American Journal of Public Health Research, 2026, Vol. 14, No. 6, 167-176
Available online at http://pubs.sciepub.com/ajphr/14/6/1
Published by Science and Education Publishing
DOI:10.12691/ajphr-14-6-1

Harmonizing Healthy Aging Metrics: Comparing IRT and Equipercentile Equating for HRS Physical Activity Scales

Peter D. Hart*

[1]Health Promotion Research, Havre, Montana, USA
[2]Kinesmetrics Lab, Tallahassee, Florida, USA
*Corresponding author: pdhart@outlook.com

Received October 01, 2026; Revised November 02, 2026; Accepted November 10, 2026

Abstract **Background:** The Health and Retirement Study (HRS) administers three core physical activity (PA) items to all respondents and two additional items to a subsample. Because the two sets of items have different category response structures, they are difficult to use in a single combined scale. **Purpose:** This study aimed to harmonize these measures by developing and validating a crosswalk linking the three core items to an expanded five-item scale using Item Response Theory (IRT) and equipercentile (EP) equating. **Methods:** The 2022 HRS dataset was first partitioned into training (N = 14,856) and independent holdout (N = 1,000) samples. Both samples were subsequently reduced by restricting analyses to respondents with valid Psychosocial Leave-Behind survey weights (N = 2,820 and N = 831, respectively). A single-group nested scales equating design was employed. Following category and scale optimization, both IRT true-score and EP equating frameworks were employed followed by cross-validation and conditional error analysis. **Results:** The response formats for all five PA items were collapsed into a common three-category structure, which demonstrated improved category functioning and model fit. Expansion of the scale from three to five items increased total test information by 20.69%. A Graded Response Model produced a true-score equating crosswalk with acceptable predictive accuracy (MAE = 1.194; RMSE = 1.388; R^2 = 0.755). A parallel EP model applied log-linear smoothing to weighted empirical distributions and generated a raw-score translation table that demonstrated superior predictive performance (MAE = 0.837; RMSE = 1.089; R^2 = 0.766). **Conclusion:** Overall, IRT provided essential psychometric justification for the optimized three-category scale, whereas EP equating yielded the most accurate score conversions with lower prediction error. The validated crosswalk enables researchers to extend the expanded five-item physical activity metric to the broader HRS cohort, increasing analytic sample size while preserving behavioral information and improving the precision of epidemiological modeling.

Keywords: *Test equating, Equipercentile equating, Physical activity, Psychometrics, Gerontology*

Cite This Article: Peter D. Hart, "Harmonizing Healthy Aging Metrics: Comparing IRT and Equipercentile Equating for HRS Physical Activity Scales." *American Journal of Public Health Research*, vol. 14, no. 6 (2026): 167-176. doi: 10.12691/ajphr-14-6-1.

1. Introduction

Physical activity (PA) is one of the most influential modifiable behaviors associated with healthy aging and the prevention of chronic disease. [1] As individuals age, maintaining functional capacity and health-related quality of life becomes increasingly influenced by modifiable behavioral factors, among which PA is particularly important. A substantial body of evidence has demonstrated that regular participation in physical activity is associated with lower risks of cardiovascular disease, type 2 diabetes, cognitive decline, disability, frailty, depression, and premature mortality among older adults. [2,3,4] Consequently, accurate measurement of physical activity behaviors has become a central concern in epidemiologic and gerontologic research.

Among large-scale aging studies, the Health and Retirement Study (HRS) serves as one of the most influential longitudinal investigations of health, retirement, and aging in the United States. [5] Sponsored by the National Institute on Aging and conducted by the University of Michigan, the HRS provides nationally representative data on adults over age 50 and has become a foundational resource for examining determinants of healthy aging. Given the extensive breadth of outcomes available within the HRS, PA measures collected within the study have frequently been used to investigate relationships between activity behaviors and cognitive functioning, chronic disease, disability progression, healthcare utilization, and mortality. [6,7,8,9]

Despite the value of the HRS PA measures, challenges remain regarding the optimal use of these variables. The standard HRS interview includes three core PA items assessing the frequency of engagement in vigorous, moderate, and light activity. In addition, respondents participating in the Psychosocial and Lifestyle Leave-Behind questionnaire (LBQ) complete supplementary activity items assessing participation in walking and sports or exercise activities. Although these additional indicators potentially provide richer behavioral information and broader coverage of the latent PA continuum, they employ response formats that differ substantially from those used in the core HRS interview. Specifically, the core PA items and the LBQ activity items are derived from different frequency-response frameworks that result in non-equivalent five-category and seven-category analytic classifications. These differences create practical and psychometric barriers to integrating all available activity information into a single unified measure.

The inability to directly combine these indicators has important methodological consequences. Researchers often rely exclusively on the three core items because they are available for the full HRS sample. However, this approach potentially sacrifices valuable information provided by walking and sports participation behaviors. Conversely, restricting analyses to respondents with complete LBQ data may reduce sample size and limit generalizability. More importantly, treating scores derived from different item sets as interchangeable without formal equating procedures may introduce measurement error and threaten comparability across studies. Similar concerns have motivated scale-linking and score-equating efforts in educational, psychological, and health outcomes research, where multiple forms or versions of an instrument must be placed onto a common metric. [10,11,12]

Score equating methodologies provide a potential solution to this problem. Among the most widely used approaches are Item Response Theory (IRT)-based equating and equipercentile (EP) equating. IRT equating leverages a latent trait framework, modeling the relationship between observed responses and an underlying construct through item parameters and trait estimates. [13,14] Because IRT methods operate on an underlying latent continuum, they are often viewed as theoretically attractive and can facilitate cross-form comparisons even when item sets differ substantially. Furthermore, IRT-based equating provides insight into item discrimination, threshold functioning, reliability, and measurement precision across trait levels. [13]

In contrast, EP equating adopts a distributional perspective by linking scores that correspond to equivalent percentile ranks across two measures.[10,15] Rather than relying on assumptions regarding latent trait structure, EP methods use empirical score distributions to establish score correspondences. Modern implementations frequently incorporate smoothing procedures to reduce the influence of sampling irregularities and sparse score frequencies, improving the stability of score conversions. [15,16] Although EP equating is often simpler to implement and may provide strong predictive performance in practice, it offers fewer psychometric insights into the underlying measurement properties of the scale being equated.

