analysis addressed critical regulatory concerns (eg, cultural applicability, recall period consistency), supported the "limited use scenario" qualification pathway, clarified clinical application boundaries, and exempted supplementary studies, ultimately facilitating regulatory approval by addressing specific technical concerns, rather than acting as a primary driver of the qualification decision.[44] This case underscores DIF's role in ensuring COA tools meet "fit-for-purpose" requirements, as measurement non-invariance can lead to endpoint bias, misestimated treatment effects, and compromised "substantial evidence" for drug approval.[43]

Despite its proven value, DIF analysis faces methodological challenges, including handling complex samples (eg, multilevel or missing data) and detecting multidimensional DIF.[40,45–47] Moreover, the lack of unified global regulatory guidelines creates inconsistencies in its application across drug development programs.[48] As PFDD continues to gain traction, the need for standardized DIF practices,[49] integrated into clinical trial protocols and statistical appendices, has

become increasingly urgent. This introduction contextualizes DIF within its conceptual, methodological, and regulatory frameworks, highlighting its irreplaceable role in enhancing measurement equity, validating clinical outcome tools, and bridging academic research with regulatory decision-making. By synthesizing existing knowledge and real-world regulatory evidence, this work aims to underscore the necessity of rigorous DIF analysis in advancing precise, patient-centered drug development and maximizing public health benefits.

Classification of DIF

The existence of DIF[1,50] can lead to biases in measurement results, which in turn affects the fairness and accuracy of evaluations.[51] The detection and control of DIF are of great significance in fields such as psychological measurement, educational assessment, and clinical outcome evaluation, an increasingly important application scenario with the widespread use of PROs in clinical trials.

Uniform DIF and Non-Uniform DIF

Based on the nature and degree of differences, DIF can be classified into uniform DIF and non-uniform DIF[52](Table 1). The core of uniform DIF lies in consistent differences in item difficulty across groups.[52] This means that regardless of the test-takers' ability level, the pass rate or score of one group on the item is consistently lower or higher than that of another group. For example, a math problem may be more difficult for female test-takers than for male test-takers across all ability levels.[50] Another example is a Quality of Life (QoL) assessment item about the importance of family in cross-cultural studies comparing Dutch and Spanish participants.[53] After controlling for the participants' underlying QoL trait level, it is found that this item elicits distinct response patterns between the two groups: Spanish participants interpret "family" as extended family and adopt a relative evaluation strategy (ranking life domains against each other), while Dutch participants define "family" as nuclear family and assess the domain independently. This indicates that regardless of their QoL trait level, Dutch and Spanish participants consistently differ in their response processes and interpretations of this item, reflecting uniform DIF, a consistent group difference in item response probability across all levels of the measured construct. Non-uniform DIF means that the variation in item difficulty with ability level is inconsistent across groups. In other words, non-uniform DIF implies that both item difficulty and item discrimination may differ across groups.[54] It is usually caused by interactions between the item and certain latent traits or background factors of the test-takers.[50] For instance, in a survey on career interests, a question may involve specific career fields or skill requirements, leading to differences in responses among test-takers with different professional backgrounds or skill levels. In such cases, the item exhibits non-uniform DIF.

Clinical Trial Example for Non-Uniform DIF

A prospective multicenter clinical trial focused on moderate-to-severe traumatic brain injury (TBI) patients verified the existence of non-uniform DIF in HRQoL assessment between patient self-reporters and proxy reporters (family members/healthcare professionals).[55] The study used the Short Form-36 (SF-36) to measure role limitations due to physical health (RP) and emotional problems (RE) in 240 TBI patients and detected meaningful non-uniform DIF in 3 out of 4 RP domain items and 1 out of 3 RE domain items after controlling for confounders (eg, age, Glasgow Coma Scale score, neurological recovery). Specifically, for the RP item "were limited in the kind of work or other activities",

Table 1 Classification, Definitions, Characteristics, and Clinical Examples of Differential Item Functioning

Classification	Type	Definition/Key Feature	Clinical Example
By pattern	Uniform DIF	Consistent group difference across all trait levels	Ethnic group difference in EORTC QLQ-C30
	Non-uniform DIF	Group difference varies by trait level	Patient vs proxy in SF-36 (TBI study)
By dimension	Unidimensional DIF	Difference in only one latent trait	Most PRO studies (SF-36, QLQ-C30)
	Multidimensional DIF	Differences across multiple traits	Rare in current clinical PROs
By detectability	Explicit DIF	Directly observed score difference	Unadjusted group score differences
	Implicit DIF	Detected only after statistical adjustment	DIF after covariate/propensity adjustment

Abbreviations: TBI, traumatic brain injury; SF-36, Short Form-36; RP, physical health; RE, emotional problems; QoL, Quality of Life.

proxies showed a consistently lower probability of giving a favorable response than patients when the rest-score (reflecting role limitation severity) was 0 (severe limitations), but this trend was reversed when the rest-score exceeded 60 (mild limitations), proxies then gave more favorable responses than patients at the same trait level. This typical non-uniform DIF was driven by the interaction between the respondent type (patient/proxy) and the latent trait of role limitation severity, which led to inconsistent item difficulty across the entire trait continuum.

Clinical Trial Example for Uniform DIF

A secondary analysis of two Phase 3 non-inferiority clinical trials on paroxysmal nocturnal hemoglobinuria (PNH) explored DIF in the European Organisation for Research and Treatment of Cancer QLQ-C30 (EORTC QLQ-C30) between Asian and non-Asian patients.[56] The study included 441 PNH patients and detected negligible uniform DIF in partial function and symptom items (eg, physical functioning item "Do you have any trouble doing strenuous activities?" and fatigue item "Were you tired?") after adjusting for propensity scores (including trial type, demographic and clinical factors). Although the uniform DIF magnitude did not reach the clinically meaningful threshold (pseudo-$R^2 < 0.018$, far below Zumbo's 0.13 criterion), it still reflected consistent group differences in item response: non-Asian patients were more likely to endorse poor health for strenuous activity difficulty, while Asian patients were more likely to report fatigue-related symptoms across all levels of the underlying physical function trait, this is a typical manifestation of uniform DIF in clinical outcome assessment, with consistent response differences between groups regardless of the latent trait level.

Unidimensional DIF and Multidimensional DIF

Unidimensional DIF refers to items with differences in only one dimension, meaning that item difficulty or discrimination is related to only one latent trait. Unidimensional DIF can usually be detected and analyzed using traditional DIF detection methods,[57] such as the MH method and logistic regression. Multidimensional DIF refers to items with differences in multiple dimensions, meaning that item difficulty or discrimination is related to multiple latent traits. The detection and analysis of multidimensional DIF are relatively complex and require the use of multidimensional DIF detection methods, such as the multidimensional IRT method and multidimensional logistic regression.[58]

Clinical Trial Example for Unidimensional DIF

Both of the above-mentioned clinical trials adopted unidimensional DIF detection methods (logistic regression for binary/ordinal items) and focused on unidimensional DIF in single latent trait domains. Sebille et al targeted the single latent trait of "role limitation severity" (physical/emotional) in SF-36's RP and RE domains and used multivariable logistic regression to detect unidimensional DIF between patients and proxies,[55] all identified DIF items were only related to the single trait of role limitation, and no other latent traits (eg, cognitive function, emotional state) were involved, which is a typical unidimensional DIF in clinical HRQoL assessment. Similarly, Schwartz et al analyzed unidimensional DIF in single trait domains (eg, physical functioning, emotional functioning, fatigue) of the EORTC QLQ-C30 between Asian and non-Asian PNH patients and used ordinal logistic regression to test the interaction between ethnicity and the total score of a single domain,[56] all detected DIF items were associated with only one latent trait, confirming the applicability of traditional unidimensional DIF detection methods in clinical trial data analysis.

To date, multidimensional DIF has been less frequently reported in clinical trial-based HRQoL assessment, which may be due to the fact that most PROMs (eg, SF-36, EORTC QLQ-C30) are designed for unidimensional domain measurement, and the latent traits of clinical outcomes are often relatively single in specific research scenarios. However, with the development of comprehensive PROMs that measure multiple latent traits simultaneously, the detection of multidimensional DIF will become an important research direction in clinical outcome evaluation.

