

HARMONIZING HEALTHY AGING METRICS: COMPARING IRT AND EQUIPERCENTILE EQUATING FOR HRS PHYSICAL ACTIVITY SCALES



Peter D. Hart, PhD
Kinesmetrics Lab
Health Promotion Research
GSA 2026 Annual Scientific Meeting



INTRODUCTION

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 scales, 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 to the active Psychosocial Leave-Behind study cohort by applying module survey weights (N = 2,820 and N = 831, respectively). A single-group nested-scales equating design was employed. The following analysis stages were conducted:

1. Category optimization (original vs. 3 category)
2. Scale optimization (3 items vs. 5 items)
3. IRT model selection (5 competing polytomous IRT models)
4. Final scale calibration to IRT metric
5. Survey-adjusted IRT calibration
6. IRT true score equating
7. EP equating with log-linear smoothing
8. Cross-validation of each framework with holdout sample
9. Analysis of conditional error metrics of each framework

All analytical phases were executed independently and in parallel using both SAS (Version 9.4) and R (Version 4.4.2).

RESULTS

The response formats for all five PA items were collapsed into a uniform three-category scale, which IRT analyses supported as superior in model fit and category functioning. The 5-tem PA scale saw a 20.7% absolute increase in measurement precision across the latent continuum as compared to the 3-item scale (listwise matched unweighted training sample, N = 3,301).

RESULTS

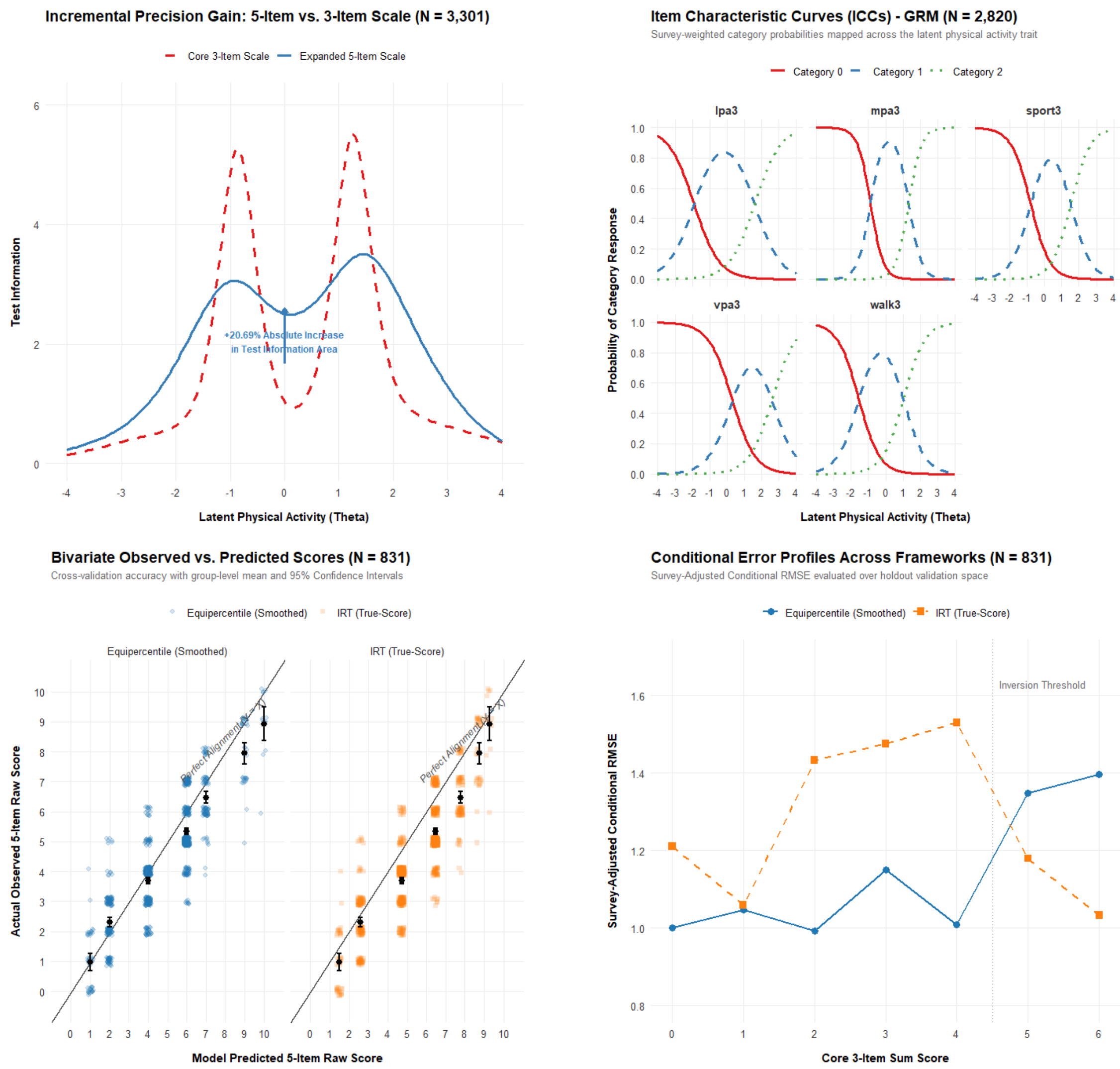
The Graded Response Model (GRM) emerged as the IRT model winner, yielding the absolute lowest values across all relative fit parameters (RMSEA = 0.036). The GRM produced a true-score raw-score crosswalk with acceptable predictive accuracy (MAE = 1.194; RMSE = 1.388; $R^2 = 0.755$).

