NHATS Technical Paper #19

Measuring Late-Life Physical Capacity in the National Health and Aging Trends Study: An Analysis of Measurement Stability and Equivalence

September 1, 2018

Suggested Citation: Chan, Kitty S., Freedman, Vicki A, and Kasper, Judy D. 2018. Measuring Late-Life Physical Capacity in the National Health and Aging Trends Study: An Analysis of Measurement Stability and Equivalence. NHATS Technical Paper #19. Johns Hopkins University School of Public Health. Available at www.NHATS.org. This technical paper was prepared with funding from the National Institute on Aging (U01AG032947).

Introduction

Considerable progress has been made in the assessment of late-life physical function in panel studies (1). These studies now routinely incorporate both self-reported ability to carry out physical movements and standardized batteries of physical performance tests (2-4) to measure the "building blocks" that underlie functioning in routine daily tasks. Because these measures aim to evaluate underlying potential or ability, we refer to them as physical capacity measures. (5,6)

Reduced physical capacity has been associated with lower survival (7), greater activity limitations and functional decline (8-10), disability risk and duration (11), higher fall risk (12), hospitalizations (8), and greater need for caregivers (12). Physical capacity differences by age, gender, race and education also have been observed (11,13). Furthermore, there is growing interest in how physical capacity trajectories relate to other health outcomes (14) and how trajectories vary for different subgroups (15).

The National Health and Aging Trends Study (NHATS), a nationally representative panel survey, offers a unique resource for examining trajectories and demographic differences in late-life physical capacity. Multiple self-reported and performance-based physical capacity measures are assessed annually for a large sample of Medicare enrollees. Previous analyses suggest that self-reported and performance-based measures are complimentary and may be used together to form an overall physical capacity score (6). However, it remains unclear whether such a composite performs comparably across groups (measure equivalence) (16) or over time (measure stability). Furthermore, while prior studies have identified age and gender DIF for activities of daily living (ADLs) and instrumental activities of daily living (IADLs) (17,18), to our knowledge, physical capacity measures have not been investigated for measure equivalence or stability over time.

This study uses Item Response Theory (IRT) methods (19,20) to evaluate self-reported and performance-based physical capacity measures administered in NHATS. Evidence of comparable

measure performance across demographic groups and data rounds would support the validity of age, gender and race-ethnicity comparisons and trajectories based on these measures.

METHODS

Data Source

Begun in 2011, NHATS is a national panel study designed to support investigation into late-life disability. The first round of NHATS gathered information in person from a nationally representative sample of over 8,000 Medicare beneficiaries ages 65 and older, with oversamples of Black individuals and at older ages (response rate 71%). Annual re-interviews, carried out to document change over time, had response rates of 85%-90%. Further details on NHATS have been published (21).

In this analysis we focus on respondents in Rounds 1, 2 and 3 (N=7609, N=6113, N=4960, respectively). For DIF analyses by respondent demographics, we used Round 1 data, which provided the largest sample. Data from all three rounds were used to examine measure stability. The Johns Hopkins Bloomberg School of Public Health Institutional Review Board approved the NHATS protocol (#2083).

Measures

Physical Capacity

Six self-reported items ask about respondents' ability to carry out nested pairs of tasks and were scored as: able to do harder; not able to do harder but able to do easier; not able to do either. Items asked about walking 3 and 6 blocks, going up 10 and 20 stairs, carrying 10 and 20 pounds, bending over and kneeling down, reaching over head and putting a heavy book on a shelf overhead, grasping small objects and opening a sealed jar. Five performance-based assessments include: balance (tandem, semi-tandem, full tandem, and one leg stand with eyes open and then with eyes shut), usual walking

speed, rapid chair stands, handgrip, and peak airflow. Each performance test is scored from 0 (worst) to 4 (best) (22). Additional details on these measures are provided in Appendix 1.

Demographics

Age group (65-69, 70-74, 75-79, 80-84, 85+ years), sex (male, female) and race-ethnicity (African American, Hispanic, White) were based on self-report.

Statistical Analysis

We used IRT methods to test individual physical capacity measures for DIF and to compare test characteristic curves (TCC) for the overall physical capacity score by demographic groups using Round 1 data. We assessed measure stability over time by evaluating the consistency of item parameters and comparing item and test characteristic curves for Rounds 1-3. We implemented IRT analyses using Multilog 7.0 (23).

IRT Modeling

The 11 physical capacity measures previously demonstrated adequate unidimensionality (one factor based on eigenvalue >1.0 criterion, factor loadings all >0.59) for IRT modeling (6). The Graded Response Model was used to estimate item discrimination (*a*) and location (*b*) parameters (24). Higher *a* values indicate better ability to discriminate persons at different levels of the measured trait. The *b* parameters refer to the location on the capacity spectrum where the item performs best. The number of location parameters is one fewer than the number of response categories (2 for self-report; 4 for performance). The underlying physical capacity scale for these location parameters is based on 0 being equal to the group mean and each unit being equal to the sample standard deviation.