The selection between IRT and EP equating approaches is therefore not always straightforward. Previous research comparing equating methods has suggested that performance frequently depends on score distributions, sample characteristics, test length, and the extent to which model assumptions are satisfied. [17,18,19] In many applications, IRT methods provide stronger theoretical justification for score linking because they explicitly model the latent construct. However, empirical equating approaches may outperform parametric methods when observed score distributions contain meaningful information not fully captured by a latent trait model. Consequently, evaluating both approaches within the same measurement context can provide valuable evidence regarding the optimal strategy for scale harmonization.

Before equating can be meaningfully conducted, however, the underlying response categories must exhibit adequate psychometric functioning. Response option structures that contain redundant, poorly differentiated, or infrequently endorsed categories can lead to unstable parameter estimation, reduced measurement precision, and diminished interpretability. [20,21,22,23] Category collapse procedures are therefore commonly employed when empirical evidence suggests that adjacent response options are not functioning as distinct levels of the latent construct. Within the HRS activity measures, the coexistence of five- and seven-category response formats raises the possibility that category optimization may be necessary prior to scale integration.

Beyond harmonization, expanding the PA instrument from three to five indicators may also provide meaningful gains in measurement precision. Within an IRT framework, the addition of psychometrically sound items generally increases test information and reduces the standard error of measurement (SEM) across portions of the latent continuum. [13] If walking and sports participation items contribute unique behavioral information beyond the traditional vigorous, moderate, and light activity indicators, then incorporating these measures may improve the assessment of PA among older adults and enhance the predictive utility of HRS-based epidemiological analyses.

Accordingly, the purpose of this study was to develop and validate a crosswalk linking the standard HRS three-item activity scale with an expanded five-item PA measure incorporating walking and sports participation indicators. Using data from the 2022 HRS and the LBQ cohort, the approach first evaluated and optimized category structures across all PA items. Next, the psychometric characteristics of the resulting scale were examined using IRT methods and the incremental measurement benefits associated with expanding the instrument quantified. Finally, the two competing equating frameworks, IRT true-score equating and log-linear smoothed EP equating, were compared using independent holdout validation and conditional error analyses. Given the relatively brief length of both PA scales and the heterogeneous activity profiles present within a nationally representative aging population, differences in equating performance between the two approaches were considered an empirical question. Nevertheless, both methods were expected to produce usable score conversions suitable for harmonization purposes.

2. Methods

Data Source and Study Sample

Data were obtained from the 2022 wave of the HRS, a nationally representative longitudinal survey of U.S. adults aged 50 years and older that is sponsored by the National Institute on Aging and administered by the University of Michigan. [5] The HRS collects extensive information on health behaviors, chronic conditions, functional status, psychosocial characteristics, and aging-related outcomes. The present study focused on five PA indicators. Three core HRS items assessed the frequency of participation in vigorous, moderate, and light activity and were available for all respondents. Two additional PA indicators assessing walking and participation in sports or exercise were collected through the HRS LBQ. Because these supplemental items were administered only to a subsample and employed a different frequency-response architecture, direct integration into a unified PA scale required score harmonization. The full 2022 HRS dataset was randomly partitioned into a model-development training sample (N = 14,856) and an independent validation sample (N = 1,000). Application of the LBQ survey weight variable restricted analyses to valid participants within the LBQ subsample, yielding effective training and holdout validation samples of N = 2,820 and N = 831, respectively. For scale optimization and item calibration procedures requiring complete responses across all five PA indicators, a synchronized complete-case dataset (N = 3,301) was constructed using listwise deletion to ensure identical respondent matrices across competing models.

Physical Activity (PA) Measures

The three core HRS PA items assessed participation in vigorous, moderate, and light activity. Supplemental LBQ items assessed walking and participation in sports or exercise. Prior to scale construction, all five indicators were evaluated in their original frequency-response formats using IRT category diagnostics. Visual inspection of category response curves (CRCs) revealed substantial category redundancy, threshold crowding, and poor separation among several intermediate response options, particularly for the 7-category LBQ items. Consequently, adjacent categories were collapsed to create a common 3-category response structure across all five indicators. For the three core HRS PA indicators (vigorous, moderate, and light activity), the original analytic classifications were collapsed into three categories representing: (0) hardly ever or never, (1) intermediate activity frequencies ranging from 1 to 3 times per month through more than once per month, and (2) daily participation. For the LBQ walking and sports/exercise indicators, the original seven-category frequency classifications were similarly reduced to three categories representing: (0) never, (1) intermediate participation frequencies ranging from not recently through more than once per week, and (2) daily participation. The resulting coding harmonized response architectures across all five indicators while preserving a theoretically meaningful progression from inactivity to daily participation. Category optimization was guided by principles of rating scale functioning and category effectiveness described by Linacre, Andrich, and related psychometric frameworks. [20,21,22,23] The resulting 3-category indicators were coded uniformly and used throughout all subsequent analyses.

Scale Optimization and Incremental Measurement Evaluation

To determine whether incorporation of the two LBQ items improved measurement precision, a baseline 3-item instrument consisting of vigorous, moderate, and light PA was compared with an expanded 5-item instrument that additionally included walking and sports/exercise participation. Both scales were calibrated on the identical complete-case sample (N = 3,301) to ensure a common latent trait metric. Test Information Functions (TIFs) were estimated over a standardized latent continuum ranging from $\theta = -4$ to $\theta = +4$. Total measurement information was quantified by integrating TIF values across the latent range. Because information and SEM exhibit an inverse relationship ($SEM = 1/\sqrt{I[\theta]}$), increases in total information were interpreted as gains in measurement precision.[13]

IRT Model Selection and Psychometric Evaluation

A series of competing unidimensional polytomous IRT models were estimated to identify the optimal measurement framework for the harmonized PA items. Candidate models included: Graded Response Model (GRM), [24] Generalized Partial Credit Model (GPCM), [25] Partial Credit Model (PCM), [26] Rating Scale Model (RSM), [22] and Graded Rating Scale Model (GRSM). [27] Model estimation was conducted using the mirt package in R. [28] Relative model fit was assessed using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Sample-Adjusted BIC (SABIC), with lower values indicating superior fit. Absolute fit was evaluated using the full-information Root Mean Square Error of Approximation (RMSEA), where values below 0.05 indicated close fit. [29]