Explicit DIF and Implicit DIF

Explicit DIF refers to differences in items across groups that can be directly observed, usually manifested as significant differences in item scores or pass rates. It is often caused by obvious differences in item language, culture, or format, and can be detected by directly comparing item scores or pass rates across groups.[59]

Implicit DIF refers to differences in items across groups that cannot be directly observed but can be detected through statistical analysis. It is usually caused by subtle interactions between the item and certain latent traits or background factors of the test-takers, requiring advanced statistical analysis methods for detection and analysis.[60]

Clinical Trial Example for Explicit and Implicit DIF

The two clinical trials also provide typical clinical scenarios for explicit and implicit DIF. Explicit DIF was reflected in the direct score differences between groups in the two studies: Sebille et al found that proxies reported significantly lower SF-36 RP (28.1 vs 43.1) and RE (36.7 vs 56.4) domain scores than patients at the descriptive statistics level, and these direct score differences were the explicit manifestation of potential DIF;[55] Schwartz et al also found that Asian PNH patients had slightly better baseline role/emotional functioning scores and worse constipation/diarrhea scores than non-Asians, which were also explicit score differences suggesting possible DIF.[56]

All DIF items identified in the two studies were implicit DIF in essence: the direct score differences between groups could not be directly attributed to DIF (as they may be confounded by clinical factors such as disease severity, neurological recovery, and treatment history), and only after controlling for confounders through advanced statistical methods (multivariable logistic regression, propensity score adjustment) could the subtle item response differences between groups (ie., implicit DIF) be detected. For example, the non-uniform DIF in TBI patients' proxy/patient assessment could only be identified after adjusting for age, Glasgow Coma Scale score, and neurological recovery; the uniform DIF in PNH patients' ethnic group comparison was only found after propensity score adjustment for trial type, LDH stratum, and transfusion history. These results confirm that implicit DIF is the main form of DIF in clinical trial data, and its detection relies on rigorous statistical control of confounding factors.

Measurement of DIF

The detection and analysis of DIF usually rely on specific theoretical models, such as CTT and IRT[61] (Table 2). CTT focuses primarily on total test scores and item scores, detecting DIF by comparing differences in item scores across different groups.[62,63] IRT, on the other hand, starts from the perspective of individuals' latent traits. It analyzes the performance of items across different ability levels by establishing ICCs, thereby detecting and explaining DIF more accurately.[62,63]

Traditional Measurement Methods

Mantel–Haenszel (MH)

The MH method is a widely adopted nonparametric approach for DIF measurement. This technique evaluates DIF presence by comparing conditional probabilities of correct responses across distinct demographic groups (eg, focal and reference groups) at matched ability levels.[64,65]

The methodological procedure involves three sequential steps. First, test-takers are stratified into homogeneous ability groups based on total test scores or relevant ability indicators. Second, within each ability stratum, a 2×2 contingency table is constructed where rows indicate group membership (reference vs focal groups) and columns represent dichotomous item responses (correct/incorrect). Third, the common odds ratio estimator α_{MH} is computed across all strata using the following formula.[66]

$$\alpha_{MH} = \frac{\sum_{k=1}^{K} \frac{a_k d_k}{N_k}}{\frac{b_k c_k}{N_k}}$$

Here, a_k and d_k represent the number of test-takers who answered the item correctly in the reference group and focal group, respectively, within the ability-level group k. b_k and c_k represent the number of test-takers who answered the item incorrectly in the reference group and focal group, respectively, within the ability-level group k. N_k denotes the total number of test-takers in the ability-level group k.

Hypothesis Testing: Determine the presence of DIF by calculating the MH statistic, which follows a chi-square distribution with 1 degree of freedom.

Table 2 Summary Table of Differential Item Functioning Measurement Methods

Category	Specific Method	Core Principle	Key Parameters/Formulas/Steps	Advantages	Limitations
Basic Theoretical Models	Classical Test Theory	Focuses on total test scores and item scores; detects DIF by comparing item score differences across groups	Relies on inter-group difference analysis of total scores and item scores	Simple principle, easy to operate	Does not consider individual latent traits; limited accuracy
	Item Response Theory (IRT)	Based on individuals' latent traits; analyzes item performance across ability levels via Item Characteristic Curves (ICCs)	Difficulty parameter (b), Discrimination parameter (a), Guessing parameter (c); logistic formulas for 1PL/2PL/3PL models	Enables accurate DIF interpretation; adaptable to various item types	Complex calculations, requires specialized software, large samples for stable estimates (>200 per group), and local independence assumptions
Traditional Measurement Methods	Mantel-Haenszel (MH)	Nonparametric method; compares conditional probabilities of correct responses between focal and reference groups at matched ability levels	1. Stratify test-takers by total scores; 2. Construct 2×2 contingency tables; 3. Calculate common odds ratio α_{kmh}; Chi-square test (df=1)	Easy to understand; suitable for dichotomous items and small samples	Low sensitivity in detecting non-uniform DIF
	Logistic Regression Method	Parametric method; models item responses as a function of group membership, total score, and their interaction	Model includes group membership, total score, and group×total score interaction term	Simultaneously detects uniform and non-uniform DIF; low sample size requirements	Strict model assumptions
Modern Measurement Methods	IRT Series Models (1PL/2PL/3PL)	Relates latent traits to response probabilities via ICCs; compares inter-group item parameter differences	1PL: Only difficulty (b); 2PL: +Discrimination (a); 3PL: +Guessing (c); logistic probability formulas	Adaptable to various item types and sample sizes; strong interpretability	Complex calculations, software-dependent, model identification constraints, and local independence assumptions
	Rasch Model	Special form of 1PL; fixes discrimination parameter at 1, adheres to fundamental measurement axioms, and requires strict data fit	Fit indices (infit/outfit mean square values) must meet thresholds; only includes difficulty parameter (b)	Enables interval scale construction; suitable for clinical assessment tools; high psychometric rigor	Strict data fit requirements, need to eliminate poorly fitting items, low flexibility in parameter estimation
Hybrid Methods	IRT-Logistic Regression Fusion	Integrates advantages of parametric models; controls latent trait distribution differences and quantifies non-uniform DIF	Model includes θ_{IRT} (latent trait) and β_3 (group×total score interaction term for non-uniform DIF quantification)	Balances latent trait control and accurate DIF quantification	Higher complexity than single methods
	Machine Learning Hybrids (eg, Random Forest)	Fuses nonparametric advantages via algorithms; identifies measurement invariance violations through resampling and permutation importance	Based on bootstrap samples; permutation importance for screening violating items	No strict model assumptions; excels at capturing complex interaction effects	Weak result interpretability; requires professional algorithm support
	Iterative Purification Hybrids	Maintains Type I error control via recursive updating of expectation-variance terms until anchor item stability	Recursively updates expectation-variance terms to ensure anchor item stability	Effective error control; improves DIF detection accuracy	Tedious process; dependent on anchor item selection
	Rasch-Tree Hybrids	Node-specific item difficulties vary across covariate-defined subgroups; detects latent interaction effects	Includes β_{jh} (node-specific difficulty parameters); suitable for subgroup analysis (eg, native vs non-native speakers)	No pre-specified grouping variables; accurately captures latent interaction effects	Complex model construction; requires covariate definition support
	Other Hybrid Methods	Integrates advantages of different models to address limitations of single methods	eg, Mixture Rasch Models, Rasch-MH Synergy, etc.	Adaptable to diverse research scenarios	Some methods are not widely adopted; high application threshold

Abbreviations: DIF, Differential Item Functioning; IRT, Item Response Theory; MH, Mantel–Haenszel; ICC, Item Characteristic Curves; PL, Parameter Logistic Model.

https://doi.org/10.2147/DDDT.S583986

The MH method is simple and easy to understand. It is suitable for dichotomous items (items with only two response options, eg, correct/incorrect) and small sample sizes, but it is less sensitive to the detection of non-uniform DIF.[67–69] The method assumes constant item difficulty across ability levels, leading to low power in detecting non-uniform DIF (Type II error risk). Besides, in unbalanced group sizes (eg, reference group n=500, focal group n=50), the MH statistic tends to overestimate DIF presence (Type I error risk).[67–69]

Logistic Regression Method

The logistic regression method is a parametric DIF detection approach.[6,70] It evaluates the presence of DIF by modeling item responses as a function of group membership, total score, and their interaction. This method can assess both uniform and non-uniform DIF simultaneously and has relatively low requirements for sample size, but it imposes strict assumptions on the model.[71,72] Besides, model assumption violations: Requires linearity between total scores and logit-transformed responses, which is often violated in clinical PROMs with skewed trait distributions (eg, fatigue scales in terminal illness).[71,72] Violation leads to Type I error inflation or Type II error (missed DIF).