Differential Item Functioning

The presence of DIF indicates important differences in the measurement properties for an item between two groups in assessing the trait of interest. Specifically, item discrimination (*a*) DIF indicates that an item has a stronger relationship with the trait in one group than the other. Item location (*b*) DIF, on the other hand, suggests that the item is 'easier,' or more likely to be endorsed at a lower level of the trait, for one group than the other.

Item-Level Effect. We used likelihood ratio difference tests to determine whether item parameters significantly differ in the two groups being compared (e.g., male vs. female). We implemented an iterative process to identify anchor items that did not show DIF using IRTLRDIF (25). Once anchor items are determined, we evaluated DIF for the remaining items. We first tested for discrimination DIF by comparing the model that constrained the *a* parameter to be equal in the two groups with the model that specifies a separate *a* parameter for each group. If no *a* DIF was observed, we tested for location DIF by comparing the model that constrained both the *a* and the *b* parameters to be equal across groups with the model that only constrained the *a* parameter. If the *a* parameter differed significantly across groups, however, no furthering testing was performed, as the interpretation of location DIF is unclear in this situation (26.) Differences in –2*log likelihood was evaluated using the chi-square distribution. $P<0.05$ indicate significant difference, with Benjamin-Hochberg adjustment for multiple comparisons (27).

We examined the nature of DIF using item characteristic curves (ICC), which plot the probability of endorsing the item over the range of physical capacity. Group differences in these curves reveal the magnitude and direction of the DIF for the item. Non-overlapping ICCs by group indicate DIF; coincident curves reflect lack of DIF. We used the criterion of 0.1 for the standardized score (score difference by group divided by item score range) to identify meaningful DIF (28).

Scale-Level Effect. The aggregate effect of item DIF is evaluated by comparing TCC, which plots the expected score over the range of capacity, in each group. A comparison of the TCCs illustrates the magnitude and direction of scale-level DIF effect, with coincident curves indicating measure equivalence and non-overlapping curves reflecting non-equivalence. As with ICCs, we used the 0.1 criterion to identify meaningful DIF for the overall physical capacity score.

RESULTS

Sample Characteristics

In Round 1, our sample was 58% female, 68% White, with approximately 20% in each age group (Table 1). Gender and race-ethnic distributions were comparable across data rounds.

Overall Score

For the overall score, no meaningful DIF was observed by age, sex, race-ethnicity, or by rounds of data collection (Figures 1-4). With respect to age, non-overlapping curves were found for the comparisons with the two older age groups, with a maximum difference of 1.5 and 2.5 points lower for the 80-84 and 85+ age groups, respectively, vs. reference group of ages 70-74. Male respondents had higher observed scores than female respondents at the same physical capacity level, with the largest score differences (between 2 and 3 points) observed within the physical capacity range of -1 to 2 (Figure 2). However, the standardized difference for the overall score was <0.1 across the capacity range for these age and gender comparisons (i.e., maximum observed score difference < 3 given score range of 32), indicating that these differences in the overall score are not meaningful.

Individual Items

Using respondents aged 70-74 as the reference group, no meaningful DIF (standardized score difference > 0.1) was observed for any measure in the comparisons with the younger age groups (65-69 and 75-79). For age group 80-84, only performance-based balance demonstrated meaningful DIF

(within the 0.0 to 1.5 capacity range). For the oldest age group (85+), 4 of the 7 measures with statistically significant DIF also had meaningful DIF, three performance-based (i.e., balance, walk, handgrip) and one self-reported (walking) measure. Meaningful DIF observed for the performance measures occurred in the ranges from -1.0 to 2.0 for balance, 0.0 to 1.0 for walking, and 0.0 to 2.0 for handgrip. Self-reported walking demonstrated meaningful DIF only for a narrow range (-0.5 to 0.0). (See Appendices 2a-2d)

Meaningful gender DIF was observed for 6 of the 8 measures that demonstrated statistically significant DIF. However, the strongest DIF was observed for performance-based handgrip and peak airflow. For both measures, meaningful DIF was observed across a broad range of physical capacity (-2 to 4 for handgrip and -1.5 to 3.5 for peak airflow). Meaningful DIF for the self-reported measures affected a more limited capacity range. Except for the grasping item which demonstrated meaningful DIF between -3.5 to -3.0 and -1.0 to 0.5, the remaining self-reported items affecting a very narrow range (between -1.0 to 0.0 for carrying; only around 0.5 for bending and around -1.0 for reaching). Overall, eight measures perform comparably by gender for most of the physical capacity range. (See Appendices 2e).