Following identification of the optimal model, item discrimination and threshold parameters were estimated on the traditional IRT metric. Item-level fit was evaluated using Orlando and Thissen's S- X^2 statistic together with associated RMSEA values. [30] Local independence was assessed using Chen and Thissen's standardized local dependence (LD) residual correlations. [31] Person-fit analyses were conducted using Drasgow's standardized log-likelihood index (Zh), with values less than −3.0 indicating substantial response-pattern aberrance. [32] Internal consistency was evaluated using Cronbach's α, while latent-score reliability was evaluated using marginal reliability coefficients ($r_{xx'}$) and empirical reliability (ρ_θ) derived from expected a posteriori (EAP) scoring. [13]

Survey-Weighted IRT Calibration

After selection of the optimal psychometric model, a survey-weighted calibration was conducted using the LBQ analytic weights to obtain nationally representative latent trait estimates. Sampling weights were incorporated directly into model estimation to ensure correspondence between calibrated parameters and the HRS target population. Survey-weighted discrimination and threshold parameters served as the fixed reference metric for all subsequent IRT equating procedures.

Equating Design

A single-group nested-scales design was employed to link the core 3-item PA scale with the expanded 5-item instrument. [10,11] Because the three core items were

contained within the larger 5-item form and both scales were administered to the same respondents, the design satisfied the assumptions required for direct score linking and equating. Two independent equating frameworks were evaluated: (1) IRT true-score equating and (2) EP equating.

Item Response Theory (IRT) True-Score Equating

IRT equating was performed using the survey-weighted GRM. First, latent physical activity trait scores (θ) were estimated using EAP scoring procedures. [13] Raw scores from the 3-item instrument were then linked to their corresponding latent trait locations. Expected raw scores on the 5-item instrument were subsequently calculated by passing each latent trait estimate through the calibrated item characteristic functions of the expanded instrument. This approach produced a true-score crosswalk translating each possible 3-item summed score into an expected 5-item score on the common latent metric. [10,13]

Equipercentile (EP) Equating

A parallel nonparametric EP equating procedure was conducted using survey-weighted observed score distributions. EP equating links scores possessing identical cumulative percentile ranks across two score distributions. [10,15] Survey-weighted cumulative distributions were first computed separately for the 3-item and 5-item scales. To reduce sampling irregularities and sparse-cell instability, observed score frequencies were smoothed using a log-linear Poisson regression model incorporating linear, quadratic, and cubic score terms. [15] The resulting theoretical percentile distributions were then used to identify score pairs with equivalent percentile standing, generating a nonparametric score-conversion table linking the two PA instruments.

Cross-Validation and Predictive Accuracy

Both equating frameworks were evaluated using the independent holdout sample (N = 831). For each participant, observed 5-item scores were compared with scores predicted by the IRT and EP crosswalks. Predictive performance was quantified using weighted mean absolute error (MAE), weighted root mean squared error (RMSE), and coefficient of determination (R^2). Lower MAE and RMSE values indicated greater predictive accuracy. To assess performance across the score continuum, conditional RMSE values were calculated separately for each possible 3-item summed score category. Conditional error profiling allowed direct comparison of framework-specific performance at lower, middle, and upper activity ranges, thereby identifying regions where either IRT or EP equating demonstrated local advantages.

Sensitivity Analyses and Software Verification

To evaluate computational reproducibility, all major analytic stages were independently replicated in both R (Version 4.4.2) and SAS Version 9.4. These analyses included category optimization, incremental information testing, IRT calibration, log-linear smoothing, and equating procedures. Concordance of parameter estimates, crosswalk conversions, and fit statistics across software environments was used as evidence of numerical stability and robustness.

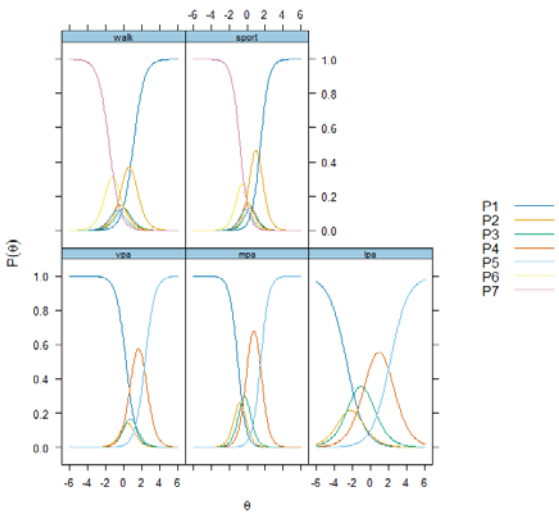
Statistical Software

All primary analyses were conducted in R (Version 4.4.2). Psychometric modeling was performed using the *mirt* package, while descriptive reliability analyses, graphical diagnostics, and equating procedures were conducted using base R and supporting statistical packages. Parallel verification analyses were conducted using SAS Version 9.4.

3. Results

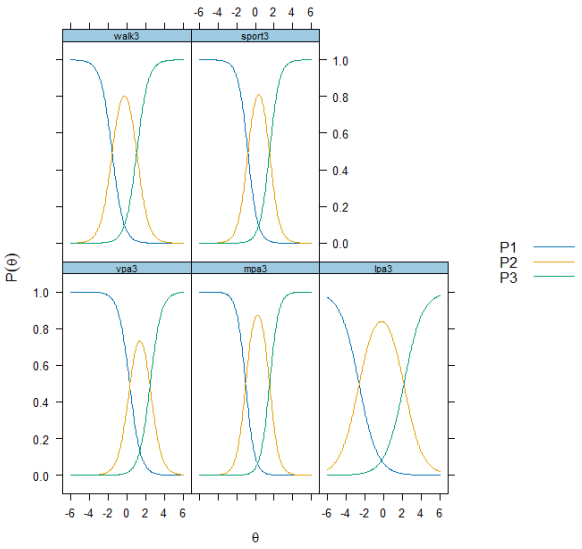
Category Optimization

Initial evaluation of the five physical activity items in their original response formats revealed substantial category redundancy and threshold crowding (Figure 1). Visual inspection of the CRCs indicated that many intermediate response options failed to function as distinct response categories. The two supplemental LBQ items, originally represented using a seven-category frequency framework, exhibited severe compression among middle response options, with several categories failing to achieve a dominant probability peak across the latent physical activity continuum. Similar patterns were observed among the core vigorous, moderate, and light activity items, where adjacent middle categories demonstrated extensive overlap and limited discriminative value. Following category consolidation, all five indicators were rescored into a common three-category structure. The optimized response architecture demonstrated substantially improved category functioning (Figure 2). Each retained response category exhibited a distinct probability peak and orderly threshold progression, indicating that respondents could meaningfully distinguish among the resulting activity-frequency classifications. The harmonized three-category framework eliminated the majority of redundant response behavior and established a common measurement foundation for subsequent scale development and equating procedures.