Modern Measurement Methods

IRT Method

IRT models test-taker responses by establishing a probabilistic framework that relates discrete item responses to latent traits (unobservable abilities such as intelligence or anxiety) through ICCs. This parametric approach detects and explains DIF by comparing group-level differences in item parameters.[73–76]

Difficulty (*b*)–Latent trait level at which the probability of correct response reaches 50%

Discrimination (*a*)–Sensitivity to distinguish test-takers across trait levels

Guessing (*c*)–Baseline probability of correct response under random guessing.

The item response theory framework models the probability $P(\theta)$ of correct response via logistic functions parameterized by latent trait. Beginning with the most parsimonious form:

$$P(\theta) - \frac{1}{1+e^{-(\theta-b)}} \text{ (1-Parameter Logistic Model, 1PL)}$$

where θ denotes test-takers' latent trait level.

While IRT accommodates diverse item types and sample sizes, computationally intensive parameter estimation requires specialized software and sufficient calibration samples; a commonly cited empirical reference is >200 per group for stable frequentist estimation. Notably, the actual sample size required for robust IRT parameter estimation is not a fixed threshold, but is jointly determined by item parameters (eg, the range and distribution of difficulty and discrimination), test length (more items typically require larger samples), and estimation methods (Bayesian estimation often has lower sample size demands compared to frequentist approaches such as maximum likelihood estimation).

Extension to the two-parameter model incorporates discrimination (*a*):

$$P(\theta) = \frac{1}{1+e^{-(a(\theta-b))}} \text{ (Two-Parameter Logistic Model, 2PL)}$$

Further generalization via the three-parameter model accounts for pseudoguessing (*c*):

$$P(\theta) = c + (1-c) \cdot \frac{1}{1+e^{-(a(\theta-b))}} \text{ (Three-Parameter Logistic Model, 3PL)}$$

The limitation of IRT Method lies in multiple-testing challenge: In multidimensional IRT for 10+ traits, family-wise error rate exceeds 0.95 without correction (eg, Holm-Bonferroni).[77]

Rasch Model

The Rasch model[78–80] is commonly used for DIF measurement and is a specific form of the 1PL, is defined by two core features: it adheres to the 1PL formula while strictly following fundamental measurement axioms, specifically, it fixes the

item discrimination parameter at 1 and imposes rigorous data fit requirements (eg, infit and outfit mean square values must align with predefined thresholds to ensure measurement validity). In contrast, the 1PL model, situated within the IRT framework, functions as a statistically simplified model that allows flexibility in parameter estimation (eg, the discrimination parameter, though conceptually uniform across items, can be estimated rather than fixed, and data fit criteria are relatively lenient compared to the Rasch model). For operational applications, distinct choices are recommended based on research objectives: if the goal is to construct an interval scale (eg, for clinical assessment tools where precise, comparable score interpretations are critical), the Rasch model should be adopted, with items exhibiting poor fit (eg, significant deviations from model expectations) eliminated to ensure the scale's psychometric rigor; if the objective focuses on parameter comparison or model selection (eg, Akaike Information Criterion-based model comparison to evaluate relative goodness-of-fit), the 1PL model is more suitable as an initial baseline model within the IRT framework, given its balance of simplicity and estimation flexibility.

Hybrid Methods

Hybrid approaches integrate the advantages of parametric models (eg, IRT) and nonparametric methods (eg, MH), and address the limitations of single methods through statistical model nesting or algorithm cascading.

For example, the canonical formulation, which integrates IRT with logistic regression, expressed as:[81]

$$\log\left(\frac{P(Y=1|\theta,G)}{1-P(Y=1|\theta,G)}\right)=\underbrace{\beta_0+\beta_1\theta_{IRT}}_{\text{IRT calibration}}+\underbrace{\beta_2 G+\beta_3(\theta\times G)}_{\text{DIF detection}}$$

This compound model achieves dual objectives: θ_{IRT} controls latent trait distribution differences, while the β_3 interaction term quantifies non-uniform DIF. G means group.

Furthermore, machine learning hybrids extend this through algorithmic fusion. For random forest implementations.[82] where B denotes bootstrap samples, and permutation importance identifies items violating measurement invariance.

Iterative purification hybrids maintain Type I error control via:[83]

$$\chi^2_{MH-adjusted}=\left(\frac{\left|\sum_{k=1}^{K}E[A_k]-\sum_{k=1}^{K}A_k\right|-0.5}{\sqrt{\sum_{k=1}^{K}Var(A_k)}}\right)^2$$

with expectation-variance terms recursively updated until anchor item stability is achieved.

Besides, Rasch-based hybrids play a critical role in clinical and educational measurement. For example, the Rasch-Tree Hybrids:[84]

$$P(X_{ij}=1|\theta_i,\beta_{jh})=\frac{\exp(\theta_i-\beta_{jh})}{1+\exp(\theta_i-\beta_{jh})}$$

where β_{jh} are node-specific item difficulties that vary across covariate-defined subgroups (eg, language learners vs native speakers). This detects latent interaction effects without pre-specifying grouping variables.

Beyond the aforementioned methods, there are a variety of hybrid approaches, such as Mixture Rasch Models,[85] Rasch-MH Synergy,[86] and so on.

Figure 1 presents the integrated framework of DIF classification and corresponding measurement methods, clearly showing the logical relationship between different DIF types and applicable detection tools.

Critical Comparison of DIF Detection Methods

While Traditional Measurement Methods and 2.2 describe individual DIF detection methods, selecting the optimal approach requires understanding their relative strengths, limitations, and suitability for specific research contexts.

Method Selection Decision Framework

For small samples (<200 per group) with dichotomous items:

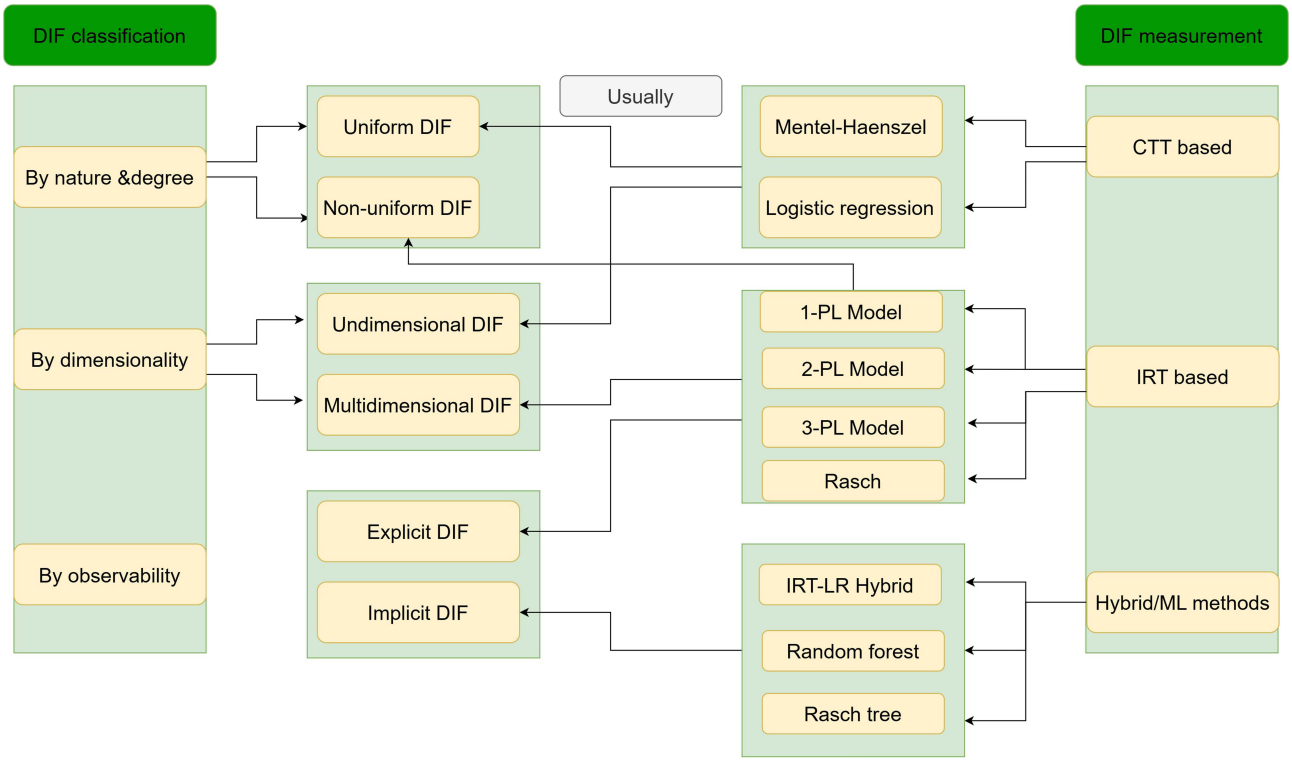

Figure 1 Classification and Measurement Framework of Differential Item functioning.
Abbreviations: DIF, Differential Item Functioning; 1PL, 1-Parameter Logistic Model; 2PL, Two-Parameter Logistic Model; 3PL, Three-Parameter Logistic Model; IRT, Item Response Theory; CTT, Classical Test Theory; ML, machine learning.