Nine items demonstrated statistically significant DIF for Black respondents and 5 for Hispanic respondents compared to White respondents. However, few were meaningful (standardized difference ≥ 0.1) and the range of physical capacity affected was typically narrow. Specifically, meaningful DIF was observed for Black respondents between 0.5 and 1.0 of the physical capacity trait for the performance walking measure and only at -3.0 for the self-reported grasping item. For Hispanic respondents, meaningful DIF was observed in a narrow range for two self-reported items: walking (at -0.5) and reaching (at -1.0). (See Appendices 2f-2g).

Item discrimination and location parameters were generally stable across Rounds 1-3 (Table 2). Despite modest variations in parameter estimates, the item characteristic curves across the 3 rounds

were coincident (see Appendix 3) indicating that the physical capacity measures performed the same way at each round.

DISCUSSION

NHATS offers annual data on a large nationally representative cohort of older adults that can be used to study differences in late-life physical capacity by demographic groups and over time. With few exceptions, study findings support the measurement equivalence of the physical capacity measures included in NHATS across age, gender and race-ethnic groups. In particular, the overall physical capacity score did not meaningfully differ by age, gender or race-ethnicity. In addition, consistent measure parameters were observed across all three rounds of data collection, indicating measure stability over time. These findings suggest that these NHATS measures, particularly the overall score, can be used to provide valid comparisons of physical capacity by major demographic groups and for describing physical capacity trajectories.

At the item level, the strongest gender DIF was observed for two performance measures, handgrip and peak airflow, and self-reported grasping. These results suggest that the relationship between each of these measures and underlying physical capacity differs for men and women. Specifically, women with the same physical capacity as their male peers, as reflected by all 11 measures, may still demonstrate weaker grip strength. These findings are consistent with prior studies which found weaker grip strength among women compared to men after controlling for body weight (29) and even when gender differences were not observed for the strength of other muscle groups (30). Interestingly, gender DIF was also observed for self-reported ability to open a sealed jar and grasping small objects, suggesting gender differences in perception of ability to perform parallel performance-based assessments of grip strength. Overall, these results suggest that neither performance nor self-report of arm strength can serve alone as valid indicators of overall physical capacity when making

comparisons by gender. Although other measures (e.g., self-reported bending) also demonstrated statistically significant gender DIF, the differences were minor, with meaningful DIF affecting a narrow range of physical capacity. Therefore, these items can still be expected to provide valid comparisons of physical capacity between older men and women.

Age DIF was mainly observed for the oldest group (85+), with meaningful DIF for three performance-based measures (walking, balance, handgrip) and one self-reported item (walking). In all instances, persons in the oldest group with the same physical capacity as the 70-74 year olds would have lower scores. Few self-reported items demonstrated meaningful DIF even for the oldest group, suggesting that these self-reported measures can provide valid data for comparing physical capacity across all age groups. These self-report measures can also ensure the inclusion in these comparisons of respondents for whom performance assessments are not obtained due to environmental challenges (e.g. insufficient space for walking course) or refusal to participate.

Individual physical capacity measures are also likely to provide valid comparisons between White, African American and Hispanic older adults. Although statistically significant DIF by race-ethnicity was found, few were meaningful and the capacity range affected by meaningful DIF was narrow. Future studies can use both the overall and individual physical capacity measures for valid race-ethnic group comparisons.

Physical capacity measures have demonstrated strong associations with important health outcomes and valid comparisons across groups and description of trajectories are important for advancing our understanding of health disparities and the dynamics of function in older adults as they age and in relation to specific health events such as hospitalizations or falls. This study found that an overall score based on all 11 NHATS performance and self-reported physical capacity measures can provide valid comparisons across age, gender and race-ethnic groups. Furthermore, evidence of measure stability suggests that valid trajectories of physical capacity can be developed. With few

exceptions (e.g., handgrip and peak airflow by gender), individual self-report and performance measures also allow valid comparisons of physical capacity across groups.

In conclusion, our study found evidence of measure equivalence for an overall physical capacity score from NHATS. Together with design features of the survey, including annual administration and oversamples of understudied demographic groups, this data resource is well-suited for future studies focused on understanding how health disparities are produced and how physical capacity and function evolve over time among older adults.

References

1. Gill TM. Assessment of function and disability in longitudinal studies. *J Am Geriatr Soc.* 2010;58(Suppl 2):S308-S312.

2. Guralnik JM, Simonsick EM, Ferrucci L, et al. A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *J Gerontol Med Sci*. 1994;49:M85–M94.

3. Nagi, SZ. Some conceptual issues in disability and rehabilitation. In: Sussman, MB., editor. Sociology and Rehabilitation. Washington, D.C.: American Sociological Association; 1965

4. Guralnik JM and Ferrucci L. Assessing the building blocks of function: utilizing measures of functional limitation. *Am J Prev Med* 2003;25(3 Suppl 2):112-121.