Note. Category response curves for the original five physical activity indicators prior to category optimization. The vigorous (vpa), moderate (mpa), and light (lpa) activity items used their original five-category response structures, whereas walking (walk) and sports/exercise participation (sport) retained their original seven-category response structures. Curves depict the probability of endorsing each response category across the latent physical activity continuum (θ). Overlapping or poorly differentiated curves indicate category redundancy and threshold crowding.

Figure 1. Category Response Curves (CRCs) for the Original HRS Physical Activity Response Structures (N = 3,301)



Note. Category response curves following category optimization and harmonization. All five indicators were collapsed to a common three-category response structure. The resulting CRCs demonstrate improved category separation and orderly threshold progression across the latent physical activity continuum (θ).

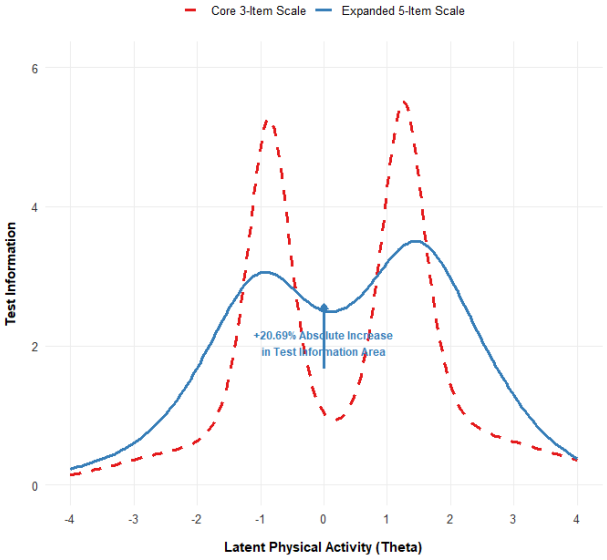
Figure 2. Category Response Curves (CRCs) for the Optimized Three-Category Physical Activity Indicators (*N* = 3,301)

Scale Optimization and Incremental Measurement Precision

The psychometric value of expanding the traditional three-item HRS PA scale was examined using a synchronized complete-case sample (*N* = 3,301). Comparison of TIFs indicated that the baseline three-item scale generated a total information area of 322.82, whereas inclusion of the walking and sports/exercise items increased total information to 389.62. This represented a 20.69% increase in measurement information across the latent physical activity continuum. As illustrated in Figure 3, the three-item instrument concentrated information within two relatively narrow regions of the trait continuum, producing a pronounced deficit in measurement precision near the population mean (θ = 0). In contrast, the expanded five-item scale filled this central information gap and provided more uniform precision across a broader range of physical activity levels. These findings provided psychometric justification for expanding the instrument prior to equating.

IRT Model Selection

Five competing polytomous IRT models were evaluated using the optimized three-category indicators (Table 1). The Graded Response Model (GRM) provided the strongest overall fit to the data, producing the lowest AIC (24,941.25), BIC (25,032.78), SABIC (24,985.12), and RMSEA (0.036) values among all candidate models. The Generalized Partial Credit Model (GPCM) demonstrated slightly poorer fit, whereas the constrained Rasch-family models (PCM, RSM, and GRSM) consistently produced inferior fit statistics. These results indicated that the PA items differed meaningfully in their discriminative capacities and were better represented by a model allowing item-specific slope parameters. Consequently, the GRM was selected for all subsequent IRT analyses.



Note. Comparison of test information across the latent physical activity continuum for the baseline three-item scale (vigorous, moderate, and light activity) and the expanded five-item scale (including walking and sports/exercise participation). Information functions were estimated using identical complete-case samples to ensure metric comparability. Total information area increased from 322.82 to 389.62, representing a 20.69% improvement in measurement precision.

Figure 3. Test Information Functions for the Core 3-Item and Expanded 5-Item Physical Activity Scales (*N* = 3,301)

Item Parameter Estimates and Psychometric Diagnostics

Table 2 presents the parameter estimates from the final GRM calibration. All five indicators demonstrated acceptable to strong discrimination parameters (*a* = 1.018-2.104). Moderate activity (mpa3) exhibited the highest discrimination (*a* = 2.104), followed by sports/exercise (sport3) participation (*a* = 1.937), whereas light activity (lpa3) demonstrated the lowest discrimination (*a* = 1.018). Threshold estimates revealed a logical ordering of activity intensity, with light activity (lpa3) functioning at the lower end of the latent continuum and vigorous activity (vpa3) targeting the highest levels of PA.

Table 1. Comparison of Polytomous IRT Model Fit Statistics (*N* = 3,301)

Model	*LL*	*AIC*	*BIC*	*SABIC*	*RMSEA*	*Parms*
Graded Response (GRM)	-12455.62	24941.25	25032.78	24985.12	0.0363	15
Generalized Partial Credit (GPCM)	-12474.92	24979.84	25071.37	25023.71	0.037	15
Partial Credit (PCM)	-12540.49	25102.99	25170.11	25135.16	0.0389	11
Graded Rating Scale (GRSM)	-12624.31	25270.61	25337.74	25302.78	0.0416	11
Rating Scale (RSM)	-12715.27	25444.53	25487.25	25465.01	0.0439	7

Note: AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; SABIC = sample-adjusted Bayesian Information Criterion; RMSEA = root mean square error of approximation. Lower values for AIC, BIC, SABIC, and RMSEA indicate superior fit.

Table 2. Graded Response Model (GRM) Item Parameter Estimates for the Harmonized Physical Activity (PA) Scale (*N* = 3,301)

Item	PA Frequency Description	Slope / Discrimination (*a*)	Threshold 1 (*b1*)	Threshold 2 (*b2*)
vpa3	Vigorous PA	1.703	0.278	2.474
mpa3	Moderate PA	2.104	-1.037	1.518
lpa3	Light PA	1.018	-2.638	2.152
walk3	Walk for 20+ Minutes	1.67	-1.594	1.055
sport3	Sports or Exercises	1.937	-0.803	1.52

Note. Parameters are reported on the traditional IRT metric. These estimates served as the calibration anchors for subsequent IRT true-score equating procedures.