A parallel EP model applied polynomial smoothing to weighted empirical distributions and generated a raw-score translation table that demonstrated stronger fit (MAE = 0.837; RMSE = 1.089; $R^2 = 0.766$). Finally, conditional error metrics showed that while EP error increased at the upper activity range, it was substantially lower than IRT at the lower to middle ranges.

Survey-Weighted Graded Response Model (GRM) Item Parameter Estimates (N = 2,820)

Item Identifier	Item Content Description	Discrimination (a)	Threshold 1 (b_1)	Threshold 2 (b_2)
vpa3	Vigorous Physical Activity	1.49	0.28	2.63
mpa3	Moderate Physical Activity	2.74	-0.88	1.33
lpa3	Light Physical Activity	1.37	-1.91	1.62
walk3	Walking for Exercise	1.66	-1.57	1.04
sport3	Sports/Structured Exercise	1.82	-0.80	1.54

Note. Latent trait parameters calibrated using complex survey weights. Total scale Marginal Reliability $r_{xx} = 0.73$; Empirical Reliability $\rho_\theta = 0.73$. Classical internal consistency Cronbach's $\alpha = 0.70$. Original 5-Category Items (vpa3, mpa3, lpa3): 0: Hardly Ever/ Never (4), 1: 1-3/month to > 1/month (1-3), 2: Every Day (7). Original 7-Category Items (walk3, sport3): 0: Never (7), 1: Not recently to >1/week (2-6), 2: Daily (1).



CONCLUSIONS

Overall, IRT provided essential psychometric justification for the optimized five-item three-category scale, whereas EP equating yielded the most accurate score conversions with minimized prediction bias. The validated crosswalk enables researchers to recover more PA data, preserve behavioral variance, and conduct more precise epidemiological modeling of aging outcomes within the HRS.

Future directions include applying the optimized 5-item equated scores across the full core longitudinal cohort to evaluate their predictive and construct validity with clinical, cognitive, and functional health outcomes.

Item Response Theory (IRT) True Score Equating Crosswalk (N = 2,820)

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

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.

Log-Linear Smoothed Equipercentile (EP) Equating Crosswalk (N = 2,820)

3-Item Score	Equated 5-Item Score	Empirical Pct (%)	Theoretical Pct (%)	Activity Volume Description
0	1	5.68	4.36	Minimal/Sedentary Volume
1	2	19.18	20.73	Low/Sporadic Volume
2	4	48.66	51.43	Below Average Volume
3	6	85.34	80.84	Average/Moderate Volume
4	7	94.48	95.53	Above Average Volume
5	9	98.99	99.43	High/Sustained Volume
6	10	100	100	Maximum/Robust Volume

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.

REFERENCES

1. Sonnega, A. (2025). Using Health and Retirement Study data: A guide for new users. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.
2. National Institute on Aging. The Health and Retirement Study (HRS) 2022 Core and Leave-Behind Psychosocial Survey Data. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan.
3. De Ayala RJ. The Theory and Practice of Item Response Theory. 2nd ed. The Guilford Press; 2022.
4. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29.
5. SAS Institute Inc. (2016). The IRT Procedure: SAS/STAT User's Guide. Cary, NC: SAS Institute Inc.
6. Samejima, F. (1969). Estimation of latent ability using a pattern of sequential responses. Psychometrika Monograph Supplement, 34(2), 1–97.
7. Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.). New York, NY: Springer.