5. Freedman VA. Adopting the ICF Language for Studying Late-life Disability : A Field of Dreams? *J Gerontol Sci*. 2009;64A(11):1172-1174.

6. Kasper JD, Chan KS, Freedman VA. Measuring Physical Capacity: An Assessment of a Composite Measure Using Self-Report and Items. *J Aging Health*. 2017;29(2):289-309.

7. Studenski S, Perera S, Patel K, et al. Gait speed and survival in older adults. *JAMA*. 2011;305(1):50-58.

8. Studenski S, Perera S, Wallace D, et al. Physical performance measures in the clinical setting . *J Am Geriatr Soc*. 2003;51(3):314-322.

9. Judge J, Schechtman K, Cress E. The relationship between physical performance measures and independence in instrumental activities of daily living: The FICSIT Group. Frailty and Injury: Cooperative Studies of Intervention Trials. *J Am Geriatr Soc*. 1996;44:1332-1341.

10. Ahacic K, Kåreholt I, Thorslund M, Parker MG. Relationships between symptoms, physical capacity and activity limitations in 1992 and 2002. *Aging Clin Exp Res*. 2007;19(3):187-193.

11. Hardy SE, Dubin JA, Holford TR, Gill TM. Transitions between States of Disability and Independence among Older Persons. *Am J Epidemiol*. 2005;161(6):575-584.

12. Montero-Odasso M, Schapira M, Soriano ER, et al. Gait velocity as a single predictor of adverse events in healthy seniors aged 75 years and older. *J Gerontol A Biol Sci Med Sci*. 2005;60(10):1304-1309.

13. Samuel LJ, Glass TA, Thorpe RJ, Szanton SL, Roth DL. Household and neighborhood conditions partially account for associations between education and physical capacity in the National Health and Aging Trends Study. *Soc Sci Med*. 2015;128:67-75.

14. Barbour KE, Lui L-Y, McCulloch CE, et al. Trajectories of Lower Extremity Physical Performance: Effects on Fractures and Mortality in Older Women. *Journals Gerontol Med Sci*. 2016;71(12):1609-1615.

15. Botoseneanu A, Allore HG, Mendes de Leon CF, Gahbauer EA, Gill TM. Sex Differences in Concomitant Trajectories of Self-Reported Disability and Measured Physical Capacity in Older Adults. *Journals Gerontol Med Sci*. 2016;71(8):1056-1062.

16. McHorney CA, Fleishman JA. Assessing and Understanding Measurement Equivalence in Health Outcome Measures: Issues for Further Quantitative and Qualitative Inquiry Health Outcome Measures Issues for Further Quantitative and Qualitative Inquiry. *Med Care*. 2006;44(11):S205-S210.

17. Fleishman JA, Spector WD, Altman BM. Impact of Differential Item Functioning on Age and Gender Differences in Functional Disability. *Journals Gerontol Soc Sci*. 2002;57B(5):S275-S284.

18. LaPlante MP. The Classic Measure of Disability in Activities of Daily Living Is Biased by Age but an Expanded IADL/ADL Measure Is Not. *Journals Gerontol Soc Sci*. 2010;65(6):720-732.

19. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications; 1991.

20. Holland PW, Wainer H. *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Assoc, Inc; 1993.

21. Kasper JD, Freedman VA. Findings from the 1st Round of the National Health and Aging Trends Study (NHATS): Introduction to a Special Issue. *J Gerontol B Psychol Sci Soc Sci* 69 (Suppl 1): S35-S41.

22. Kasper J, Freedman V, Niefeld M. *National Health and Aging Trends Study Construction of Performance-Based Summary Measures of Physical Capacity in the National Health and Aging Trends Study*. Baltimore, MD; 2012.

23. Thissen D, Chen W, Bock D. Multilog verson 7.0 for Windows. 2002.

24. Samejima F. Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika*. 1969;Monograph.

25. Thissen D. IRTLRDIF v.2.0b: Software for the computation of the statistic involved in item response theory likelihood-ratio tests for differential item functioning. 2001.

26. Thissen D, Steinberg L, Wainer H. Detection of differential item functioning using the parameters of item response models. In: Holland P, Wainer H, eds. *Differential Item Functioning*. Vol Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:67-114.

27. Thissen D, Steinberg L, Kuang D. Quick and Easy Implementation of the Benjamin-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *J Educ Behav Stat*. 2002;27(1):77-83.

28. Perkins AJ, Stump TE, Monahan PO, McHorney CA. Assessment of differential item functioning for demographic comparisons in the MOS SF-36 health survey. *Qual Life Res*. 2006;15:331-348.