The Mantel–Haenszel method remains the preferred choice due to its nonparametric nature and minimal sample size requirements. However, researchers must acknowledge its insensitivity to non-uniform DIF and potential Type I error inflation with unbalanced group sizes.

For ordinal items requiring simultaneous uniform/non-uniform DIF detection:

Logistic regression offers the best balance of statistical power and interpretability. Its primary limitation, strict linearity assumptions—can be mitigated through residual diagnostics and model fit assessment.

For large-scale PRO instrument development:

IRT-based methods (1PL/2PL/3PL) provide superior precision and interval-scale construction capability. The trade-off is substantial: requiring >200 subjects per group for stable estimation, specialized software expertise, and rigorous assessment of local independence assumptions.

For complex multidimensional constructs:

Hybrid methods (IRT-logistic regression fusion, Rasch-tree approaches) demonstrate superior performance in detecting latent interaction effects without pre-specified grouping variables. However, their computational complexity and reduced result interpretability present practical barriers.

Critical Synthesis

No single method universally outperforms others. The choice depends on four key factors: (1) sample size and statistical power; (2) item type (dichotomous vs ordinal); (3) research objectives (exploratory screening vs confirmatory validation); and (4) available technical expertise. Notably, for regulatory submissions, we recommend employing multiple methods (eg, MH for screening, IRT for confirmation) to enhance evidentiary robustness, as demonstrated in the COA #000079 qualification process.

DIF in PFDD: Regulatory Requirements, Case Evidence, and Decision-Making Impact

In clinical drug trials, DIF analysis holds significant application value. On one hand, it can help assess whether differences in PRO measurements between different treatment groups truly reflect treatment effects, rather than being caused by biases in the measurement tools. On the other hand, DIF analysis can also be used to explore differences in treatment responses among different patient subgroups, providing a basis for precise drug treatment.

In PFDD, PRO tools are important means to evaluate patients' health status and quality of life. However, due to the diversity of patient populations, PRO tools may exhibit DIF across different patient groups. The Role of DIF in PFDD see flow chart (Figure 2).

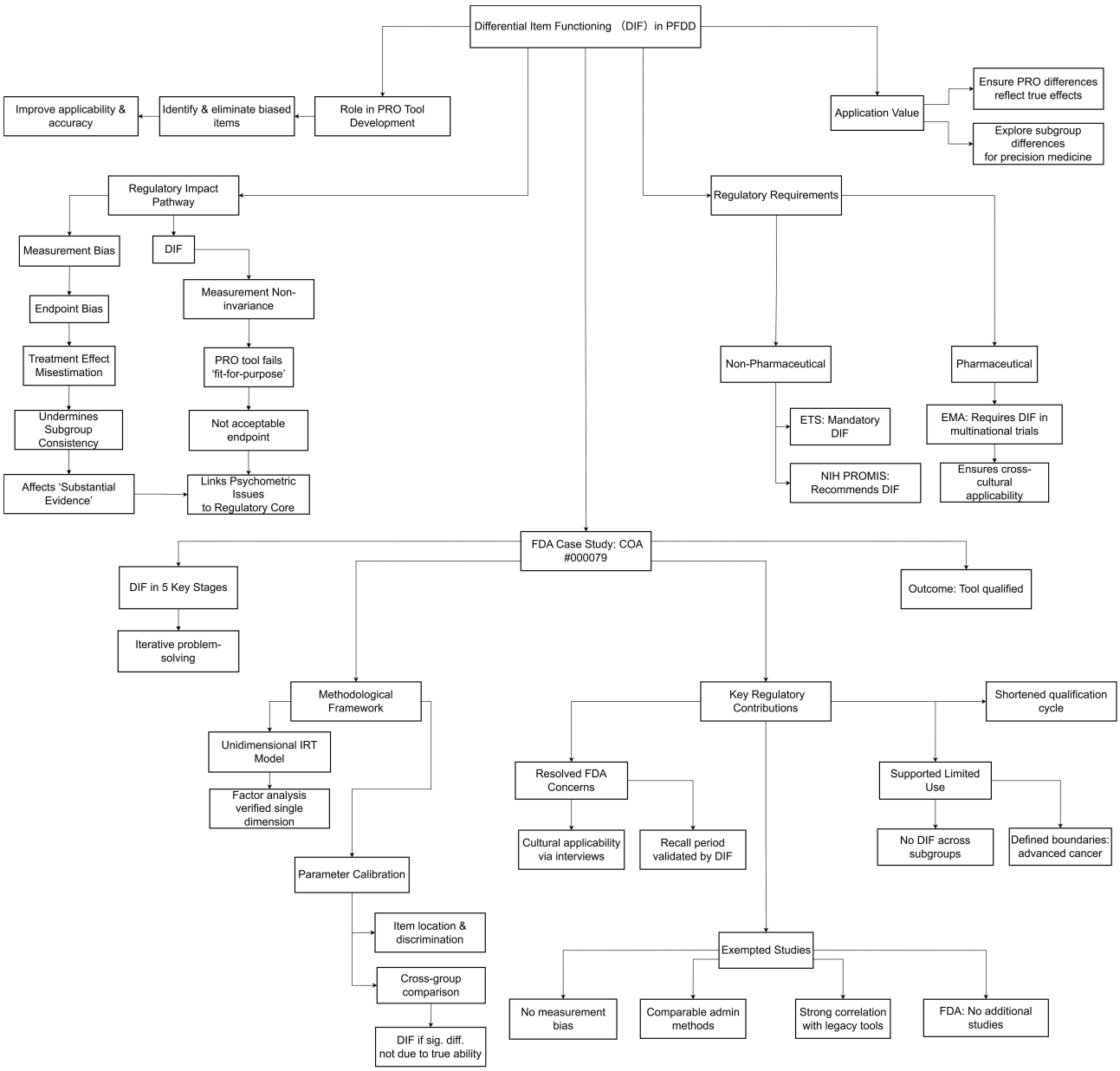

Figure 2 The Role of Differential Item Functioning in Patient-focused drug development: Regulatory Requirements, Case Evidence, and Decision-Making Impact.

https://doi.org/10.2147/DDDT.S583986

Cross-Domain Regulatory Requirements for DIF

Regulatory authorities in many countries attach great importance to the detection and management of DIF to ensure the fairness and effectiveness of testing and evaluation tools.

In Non-Pharmaceutical R&D Fields

The Educational Testing Service in the United States stipulates that DIF analysis must be conducted during test development and incorporated into the routine analysis process. Additionally, the National Institutes of Health in the United States has issued PROMIS guidelines and standards, recommending methods and procedures for DIF assessment. These measures indicate that regulatory authorities not only pay attention to the existence of DIF but also its impact on practical decision-making and outcomes.

In Pharmaceutical R&D Fields

The EMA emphasizes cross-cultural validation of evaluation tools in multinational trials to ensure their applicability across different countries and cultures, and recommends DIF analysis as a quantitative method to support such validation (though not mandating it as a standalone regulatory requirement).[87]

Impact of DIF on FDA Regulatory Decision-Making: A Case Study of Clinical Outcome Assessment Tool No. 000079

To date, neither the U.S. FDA nor China's CDE has issued explicit regulatory guidelines on DIF for COA tools. However, empirical observations from the FDA's review processes reveal that DIF analysis plays a pivotal role in underpinning regulatory decisions, particularly in validating the reliability, generalizability, and scientific rigor of COA instruments. This section takes the FDA's qualification review of COA Tool No.000079 (PROMIS® Oncology Physical Function Scale Qualification Program) as a paradigmatic case,[88] systematically analyzing how DIF analysis permeated the entire lifecycle of scale development and qualification, and ultimately facilitated the sponsor's successful attainment of regulatory approval.

Overview of DIF Analyses in the Qualification Process of COA Tool No. 000079

DIF analysis was integrated into five key stages of qualification. Each stage targeted specific regulatory concerns. Table 3 summarizes the purpose, analytical targets, and core findings of each DIF analysis, demonstrating its iterative and problem-solving role in addressing regulatory requirements.