Table 3. Item-Fit Statistics for the Final Graded Response Model (GRM) (*N* = 3,301)

Item	$S\text{-}X^2$	*Df* $S\text{-}X^2$	*RMSEA* $S\text{-}X^2$	*p*
vpa3	178.94	9	0.076	< 0.001
mpa3	51.752	8	0.041	< 0.001
lpa3	78.644	10	0.046	< 0.001
walk3	90.774	9	0.052	< 0.001
sport3	84.041	9	0.05	< 0.001

Note. $r_{xx'}$ = 0.730. $r_{xx'}$ = marginal reliability coefficient. $S\text{-}X^2$ denotes Orlando and Thissen's specialized item-fit chi-square statistic based on observed score subpopulations.[30] Standalone item fit is assessed via the associated RMSEA $S\text{-}X^2$, where values < 0.06 indicate adequate fit.

Table 4. Survey-Weighted GRM Item Parameter Estimates (*N* = 2,820)

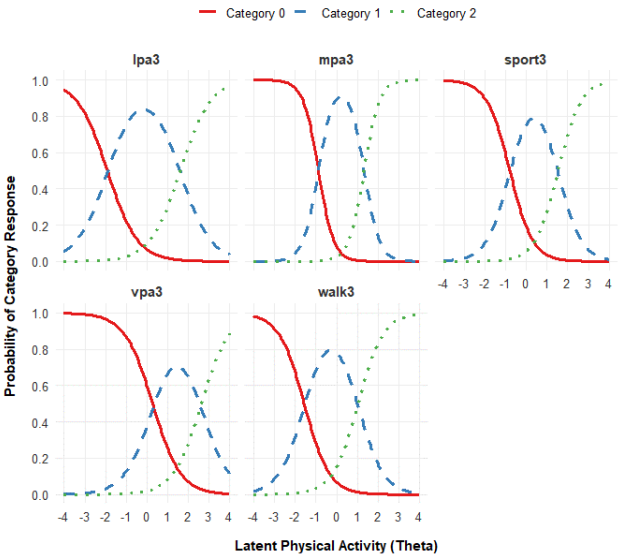
Item	Item Content Description	Discrimination (*a*)	Threshold 1 (b_1)	Threshold 2 (b_2)
vpa3	Vigorous PA	1.49	0.28	2.63
mpa3	Moderate PA	2.74	-0.88	1.33
lpa3	Light PA	1.37	-1.91	1.62
walk3	Walk for 20+ Minutes	1.66	-1.57	1.04
sport3	Sports or Exercises	1.82	-0.80	1.54

Note. Latent trait parameters calibrated using complex survey weights. Total scale Marginal Reliability $r_{xx'}$ = 0.730; Empirical Reliability ρ_θ = 0.729. Classical internal consistency Cronbach's α = 0.70.

Item-level diagnostics further supported model adequacy (Table 3). Four of the five indicators demonstrated $S\text{-}X^2$ RMSEA values below 0.06, indicating good standalone item fit. Although vigorous activity (vpa3) exhibited a slightly elevated RMSEA value (0.076), overall fit remained acceptable for a large behavioral data set. Pairwise local dependence diagnostics provided no evidence of meaningful violations of local independence, and person-fit analyses identified only 47 respondents (1.42%) with highly aberrant response patterns. The final scale demonstrated acceptable marginal reliability ($r_{xx'}$ = 0.730).

To obtain nationally representative parameter estimates, the final GRM was recalibrated using LBQ survey weights (*N* = 2,820). Survey-weighted item parameters are displayed in Table 4. Discrimination parameters ranged from 1.373 to 2.737, with moderate activity (mpa3) again emerging as the most informative indicator. Reliability estimates remained stable across weighted and unweighted calibrations ($r_{xx'}$ = 0.730; ρ_θ = 0.729; Cronbach's α = 0.70), supporting the robustness of the final measurement framework.

Figure 4 presents the final survey-weighted CRCs. Collectively, the curves demonstrate orderly category transitions and broad coverage of the latent physical activity spectrum, confirming that the expanded five-item instrument provides meaningful discrimination across a wide range of activity levels.



Note. Survey-weighted category response curves derived from the final Graded Response Model. Curves depict the probability of endorsing each response category as a function of latent physical activity (θ). The orderly progression of categories supports the psychometric adequacy of the harmonized three-category response framework and demonstrates broad coverage across the latent activity continuum.

Figure 4. Final Survey-Weighted Category Response Curves (CRCs) for the Five-Item Graded Response Model (*N* = 2,820)

Development of Equating Crosswalks

Using the survey-weighted GRM parameters, an IRT true-score equating crosswalk was developed linking the three-item HRS scale to the expanded five-item metric (Table 5). For each possible three-item summed score, latent trait estimates (θ) were converted into expected raw scores on the expanded instrument. Expected five-item scores ranged from 1.58 for a three-item score of 0 to 9.23 for a three-item score of 6. The resulting crosswalk provided a theoretically grounded translation framework anchored on the common latent trait continuum.

A parallel EP equating procedure was then performed using survey-weighted observed score distributions (Table 6). To stabilize percentile estimates, distributions were smoothed using cubic log-linear Poisson regression. The resulting crosswalk translated three-item summed scores into five-item equated scores through percentile-rank matching. The final EP translation table produced score conversions ranging from 1 to 10 across the seven possible three-item raw score levels.

Table 5. Survey-Weighted IRT True-Score Equating Crosswalk (*N* = 2,820)

3-Item Score	Theta	Expected 5-Item (Raw)	Expected 5-Item (Integer)	Theoretical Pct (%)	Empirical Pct (%)	PA Level
0	-1.56	1.58	2	5.99	5.64	Very Low
1	-0.99	2.76	3	16.14	19.06	Low
2	-0.17	5.01	5	43.28	48.37	Below Average
3	0.38	6.50	6	64.62	84.82	Average
4	0.94	7.81	8	82.73	93.91	Above Average
5	1.54	8.74	9	93.79	98.39	High
6	2.06	9.23	9	98.02	99.40	Very High

Note. Crosswalk coordinates are derived parametrically from the finalized survey-weighted GRM. Expected 5-Item (Integer) represents the rounded, true-score translation mapping recommended for discrete regressions. Latent trait coordinates (θ) are scaled to a standard population metric with a mean of 0 and a standard deviation of 1. PA level descriptors represent approximate locations on the latent PA continuum derived from the survey-weighted GRM and should not be interpreted as clinical or guideline-based PA classifications.