29. Desrosiers J, Bravo G, Hebert R, Dutil E. Normative Data for Grip Strength of Elderly Men and Women. *Am J Occup Ther*. 1995;49(7):637-644.

30. Rice CL, Cunningham DA, Paterson DH, Rechnitzer PA. Strength in an Elderly Population. *Arch Phys Med Rehabil*. 1989;70:391-397.

Table 1. NHATS respondent characteristics, Rounds 1-3 (unweighted N)

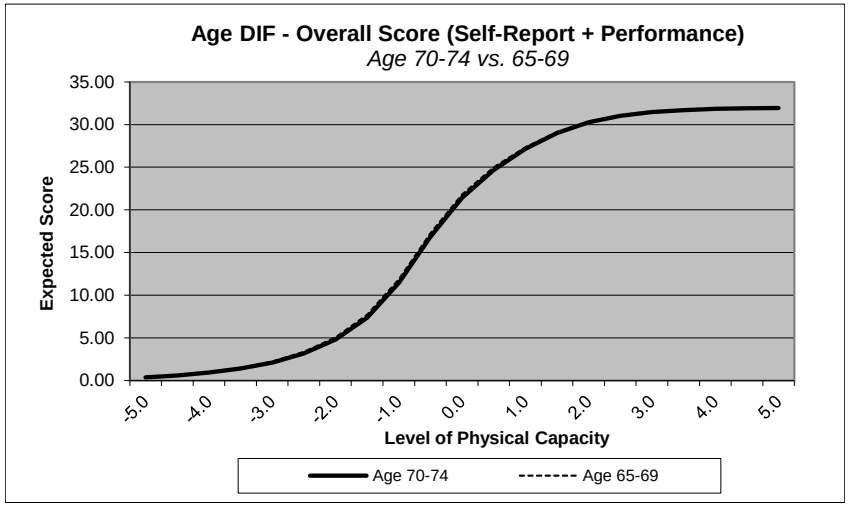
Respondent Demographics	Data Round (Year)		
	Round 1 (2011) N=7609	**Round 2 (2012) N=6113**	**Round 3 (2013) N=4960**
Age, n (%)			
65-69	1,409 (18.5)	903 (14.8)	470 (9.5)
70-74	1,579 (20.8)	1,249 (20.4)	1,115 (22.5)
75-79	1,513 (19.9)	1,259 (20.6)	1,018 (20.5)
80-84	1,505 (19.8)	1,229 (20.1)	1,025 (20.7)
85+	1,603 (21.1)	1,473 (24.1)	1,332 (26.9)
Sex, n (%)			
Female	4,438 (58.3)	3,571 (58.4)	2,931 (59.1)
Race-ethnicity, n (%)			
White	5,186 (68.2)	4,214 (68.9)	3,482 (70.2)
African American	1,662 (21.8)	1,326 (21.7)	1,047 (21.1)
Hispanic	454 (6.0)	353 (5.8)	268 (5.4)
Other	307 (4.0)	220 (3.6)	163 (3.3)

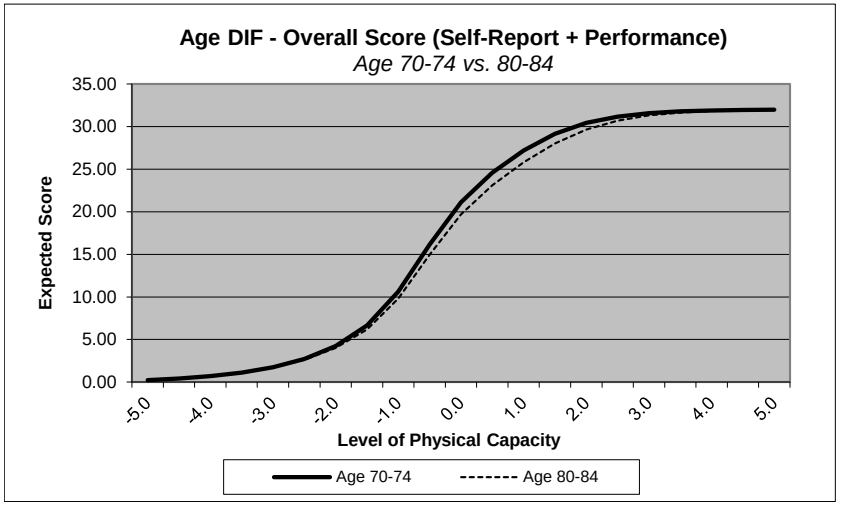
Table 2. Discrimination and Location Parameters by Data Round