Table 3 Differential Item Functioning Analyses Conducted During the Qualification Process of Clinical Outcome Assessment Tool No. 000079 (PROMIS® Oncology Physical Function Scale Qualification Program)

Order	Stage	DIF Analysis Purpose	Analytic Target	Results and Conclusions
1	General Entry Library Development	Ensure entries are unbiased across disease groups	21,133 participants (including clinical populations such as cancer and Chronic Obstructive Pulmonary Disease, COPD)	After excluding 44 entries with significant DIF, the remaining entries showed good consistency in measurements across different disease groups.
2	Development of tumor-specific entry libraries	Verify the consistency of entry parameters between the general population and the cancer population	521 cancer patients + general population data	Most entries show minimal DIF, allowing the same entry parameters to be used for both general and cancer populations without separate calibration.
3	8c Entry Version Filter	Addressing FDA's concerns about "unclear recall period"	31 candidate entries (7-day recall vs no recall)	No entries showed significant DIF, confirming the comparability of the two recall periods. The 7-day recall period was ultimately selected as the standard.
4	Test method validation	Confirm that different testing methods (paper, electronic, or telephone) are unbiased	923 patients (COPD, depression, rheumatoid arthritis)	No significant DIF was found among different testing methods, with ICC (within-group correlation coefficient) ranging from 0.85 to 0.93, indicating reliable and comparable results.
5	Qualification Certification Supplementary Materials	Supports cross-cultural/cross-population applicability of the 8C Entry Version of the Scale	8 final entries (patients with different cultural backgrounds/ tumor types)	The absence of significant DIF demonstrates the scale's applicability to patients with advanced-stage tumors from diverse cultural backgrounds and tumor types.

Methodological Framework of DIF Analysis for COA Tool Qualification

The DIF analysis for COA Tool No.000079 adhered to a rigorous methodological framework, grounded in IRT and targeted parameter comparison, to ensure the validity of its findings.

Selection of the Underlying Measurement Model

A unidimensional IRT model was employed as the analytical foundation. Prior factor analysis had verified that while physical function encompasses subdomains (eg, mobility, upper limb function, trunk function), these subdomains exhibit high intercorrelation and can be aggregated into a single overarching measurement dimension "overall physical function level". This aligns with the core "unidimensionality" assumption of IRT models, ensuring the methodological appropriateness of DIF detection.

Parameter Calibration and DIF Determination Criteria

First, IRT parameters of 168 candidate items were calibrated, including: item location (reflecting the level of physical function difficulty required to endorse the item); and item discrimination (representing the item's ability to distinguish between individuals with different physical function levels). Subsequently, cross-group comparisons of IRT parameters were conducted for key population subgroups (eg, cancer vs chronic obstructive pulmonary disease patients, cancer patients vs healthy controls). An item was classified as exhibiting DIF if: there were statistically significant differences in IRT parameters across groups; and such differences could not be attributed to genuine variations in physical function levels between groups (eg, an item's difficulty parameter was significantly higher in cancer patients than in healthy individuals, independent of actual physical function disparities). Items without significant parameter differences were deemed DIF-free.

Key Regulatory Contributions of DIF Analysis

DIF analysis directly addressed critical FDA concerns, supported qualification pathway selection, defined clinical application boundaries, and streamlined the regulatory process, as detailed below.

Resolving Regulatory Concerns: Cultural Applicability and Recall Bias

In 2016, the FDA raised two objections: (1) potential cultural inapplicability of certain items across diverse populations; and (2) unreliable results from unclear recall periods. To address these concerns: 8 cross-culturally appropriate items were selected via international cognitive interviews, and subsequent DIF analysis confirmed no measurement bias across patients from different cultural backgrounds; DIF analysis validated the consistency of item functioning between the "7-day recall period" and "no recall period" demonstrating result comparability. The FDA ultimately endorsed the 7-day recall period as the standard, eliminating the regulatory barrier of "unclear recall periods".

Supporting "Limited Use Scenario" Qualification and Defining Application Boundaries

In 2019, the FDA proposed two qualification pathways: "limited use scenario" and "broad use scenario," with the former requiring evidence of the tool's validity in specific populations/scenarios. DIF analysis provided critical empirical support for the "limited use scenario" pathway: no significant DIF was detected across patients with different tumor types (eg, breast cancer, lung cancer, hematological malignancies), performance status scores (0–3), or treatment stages (preoperative, postoperative, 1-year follow-up); consistent item functioning was confirmed across administration methods. These findings directly validated the scale's reliability in the targeted "limited scenario", "adult advanced cancer patients receiving active treatment, without cultural or language restrictions", leading to the FDA's approval of qualification under this pathway.

Notably, DIF analysis also clarified the scale's clinical application boundaries. The analysis was restricted to "advanced cancer patients (Stage III–IV) with performance status scores 0–3 and receiving active treatment," excluding early-stage cancer patients and those with performance status score 4 (completely bedridden). Consequently, the FDA required the sponsor to conduct additional studies to verify the scale's applicability in these untested populations.

Exempting Supplementary Studies and Shortening the Qualification Cycle

A common regulatory requirement during COA tool qualification is the conduct of supplementary studies to confirm validity. However, DIF analysis and associated evidence for COA Tool No.000079 were sufficiently robust to obviate this need: the scale exhibited no measurement bias in the target population; results were comparable across administration methods (paper-based, electronic, telephone) and recall periods; strong correlations were observed with established tools (eg, SF-36, HAQ-DI), with Pearson correlation coefficients ranging from 0.80 to 0.88, confirming concurrent validity. Following an October 2018 meeting, the FDA explicitly stated that "no additional studies were required," only requesting a revision plan related to DIF analysis, significantly shortening the qualification cycle.

In summary, DIF analysis served as a cornerstone of evidence for the FDA qualification of COA Tool No.000079. It addressed key regulatory concerns, supported the selection of the appropriate qualification pathway, defined clinical application boundaries, and streamlined the review process. Ultimately, DIF analysis enabled the scale to be recognized as a "qualified tool" for assessing physical function in advanced cancer clinical trials, contributing to the standardization of clinical trial endpoints and providing valuable insights for the regulatory evaluation of COA tools more broadly.

Regulatory Impact Pathway of DIF

DIF exerts an impact on regulatory decisions through a clear pathway: DIF leads to measurement bias, which further causes endpoint bias. Endpoint bias results in misestimation of treatment effects, which undermines the consistency of subgroups and affects the formation of "substantial evidence" for drug approval. Notably, DIF is not a standalone criterion for regulatory approval, but its induced measurement bias can undermine the scientific rigor of the substantial evidence required for approval, thus becoming a key consideration in regulatory review.

At the same time, measurement non-invariance makes PRO tools fail to meet the "fit-for-purpose" requirement, thus being unable to become acceptable endpoints for regulatory review. This pathway directly links psychometric issues to regulatory core concerns, highlighting the criticality of DIF in drug development.

Balanced Perspective on DIF's Regulatory Role

Despite the growing recognition of DIF in regulatory practice, several limitations to its regulatory weight should be acknowledged. First, FDA's PFDD Guidance 3[49] and EMA's Reflection Paper on Patient Experience Data[20] only mention DIF as an "optional" method for cross-cultural validation, not a mandatory requirement for all PRO/COA tools. Second, in the FDA's qualification process for COA tools (eg, COA #000079), DIF analysis served as "supporting evidence" to address specific concerns (eg, cultural applicability) rather than a "prerequisite" for qualification. Third, some scholars[54,89] have argued that overemphasizing DIF may impose unnecessary methodological burdens on drug development, especially for small-scale trials or rare disease studies where sample size constraints limit robust DIF detection. These perspectives highlight that DIF's regulatory role is contextual and supplementary, rather than universal and decisive.

DIF's Impact Pathway in Regulatory Decision-Making

To clarify DIF's role in PFDD, we present a conceptual framework illustrating how DIF analysis influences regulatory decisions (Figure 3). This pathway demonstrates that DIF is not merely a statistical exercise but a critical quality assurance mechanism with tangible regulatory consequences.

The DIF-Regulatory Impact Chain

Level 1: Measurement Quality Assurance

DIF detection identifies items functioning differently across subgroups, flagging potential measurement bias that could compromise PRO data validity. Items exhibiting significant DIF require investigation, whether through qualitative research (cognitive interviews) to understand response process differences or statistical adjustment (item removal, scoring modifications).