Table 6. Log-Linear Smoothed Equipercentile (EP) Equating Crosswalk (*N* = 2,820)

3-Item Score	Equated 5-Item Score	Empirical Pct (%)	Theoretical Pct (%)	PA Level
0	1	5.68	4.36	Very Low
1	2	19.18	20.73	Low
2	4	48.66	51.43	Below Average
3	6	85.34	80.84	Average
4	7	94.48	95.53	Above Average
5	9	98.99	99.43	High
6	10	100	100	Very High

Note. Crosswalk translation values are derived non-parametrically by directly matching empirical percentile ranks across the two scales. Theoretical Pct (%) represents the smoothed distribution profiles optimized via a cubic log-linear Poisson regression to eliminate localized sampling artifacts and stabilize conversion steps. PA level descriptors represent relative standing within the harmonized PA distribution and should not be interpreted as clinical or guideline-based PA classifications.

Table 7. Holdout Validation Statistics Comparing IRT and Equipercentile Equating Frameworks (N = 831)

Metric	EP	IRT
MAE	0.837	1.194
RMSE	1.089	1.388
R^2	0.766	0.755

Note. MAE = mean absolute error; RMSE = root mean squared error; R^2 = coefficient of determination. Statistics were calculated using survey weights in the independent holdout validation sample. Lower MAE and RMSE values indicate superior predictive performance.

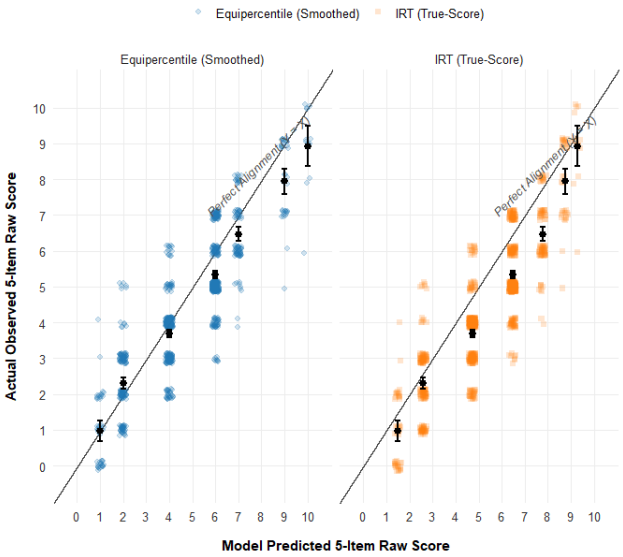
Holdout Cross-Validation

Both equating frameworks were evaluated using an independent holdout sample (N = 831). Validation results are shown in Table 7. The EP crosswalk demonstrated superior predictive performance overall, producing lower mean absolute error (MAE = 0.837) and root mean squared error (RMSE = 1.089) values than the IRT framework (MAE = 1.194; RMSE = 1.388). The EP approach also yielded a slightly higher coefficient of determination (R^2 = 0.766) relative to the IRT model (R^2 = 0.755).

Visual inspection of observed-versus-predicted score relationships further supported these findings (Figure 5). Predicted scores from both methods closely tracked observed values across the activity spectrum; however, EP predictions aligned more closely with the identity line throughout the low and middle score ranges where the majority of respondents were concentrated. Confidence interval estimates surrounding group means remained narrow, indicating stable prediction performance across score levels.

Conditional Error Analysis

To better understand score-specific performance, conditional RMSE values were calculated for each possible three-item summed score (Table 8). Across scores 0 through 4, the EP framework consistently produced lower error than the IRT framework, with the largest advantage occurring at a three-item score of 4 (RMSE = 1.008 vs. 1.530). These findings indicate that EP equating provided substantially greater precision across the lower and middle portions of the physical activity continuum.



Note. Side-by-side validation plots comparing observed five-item physical activity scores with scores predicted by the equipercentile (EP) and Item Response Theory (IRT) equating frameworks. Translucent points represent individual observations with jitter applied to improve visualization of score density. Solid black points represent group-level mean observed scores for each predicted score value, with vertical error bars indicating 95% confidence intervals. The diagonal reference line denotes perfect agreement between predicted and observed scores.

Figure 5. Cross-Validation of Equipercentile and IRT Equating Models Using an Independent Holdout Sample (N = 831).

Table 8. Survey-Adjusted Conditional RMSE Values Across the Core 3-Item Physical Activity Score Range (N = 831)

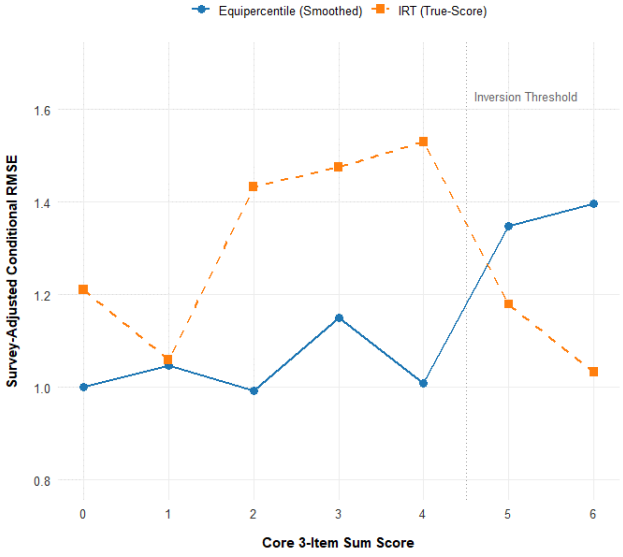
Core 3-Item Score	EP Conditional RMSE	IRT Conditional RMSE	Delta (EP vs. IRT)	Locally Optimal Framework
0	1.00	1.21	-0.21	EP Dominant
1	1.05	1.06	-0.01	EP Dominant
2	0.99	1.43	-0.44	EP Dominant
3	1.15	1.48	-0.33	EP Dominant
4	1.01	1.53	-0.52	EP Dominant
5	1.35	1.18	0.17	IRT Dominant
6	1.40	1.03	0.36	IRT Dominant

Note. Overall IRT: MAE = 1.194; RMSE = 1.388; R^2 = 0.755. Overall EP: MAE = 0.837; RMSE = 1.089; R^2 = 0.766. Lower RMSE values indicate greater local predictive accuracy.