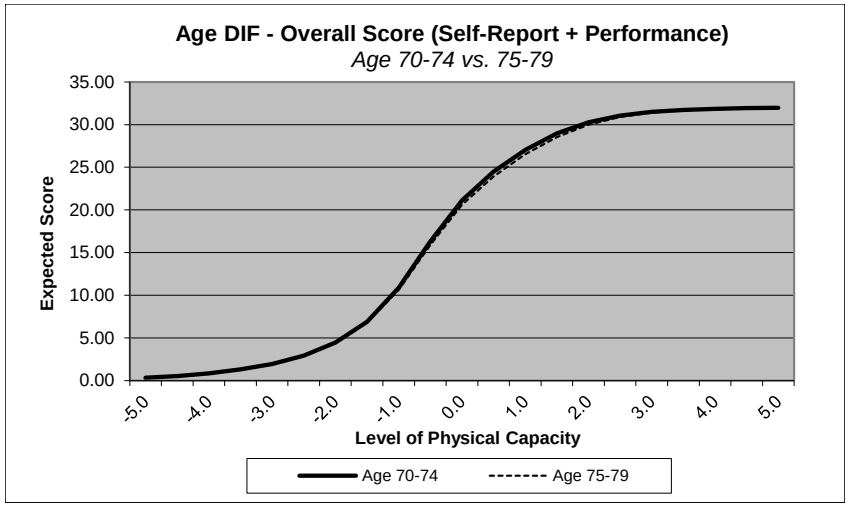
		Item Discrimination	Item Locations			
Measure	**Round**	**A**	**B1**	**B2**	**B3**	**B4**
Performance						
Balance	1	2.16	-1.34	-0.32	0.41	1.18
	2	2.34	-1.24	-0.24	0.48	1.21
	3	2.27	-1.18	-0.21	0.53	1.23
Walk speed	1	2.25	-1.48	-0.20	0.52	1.24
	2	2.55	-1.36	-0.17	0.54	1.22
	3	2.64	-1.28	-0.18	0.47	1.12
Chair stand	1	2.13	-0.71	0.09	0.68	1.34
	2	2.24	-0.65	0.07	0.63	1.33
	3	2.24	-0.66	0.02	0.54	1.22
Grip strength	1	1.37	-1.79	-0.37	0.51	1.48
	2	1.43	-1.78	-0.34	0.53	1.49
	3	1.42	-1.75	-0.32	0.50	1.48
Peak airflow	1	1.44	-3.14	-0.49	0.49	1.45
	2	1.49	-2.85	-0.49	0.49	1.42
	3	1.50	-2.76	-0.55	0.41	1.36

Self-Report						
Walking	1	4.03	-0.37	-0.12		
	2	3.86	-0.34	-0.07		
	3	4.08	-0.29	-0.02		
Climbing	1	3.76	-0.66	-0.32		
	2	3.92	-0.61	-0.26		
	3	3.73	-0.62	-0.27		
Carrying/Lifting	1	4.10	-0.68	-0.28		
	2	4.19	-0.65	-0.24		
	3	3.88	-0.65	-0.23		
Bending	1	2.35	-0.73	0.44		
	2	2.49	-0.72	0.46		
	3	2.43	-0.79	0.46		
Reaching	1	2.82	-1.08	-0.70		
	2	2.71	-1.09	-0.70		
	3	2.80	-1.09	-0.65		
Grasping	1	1.40	-2.49	-0.70		
	2	1.34	-2.61	-0.76		
	3	1.37	-2.70	-0.75		