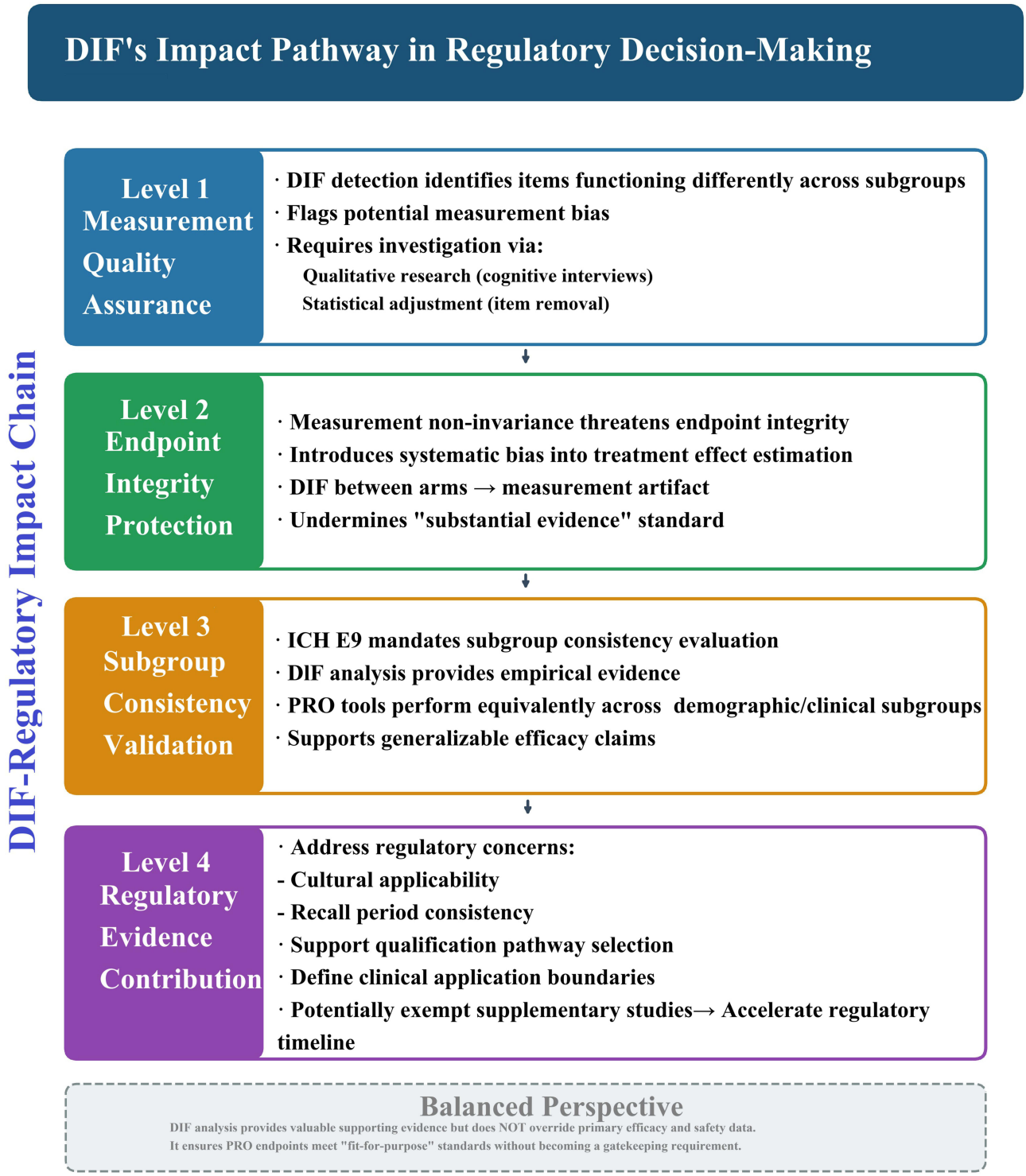

Figure 3 The DIF-Regulatory Impact Chain: A conceptual framework illustrating how DIF analysis influences regulatory decisions at four levels.
Abbreviation: DIF, Differential Item Functioning.

Level 2: Endpoint Integrity Protection

Measurement non-invariance threatens endpoint integrity by introducing systematic bias into treatment effect estimation. When DIF exists between treatment arms, observed score differences may reflect measurement artifact rather than true treatment effects, undermining the "substantial evidence" standard required for drug approval.

https://doi.org/10.2147/DDDT.S583986

Level 3: Subgroup Consistency Validation

Regulatory guidelines (ICH E9) mandate evaluation of treatment effect consistency across subgroups. DIF analysis provides empirical evidence that PRO tools perform equivalently across demographic/clinical subgroups, supporting generalizable efficacy claims.

Level 4: Regulatory Evidence Contribution

As demonstrated in the COA #000079 case, robust DIF analysis can: (a) address specific regulatory concerns (cultural applicability, recall period consistency); (b) support qualification pathway selection (limited vs broad use scenarios); (c) define clinical application boundaries; and (d) potentially exempt supplementary studies, directly accelerating the regulatory timeline.

Balanced Perspective

While DIF analysis contributes valuable supporting evidence, we emphasize that it does not override primary efficacy and safety data. Rather, DIF ensures that PRO endpoints meet "fit-for-purpose" standards, enhancing the overall evidentiary package without becoming a gatekeeping requirement.

Conclusion

DIF represents a critical psychometric safeguard for ensuring measurement fairness and validity in PFDD. As this review demonstrates, DIF is not merely a technical statistical issue but a regulatory-impacting factor that directly affects the reliability of PRO data used in drug approval decisions.

The evidence presented reveals three key insights. First, DIF detection methods have evolved from traditional approaches (Mantel-Haenszel, logistic regression) to sophisticated modern techniques (IRT-based methods, hybrid approaches, machine learning algorithms), each with distinct advantages and limitations that must be carefully considered in specific research contexts. Second, regulatory bodies, particularly the FDA and EMA, increasingly recognize DIF analysis as essential evidence supporting COA tool qualification, as exemplified by the PROMIS® Oncology Physical Function Scale case where DIF analysis directly facilitated regulatory approval. Third, despite methodological advances, significant challenges persist, including multidimensional DIF detection, handling missing data, and establishing unified global regulatory standards.

Moving forward, we recommend three actionable strategies: (1) incorporate DIF control procedures into clinical trial protocols at the design stage; (2) submit comprehensive PRO measurement bias analyses as dedicated statistical appendices in regulatory submissions; and (3) develop standardized DIF analysis guidelines aligned across global regulatory frameworks. Addressing these priorities will strengthen DIF's role in advancing precise, patient-centered drug development while acknowledging that DIF remains a supplementary technical consideration rather than a primary driver of drug approval decisions.

Abbreviations

DIF, Differential Item Functioning; PFDD, Patient-focused drug development; COA, Clinical outcome assessment; PRO, Patient-reported outcome; FDA, U.S. Food and Drug Administration; R&D, Research and development; CDE, China Center for drug evaluation; EMA, European Medicines Agency; MH, Mantel-Haenszel; IRT, Item Response Theory; CTT, Classical Test Theory; ICCs, item characteristic curves; 1PL, 1-Parameter Logistic Model; 2PL, Two-Parameter Logistic Model; 3PL, Three-Parameter Logistic Model; TBI: traumatic brain injury; SF-36: Short Form-36; RP, physical health; RE, emotional problems; QoL, Quality of Life.

Data Sharing Statement

The datasets used and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

National Natural Science Foundation of China (82173615 and 82373682). Commissioned Project by China Center for Food and Drug International Exchange: "Research on Clinical Trial Statistical Methodology for Patient-Centered Drug Development".