At the highest activity levels, however, the pattern reversed. For scores of 5 and 6, the IRT framework produced lower conditional RMSE values than the EP approach, suggesting improved stability near the upper boundary of the score distribution. Figure 6 illustrates this crossover effect, with the conditional error trajectories intersecting between scores 4 and 5. Thus, while IRT equating exhibited advantages at extreme activity levels, EP equating provided superior overall performance across the range most heavily populated by older adults.

Sensitivity Analyses

Parallel analyses conducted independently in SAS 9.4 and R 4.4.2 produced virtually identical parameter estimates, equating tables, and model fit statistics. The consistency of findings across software environments supports the computational stability and reproducibility of the equating procedures and resulting crosswalks.



Note. Survey-adjusted conditional root mean squared error (RMSE) values calculated separately for each possible core three-item physical activity score. Lower values indicate greater predictive accuracy. The vertical reference line identifies the approximate score region where predictive superiority shifts from equipercentile equating to IRT true-score equating. Conditional accuracy favored equipercentile equating across the lower and middle portions of the score distribution, whereas IRT demonstrated lower conditional error near the upper boundary.

Figure 6. Conditional Root Mean Squared Error Profiles for Equipercentile and IRT Equating Frameworks (N = 831)

4. Discussion

The purpose of this study was to develop and validate a harmonized PA metric linking the standard three-item HRS PA scale to an expanded five-item instrument incorporating walking and sports/exercise participation. Using a nationally representative sample of older U.S. adults, response-category functioning, psychometric value of instrument expansion, and comparison of two competing equating frameworks were evaluated. Several important findings emerged. First, substantial category redundancy was observed across the original response formats. The CRCs demonstrated that many intermediate response options failed to function as psychometrically distinct levels of the latent PA construct, particularly among the seven-category supplemental items. Collapsing all indicators into a common three-category structure substantially improved category separation, threshold ordering, and response interpretability. These findings suggest that older adults may not reliably distinguish among numerous intermediate activity frequency categories, and that a more parsimonious response structure may provide a cleaner representation of underlying PA behavior. Importantly, the optimized three-category framework established a common measurement architecture that enabled integration of the core and supplemental HRS activity indicators.

Second, expansion of the traditional three-item scale provided meaningful psychometric benefits. The addition of walking and sports/exercise participation increased total test information by more than 20% and improved measurement coverage across the latent continuum.

Although the original three-item instrument captured broad variation in physical activity, it exhibited a notable deficit in precision near the population mean. The expanded five-item scale effectively filled this gap, resulting in more balanced and consistent measurement across the range of activity levels typically observed among older adults. From a practical perspective, these findings suggest that walking and sports/exercise behaviors contribute unique information beyond the traditional vigorous, moderate, and light activity indicators and should be incorporated whenever available. Third, the psychometric evaluation strongly supported the GRM as the most appropriate measurement framework for the harmonized scale. The GRM consistently outperformed alternative polytomous IRT models and produced stable item parameters, acceptable reliability estimates, and satisfactory item-level fit. Moderate activity emerged as the most discriminating indicator of latent activity status, whereas vigorous activity primarily differentiated highly active individuals near the upper end of the continuum. Collectively, these results support the interpretation of the five-item instrument as a coherent, unidimensional measure of PA among older adults.

The primary objective of the study, however, was the development and evaluation of score-linking procedures. Both the IRT and EP frameworks produced viable score conversion tables that successfully linked the two scales. Nonetheless, important differences emerged when the methods were subjected to independent cross-validation. Overall predictive accuracy favored the EP framework, which demonstrated lower MAE and RMSE values and a slightly higher coefficient of determination than the IRT approach. These findings indicate that direct percentile-rank matching provided a closer approximation of observed score distributions in the validation sample. The superior overall performance of the EP approach likely reflects several characteristics of the present application. First, both instruments were relatively short, containing only three and five indicators, respectively. Under such conditions, observed score distributions may contain information that is not fully captured by a latent trait model. Second, the majority of respondents clustered within the middle range of the physical activity continuum. Because equipercentile equating directly aligns empirical score distributions, it naturally maximizes accuracy within the regions where most observations occur. This advantage was evident in the conditional error analyses, where EP equating consistently outperformed IRT across the low-to-middle score range.

Despite the stronger overall performance of EP equating, the IRT framework demonstrated a notable advantage at the highest activity levels. Near the upper boundary of the score distribution, conditional error was consistently lower for IRT than for EP equating. This pattern is consistent with the theoretical strengths of latent trait modeling at score extremes. Whereas equipercentile equating relies on sparse percentile information in the distribution tails, IRT utilizes the continuous latent trait continuum and can therefore produce more stable estimates when observed data become increasingly limited. The conditional error crossover observed in the present study highlights an important tradeoff between the two approaches: EP provides superior population-level prediction, whereas IRT may offer advantages for highly active subgroups located at the extremes of the distribution. A particularly important contribution of this study is the creation of a validated crosswalk that allows researchers to recover information from respondents who completed only the core HRS PA items. Historically, investigators have faced a choice between restricting analyses to the smaller LBQ sample or relying exclusively on the less informative three-item measure. The present crosswalk offers a practical alternative by enabling harmonization across these measurement forms while preserving a considerable amount of the additional behavioral information contained in the expanded scale. Consequently, future HRS investigations examining aging, disability, cognition, mortality, and other health outcomes may be able to utilize more information than previously possible without sacrificing comparability.

This study possesses several notable strengths. The analyses were conducted using a nationally representative sample of older U.S. adults and incorporated HRS survey weights to ensure population-level relevance. An independent holdout sample was used for validation, providing a more rigorous assessment of predictive performance than would be possible using calibration data alone. Multiple competing IRT frameworks were systematically evaluated before model selection, and both parametric and nonparametric equating approaches were examined. Finally, all major analytical procedures were independently replicated across software environments, providing strong evidence for computational reproducibility.