Figure 1. Age DIF for Overall Score, Round 1







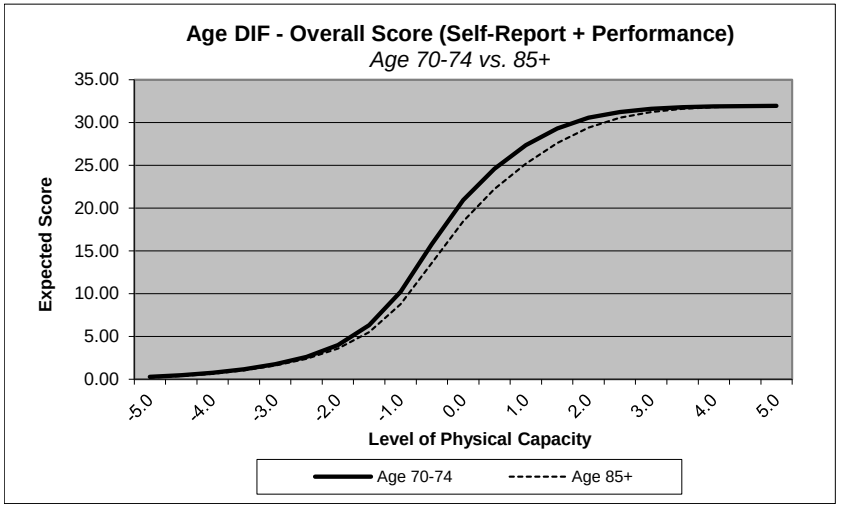


Figure 2. Gender DIF for Overall Score, Round 1

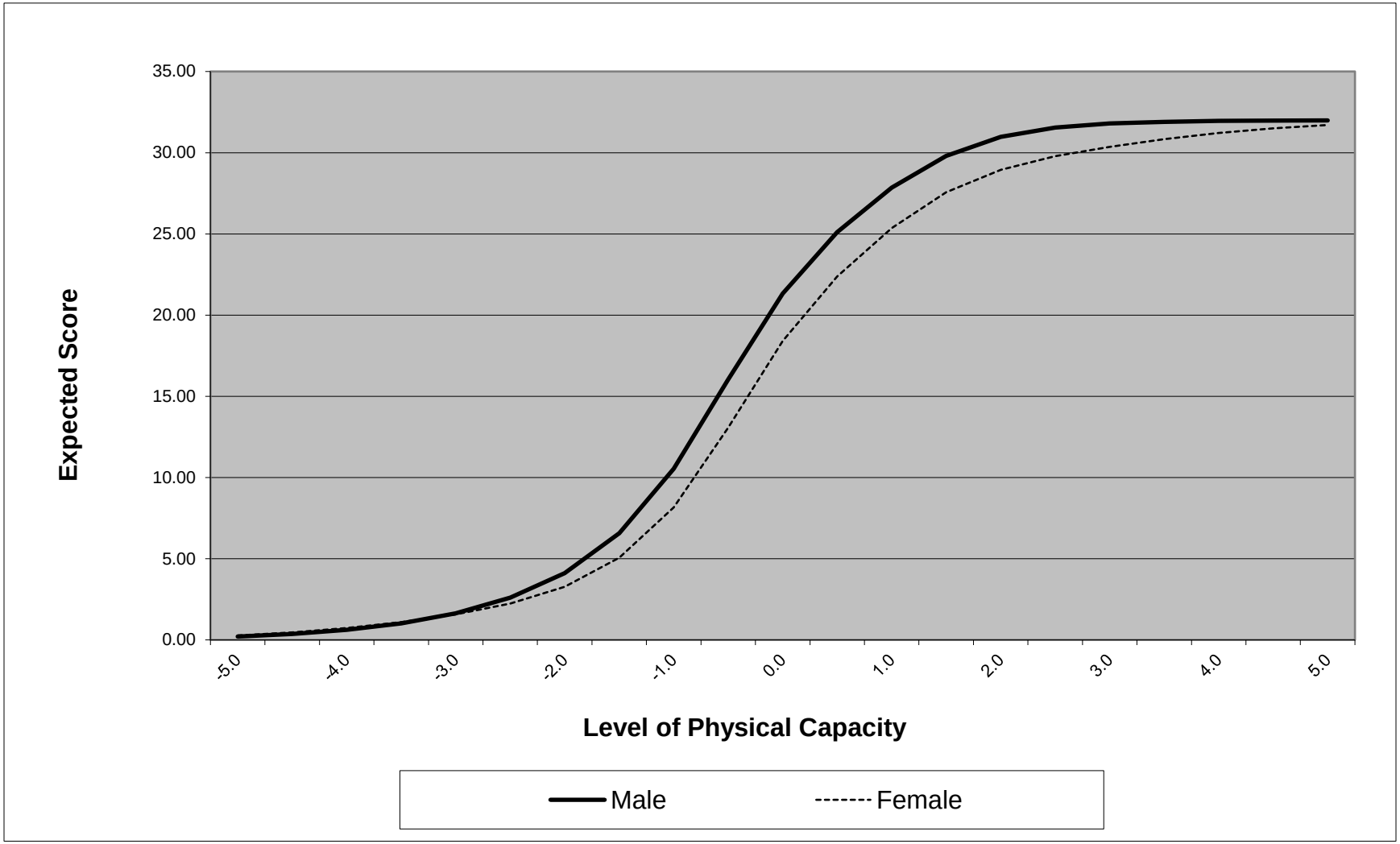
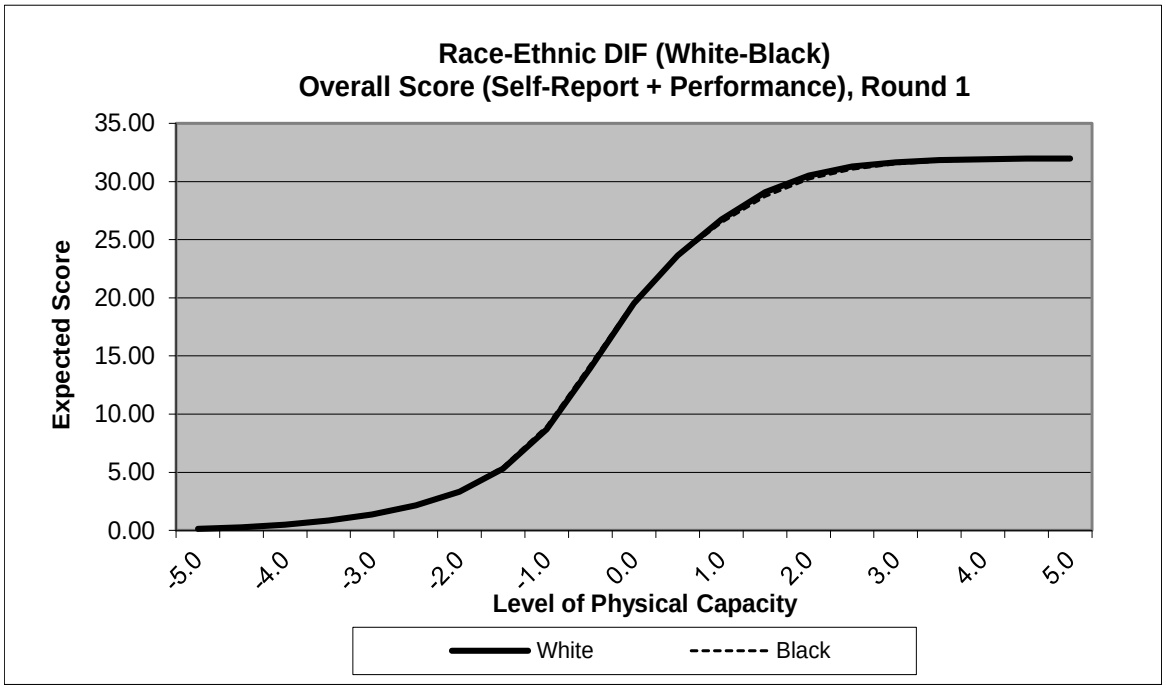