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Crane PK, Gibbons LE, Narasimhalu K, et al. Rapid detection of differential item functioning in assessments of health-related quality of life: the functional assessment of cancer therapy. *Qual Life Res*. 2007;16(1):101–114. doi:10.1007/s11136-006-0035-7
2. Osterlind SJ, Everson HT. Differential item functioning. In: *Item Response Theory*. Thousand Oaks, CA: Sage; 2009.
3. Zumbo BD. A handbook on the theory and methods of Differential Item Functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (Ordinal) item scores. Ottawa National Defense Headquarters; 1999.
4. Schatschneider C, Lane KL, Oakes WP, et al. The student risk screening scale: exploring dimensionality and differential item functioning. *Educ Assess*. 2014;19(3):185–203. doi:10.1080/10627197.2014.934608
5. Wu X, Shu Y, Zheng Y, et al. Recent advances in nanomedicine: cutting-edge research on nano-PROTAC delivery systems for cancer therapy. *Pharmaceutics*. 2010;17(8):30.
6. Penton H, Dayson C, Hulme C, et al. An investigation of age-related differential item functioning in the EQ-5D-5L using item response theory and logistic regression. *Value Health*. 2022;25(9):1566–1574. doi:10.1016/j.jval.2022.03.009
7. Smit EB, Bouwstra H, van der Wouden JC, et al. Development of a patient-reported outcomes measurement information system (PROMIS®) short form for measuring physical function in geriatric rehabilitation patients. *Qual Life Res*. 2020;29(9):2563–2572. doi:10.1007/s11136-020-02506-5
8. Eremenco S, Pease S, Mann S, et al. Patient-Reported Outcome (PRO) consortium translation process: consensus development of updated best practices. *J Patient-Report Outcomes*. 2017;2(1):12. doi:10.1186/s41687-018-0037-6
9. Cao K, Quan X, Hou Y. From the formation of conceptual framework to regulatory decision-making: considerations for the developments of patient-reported outcome instruments. *Drug Design Develop Ther*. 2024;18:5759–5771. doi:10.2147/DDDT.S490289
10. Ameer B. Patient-reported outcomes: listening for what is most important in clinical care and patient-focused drug development. *J Clin Pharmacol*. 2021;61(7):845–847. doi:10.1002/jcph.1867
11. Oehrlein EM, Perfetto EM, Love TR, et al. Patient-reported outcome measures in the Food and Drug Administration Pilot Compendium: meeting today's standards for patient engagement in development? *Value Health*. 2018;21(8):967–972. doi:10.1016/j.jval.2018.01.004
12. Fiero MH, Roydhouse JK, Vallejo J, et al. US Food and Drug Administration review of statistical analysis of patient-reported outcomes in lung cancer clinical trials approved between January, 2008, and December, 2017. *Lancet Oncol*. 2019;20(10):e582–e589. doi:10.1016/S1470-2045(19)30335-3
13. Gnanasakthy A, Barrett A, Evans E, et al. A review of patient-reported outcomes labeling for oncology drugs approved by the FDA and the EMA (2012-2016). *Value Health*. 2019;22(2):203–209. doi:10.1016/j.jval.2018.09.2842
14. Coles TM, Lin L, Weinfurt K, et al. Do PRO measures function the same way for all individuals with heart failure? *J Card Fail*. 2023;29(2):210–216. doi:10.1016/j.cardfail.2022.05.017
15. Teresi JA, Wang C, Kleinman M, et al. Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS(R)) measures: methods, challenges, advances, and future directions. *Psychometrika*. 2021;86(3):674–711. doi:10.1007/s11336-021-09775-0
16. Carle AC, Mara CA. Differential item functioning in patient reported outcomes research. *Dev Med Child Neurol*. 2016;58(11):1100–1101. doi:10.1111/dmcn.13165
17. FDA. Patient-focused drug development: incorporating clinical outcome assessments into endpoints for regulatory decision-making. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-incorporating-clinical-outcome-assessments-endpoints-regulatory. Accessed May 21, 2026.
18. FDA. Patient-focused drug development: selecting, developing, or modifying fit-for-purpose clinical outcome assessments. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-selecting-developing-or-modifying-fit-purpose-clinical-outcome. Accessed May 21, 2026.
19. FDA. Qualification process for drug development tools. Available from: https://www.fda.gov/media/133511/download. Accessed May 21, 2026.
20. Agency E M. Reflection paper on patient experience data. EMA/CHMP/PRAC/148869/2025. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-patient-experience-data_en.pdf. Accessed May 21, 2026.
21. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). ICH harmonised tripartite guideline: statistical principles for clinical trials (E9). Available from: https://database.ich.org/sites/default/files/E9_Guideline.pdf. Accessed May 21, 2026.

22. Holland PW, Wainer H. Differential item functioning. *Int Encyclopedia Educ*. 1995;11(7):36–44.
23. Camilli G, Shepard L. Methods for identifying biased test items; 1994.
24. Teresi JA, Fleishman JA. Differential item functioning and health assessment. *Qual Life Res*. 2007;16(1 Supplement):33–42. doi:10.1007/s11136-007-9184-6
25. Dorans NJ, Holland PW. DIF detection and description: mantel-Haenszel and standardizatioN1,2. ETS Research Report Series; 2014.
26. Roussos L, Stout W. A multidimensionality-based DIF analysis paradigm. *Appl Psychol Measure*. 2016;20(4):355–371. doi:10.1177/014662169602000404
27. Traub R. Book review: applications of item response theory to practical testing problems. *Appl Psychol Measure*. 1981;5:539–543. doi:10.1177/014662168100500412
28. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. *Contemp Sociol*. 1991;21(2).
29. Ackerman TA. Item Response Theory: parameter Estimation Techniques. *J Am Statis Assoc*. 1993;88(422):707–708. doi:10.2307/2290371
30. Embretson S, Reise S. Item response theory for psychologists; 2000.
31. Mantel NJ, Haenszel WH. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*. 1959;22(4):719–748.
32. Shealy R, Stout W. A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*. 1993;58(58–2):159–194. doi:10.1007/BF02294572
33. Lee S. Detecting differential item functioning using the logistic regression procedure in small samples. *Appl Psychol Measure*. 2017;23117167.
34. Shimizu Y, Zumbo B. A logistic regression for differential item functioning primer. *JLTA Journal Kiyo*. 2005;7:110–124. doi:10.20622/jltaj.7.0_110
35. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. The SAGE Encyclopedia of Research Design; 1981.
36. Bond TG, Yan Z, Heene M. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 4th ed. New York: Routledge; 2020.
37. Glencross M. *Rating Scale Analysis*. Stata Users Group; 2009.
38. Chalmers RP. mirt: a multidimensional item response theory package for the R Environment. *J Stat Software*. 2012;48. doi:10.18637/jss.v048.i06
39. Huelmann T, Debelak R, Strobl C. A comparison of aggregation rules for selecting anchor items in multigroup DIF analysis. *J Educ Measure*. 2019;4.
40. Strobl C, Kopf J, Zeileis A. A new method for detecting differential item functioning in the Rasch model. *Psychometrika*. 2015;80(2):289–316. doi:10.1007/s11336-013-9388-3
41. Teresi JA, Ramirez M, Lai J, et al. Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. *Psychol Sci Q*. 2008;50(4):538.
42. Teresi JA, Ocepek-Welikson K, Kleinman M, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. *Psychol Sci Q*. 2009;51(2):148–180.
43. EMA. Appendix 2 to the guideline on the evaluation ofanticancer medicinal products in man. Available from: https://www.ema.europa.eu/en/documents/other/appendix-2-guideline-evaluation-anticancer-medicinal-products-man_en.pdf. Accessed May 21, 2026.
44. Reeve BB, Hays RD, Bjorner JB, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care*. 2007;45(5 Suppl 1):S22–S31. doi:10.1097/01.mlr.0000250483.85507.04
45. Teresi JA, Jones RN. Methodological issues in examining measurement equivalence in patient reported outcomes measures: methods overview to the two-part series, "measurement equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) short forms". *Psychol Test Assess Model*. 2016;58(1).
46. Svetina Valdivia D, Huang S, Botter P. Detecting differential item functioning in presence of multilevel data: do methods accounting for multilevel data structure make a DIFference? *Front Educ*. 2024;9(9):1389165. doi:10.3389/feduc.2024.1389165
47. Bauer DJ. A more general model for testing measurement invariance and differential item functioning. *Psychol Methods*. 2017;22(3):507–526. doi:10.1037/met0000077
48. Storf M, Garmisch-Partenkirchen A. The impact of FDA and EMA guidances regarding Patient Reported Outcomes (PRO) on the drug development and approval process. *J Regulatory Sci*. 2020;8(2):45–56.
49. FDA. Patient-focused drug development: selecting, developing, or modifying fit-for-purpose clinical outcome assessments. Available from: https://www.fda.gov/media/159500/download. Accessed May 21, 2026.
50. Del Carmen Aguilar-Díaz F, Foster Page LA, Thomson NM, et al. Differential item functioning of the Spanish version of the Child Perceptions Questionnaire. *J Investig Clin Dentistry*. 2013;4(1):34–38. doi:10.1111/j.2041-1626.2012.00132.x
51. Tavakol M, Stewart C, Sharpe CC. Ensuring fairness in assessment in health professions education: rapid analysis tools to detect differential item functioning across groups. *Int J Med Educ*. 2024;15:80–83. doi:10.5116/ijme.6694.de69
52. Ikeanumba chukwuemeka OS. Differential item functioning detection methods: an overview. *Int J Res Publ Rev*. 2024. doi:10.55248/GENGPI.5.0224.0505
53. Benítez I, Vijver FVD, Padilla JL. A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociol Methods Res*. 2019;(1):1030601385.
54. Ayilara OF, Sajobi TT, Barclay R, et al. A comparison of methods to address item non-response when testing for differential item functioning in multidimensional patient-reported outcome measures. Quality of life research: an international journal of quality of life aspects of treatment. *Care Rehabil*. 2022;31(9):2837–2848. doi:10.1007/s11136-022-03129-8
55. Sébille V, Dubuy Y, Feuillet F, et al. Does differential item functioning jeopardize the comparability of health-related quality of life assessment between patients and proxies in patients with moderate-to-severe traumatic brain injury? *Neurocrit Care*. 2023;39(2):339–347. doi:10.1007/s12028-023-01705-5
56. Schwartz CE, Stark RB, Borowiec K, et al. No impact of Asian ethnicity on EORTC QLQ-C30 scores: group differences and differential item functioning in paroxysmal nocturnal hemoglobinuria. *Health Quality Life Outcomes*. 2021;19(1):228. doi:10.1186/s12955-021-01860-3
57. Gurdil H, Demir E. The use of multidimensional item response theory estimations in controlling differential item functioning. *Measurement*. 2024;1–14.
58. Amaechi CE, Onah FE. Detection of uniform and non-uniform gender differential item functioning in economics multiple choice standardized test in Nigeria. *J Nigerian Acad Educ*. 2020;2(15):224–233.
59. Karadavut T. Characterizing the latent classes in a mixture IRT model using DIF. *Appl Measure Educ*. 2021;34(4):301–311. doi:10.1080/08957347.2021.1987900