Several limitations should also be acknowledged. First, the crosswalk was derived from a specific wave of the HRS and therefore reflects the activity distributions and measurement characteristics present within that cohort. Although the HRS design is highly stable, future waves may warrant additional evaluation. Second, the harmonized scale remains based on self-reported activity behaviors and is therefore subject to the recall and reporting biases inherent to self-report measurement. Third, the crosswalk was developed within a single aging population and may not generalize directly to younger populations, international cohorts, or alternative PA instruments. Finally, although conditional error analyses demonstrated favorable performance, all equating procedures involve some degree of measurement uncertainty and should be interpreted as approximations rather than exact score substitutions.

Future research should evaluate the performance of the crosswalk across multiple HRS waves and determine whether linking relationships remain stable longitudinally. Additional studies could also examine measurement invariance across demographic subgroups, including age, sex, race/ethnicity, and health status, to determine whether equating performance varies across population segments. Beyond the HRS, the methodological framework employed in this study may be useful for harmonizing PA measures across large epidemiological cohorts that utilize different item structures or response formats. Finally, the validated crosswalk developed here provides a foundation for subsequent investigations examining whether the expanded harmonized metric improves prediction of important aging outcomes relative to the traditional three-item scale.

5. Conclusions

This study successfully harmonized the HRS core PA measures with supplemental walking and sports/exercise indicators through the development of validated IRT and EP crosswalks. Psychometric analyses supported the use of a common three-category response architecture and demonstrated meaningful gains in measurement precision when the physical activity instrument was expanded from three to five items. Although both equating frameworks produced usable score conversions, log-linear smoothed EP equating yielded the strongest overall predictive performance, while IRT exhibited advantages at the upper extremes of the activity continuum. The resulting crosswalk enables researchers to integrate previously fragmented PA information, extend the expanded five-item metric to a substantially larger segment of the HRS cohort, preserve behavioral variance, and improve the precision of future studies examining healthy aging and related health outcomes.

References

[1] Bangsbo J, Blackwell J, Boraxbekk CJ, et al. Copenhagen Consensus statement 2019: physical activity and ageing. Br J Sports Med. 2019; 53(14): 856-858.

[2] Piercy KL, Troiano RP, Ballard RM, et al. The Physical Activity Guidelines for Americans. JAMA. 2018; 320(19): 2020-2028.

[3] Hart PD. Self-rated health, physical activity, measures of physical functioning and all-cause mortality among older US adults. J Phys Act Res. 2025; 10(1): 75-81.

[4] Hart PD. A healthy lifestyle index predicts all-cause mortality in older adults independent of psychological distress, chronic conditions, and functional limitations: NHIS 1997 to 2018. Am J Appl Psychol. 2025; 13(1): 17-22.

[5] Sonnega A, Faul JD, Ofstedal MB, Langa KM, Phillips JW, Weir DR. Cohort Profile: the Health and Retirement Study (HRS). Int J Epidemiol. 2014; 43(2): 576-585.

[6] Hart PD. Using structure coefficients and multivariate multiple regression to predict psychosocial outcomes with health and performance variables in older adults. Am J Public Health Res. 2026; 14(4): 103-112.

[7] Dhakal U, McLaughlin SJ, Kim S, Vivoda JM, Roberts AR, Brown JS. It's Not Too Late: Physical and Social Activity Slow Cognitive Decline in Those Living With Mild Cognitive Impairment. J Appl Gerontol. 2025; 44(11): 1843-1853.

[8] Ojijieme NG, Feng T, Chui CM, Qi X, Liu Y. Physical activity dynamically moderates the impact of multimorbidity on the trajectory of healthy aging over sixteen years. BMC Geriatr. 2024; 24(1): 565.

[9] Zhu Z, Zhou X, Chen M, et al. Interaction between physical activity and deficit-based frailty on all-cause mortality in older adults: a prospective study of five population-based cohorts. Lancet Reg Health West Pac. 2025; 66: 101780.

[10] Kolen MJ, Brennan RL. Test Equating, Scaling, and Linking: Methods and Practices. 3rd ed. Springer; 2014.

[11] von Davier AA. Statistical Models for Test Equating, Scaling, and Linking. Springer; 2011.

[12] Dorans NJ. Linking scores from multiple health outcome instruments. Qual Life Res. 2007; 16(suppl 1): 85-94.

[13] De Ayala RJ. The Theory and Practice of Item Response Theory. 2nd ed. The Guilford Press; 2022.

[14] Embretson SE, Reise SP. Item Response Theory for Psychologists. Lawrence Erlbaum Associates; 2000.

[15] Albano AD. equate: Observed-Score Linking and Equating in R. Appl Psychol Meas. 2016; 40(5): 361-362.

[16] Aiello EN, Curti B, De Luca G, et al. Convergence and equating norms between the Telephone Interview for Cognitive Status (TICS), the MMSE and the MoCA in an Italian population sample. Aging Clin Exp Res. 2025; 37(1): 154.

[17] Brossman BG, Lee WC. Observed score and true score equating procedures for multidimensional item response theory. Appl Psychol Meas. 2013; 37(6): 460-481.

[18] Kim JS, Hanson BA. Test equating under the multiple-choice model. Appl Psychol Meas. 2002; 26(3): 225-270.

[19] Livingston SA. Equating Test Scores (Without IRT). Educational Testing Service; 2004.

[20] Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas. 2002; 3(1): 85-106.

[21] Bond TG, Yan Z, Heene M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 4th ed. Routledge; 2020.

[22] Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978; 43(4): 561-573.

[23] Wright BD, Masters GN. Rating Scale Analysis. MESA Press; 1982.

[24] Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores. Psychometrika Monograph Supplement. 1969.

[25] Muraki E. A generalized partial credit model: application of an EM algorithm. Applied Psychological Measurement. 1992; 16(2): 159-176.

[26] Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982; 47(2): 149-174.

[27] Muraki E. Generalized rating scale model. Applied Psychological Measurement. 1990; 14(1): 59-71.

[28] Chalmers RP. mirt: A multidimensional item response theory package for the R environment. J Stat Softw. 2012; 48(6): 1-29.

[29] Maydeu-Olivares A. Goodness-of-fit assessment of item response theory models. Measurement. 2013; 46(9): 3646-3654.

[30] Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement. 2000; 24(1): 50-64.

[31] Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. J Educ Behav Stat. 1997; 22(3): 265-289.

[32] Drasgow F, Levine MV, Williams EA. Appropriateness measurement with polychotomous item response models and standardized indices. Br J Math Stat Psychol. 1985; 38: 67-86.

© The Author(s) 2026. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).