Figure 3. Race-ethnic DIF for Overall Score, Round 1



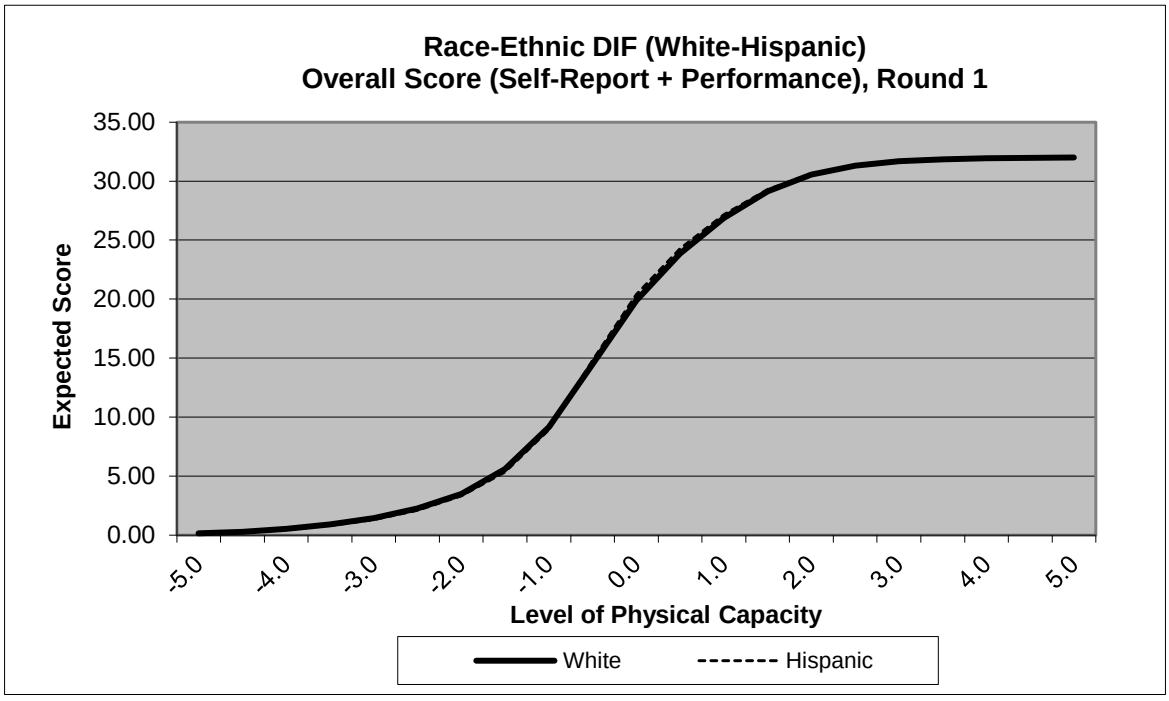


Figure 4. NHATS Overall Score for Rounds 1--3