60. Doan M, Atar B. Comparing differential item functioning based on multilevel mixture item response theory, mixture item response theory and manifest groups. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*. 2024;15(2):120–137. doi:10.21031/epod.1457880
61. Diaz E, Brooks G, Johanson G. Detecting differential item functioning: item response theory methods versus the mantel-haenszel procedure. *Int J Assessment Tools Educ*. 2021;8(2):376–393. doi:10.21449/ijate.730141
62. Guo H, Lu R, Johnson MS, et al. Alternative methods for item parameter estimation: from CTT to IRT. ETS Research Report Series; 2022.
63. Gorter R, Fox JP, Riet GT, et al. Latent growth modeling of IRT versus CTT measured longitudinal latent variables. *Statist Methods Med Res*. 2019;29(4):720281243.
64. Paek I, Holland P. A note on statistical hypothesis testing based on log transformation of the Mantel-Haenszel common odds ratio for differential item functioning classification. *Psychometrika*. 2015;80(2):406–411. doi:10.1007/s11336-013-9394-5
65. Mollazehi M, Abdel-Salam ASG. Understanding the alternative Mantel-Haenszel statistic: factors affecting its robustness to detect non-uniform DIF. *Commun Statistics - Theory Methods*. 2025;54(4):1135–1159. doi:10.1080/03610926.2024.2330668
66. Shi -Y-Y. Comparative Study on Functional Difference Detection Methods for Cognitive Assessment Items. Jiangxi Normal University; 2024. doi:10.27178/d.cnki.gjxsu.2024.001004.
67. Fidalgo AM. A new approach for differential item functioning detection using Mantel-Haenszel methods. The GMHDIF program. *Spanish J Psychol*. 2011;14(2):1018–1022. doi:10.5209/rev_sjop.2011.v14.n2.47
68. Liu I, Suesse T, Harvey S, et al. Generalized Mantel-Haenszel estimators for simultaneous differential item functioning tests. *Educ Psychol Measure*. 2023;83(5):1007–1032. doi:10.1177/00131644221128341
69. Uttaro T. Influences on the Mantel-Haenszel chi-square in detection of differential item functioning under Rasch conditions. *Perceptual Motor Skills*. 1995;80(3 Pt 2):1071–1074. doi:10.2466/pms.1995.80.3c.1071
70. Lee S. Detecting differential item functioning using the logistic regression procedure in small samples. *Appl Psychol Meas*. 2017;41(1):30–43. doi:10.1177/0146621616668015
71. Chen H, Jin K. Applying logistic regression to detect differential item functioning in multidimensional data. *Front Psychol*. 2018;9:1302. doi:10.3389/fpsyg.2018.01302
72. Sun X, Wang S, Guo L, et al. Using a generalized logistic regression method to detect differential item functioning with multiple groups in cognitive diagnostic tests. *Appl Psychol Measure*. 2023;47(4):328–346. doi:10.1177/01466216231174559
73. Ackerman TA, Ma Y. Examining differential item functioning from a multidimensional IRT perspective. *Psychometrika*. 2024;89(1):4–41. doi:10.1007/s11336-024-09965-6
74. Cho SJ, Suh Y, Lee WY. After differential item functioning is detected: IRT item calibration and scoring in the presence of DIF. *Appl Psychol Meas*. 2016;40(8):573–591. doi:10.1177/0146621616664304
75. Walker CM, Gocer SS. Using a multidimensional irt framework to better understand Differential Item Functioning (DIF): a tale of three DIF detection procedures. *Educ Psychol Meas*. 2017;77(6):945–970. doi:10.1177/0013164416657137
76. Pattanaik S, John MT, Chung S. Assessment of differential item functioning across English and Spanish versions of the Orofacial Esthetic Scale. *J Oral Rehabil*. 2021;48(1):73–80. doi:10.1111/joor.13106
77. Lima JRS. Multidimensional IRT models for hierarchical latent structures; 2019.
78. Nielsen T, Elklit A, Brodsgaard Nielsen S, et al. Measurement of level of PTSD with the International Trauma Questionnaire (ITQ): bias and precision when using full ordinal or dichotomized items. *Eur J Psychotraumatol*. 2025;16(1):2514873. doi:10.1080/20008066.2025.2514873
79. Zhang Y, Jiang C, Jiang W, et al. Development and clinimetric validation of the Brief Brain Fog Scale (BBFS) for post-COVID cognitive symptoms. *J Psychosom Res*. 2025;198:112380. doi:10.1016/j.jpsychores.2025.112380
80. Christensen KS. Validating the 15-item stress anxiety depression scale (SAD-15) using Rasch analysis. *J Psychosom Res*. 2025;197:112349. doi:10.1016/j.jpsychores.2025.112349
81. Yesiltas G, Paek I. A log-linear modeling approach for differential item functioning detection in polytomously scored items. *Educ Psychol Meas*. 2020;80(1):145–162. doi:10.1177/0013164419853000
82. Kraus EB, Wild J, Hilbert S. Using interpretable machine learning for differential item functioning detection in psychometric tests. *Appl Psychol Meas*. 2024;48(4–5):167–186. doi:10.1177/01466216241238744
83. Zhong J, Ma H, Wang X, et al. Rasch analysis of the Chinese version of the clinically useful depression outcome scale in patients with major depressive disorder. *BMC Psychol*. 2023;11(1):218. doi:10.1186/s40359-023-01255-7
84. Yuksel S, Elhan AH, Gokmen D, et al. Analyzing differential item functioning of the Nottingham Health Profile by Mixed Rasch Model. *Turk J Phys Med Rehabil*. 2018;64(4):300–307. doi:10.5606/tftrd.2018.2796
85. Henninger M, Debelak R, Strobl C. A new stopping criterion for Rasch trees based on the Mantel-Haenszel effect size measure for differential item functioning. *Educ Psychol Meas*. 2023;83(1):181–212. doi:10.1177/00131644221077135
86. Woods CM, Harpole J. How item residual heterogeneity affects tests for differential item functioning. *Appl Psychol Meas*. 2015;39(4):251–263. doi:10.1177/0146621614561313
87. Danielsson L, Elfstrom ML, Galan HJ, et al. Measurement properties of the Swedish clinical outcomes in routine evaluation outcome measures (CORE-OM): Rasch analysis and short version for depressed and anxious out-patients in a multicultural area. *Health Qual Life Outcomes*. 2022;20 (1):30. doi:10.1186/s12955-022-01937-7
88. FDA. Clinical Outcome Assessments (COA) Qualification ProgramDDT COA #000079: PROMIS® Physical Function in OncologyQualification Plan. Available from: https://www.fda.gov/media/137966/download. Accessed May 21, 2026.
89. Teresi JA, Wang C, Kleinman M, et al. Differential item functioning analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS) measures: methods, challenges, advances, and future directions. *Psychometrika*. 2021;86:1–38. doi:10.1007/s11336-021-09748-3