## **Supplement A.** Supplementary methods & results

### Item Development

***Item bank:*** To form an item bank, we reviewed the literature on physical function measurements in aging populations and identified existing scales/tools designed to measure physical functioning or disability. In addition, survey questionnaires in 46 health studies/population-based cohort studies that included physical function measures were reviewed. Items identified through the process of item generation were organized according to two domains: functional limitations and disability, based on Nagi's disablement model.[1)] The functional limitations domain includes items associated with basic physical and mental actions, which serve as a foundation for performing routine tasks in daily life (e.g., ambulation, reaching, bending, and climbing). Items in the functional limitation domain were further categorized into the lower extremity and the upper extremity in accordance with linked body segments. On the other hand, the disability domain includes items associated with essential activities required for living independently and performing a socially defined role (e.g., self-care, household management, and socializing with others). Disability items were further divided into three sub-domains: activities of daily living (ADL), instrumental daily living (IADL), and social function. Items with very similar themes/contents were combined into one item. Consequently, a 144-item bank consisting of five sub-domains (lower body function, upper extremities, ADL, IADL, and social function) was constructed.

***Evaluation by experts:*** Six experts in the fields of geriatrics, geriatric nursing, psychology, and preventive medicine reviewed the item bank of 144 items and then evaluated the extent to which each of the items was relevant to physical functioning in people aged 50 years or older. The relevance was rated on a 5-point Likert scale: 1 "not relevant," 2 "slightly relevant," 3

"somewhat relevant," 4 "quite relevant," and 5 "highly relevant." To quantify the content validity of items, we computed the item-level content validity index (I-CVI), by which items key to assessing physical function in aging populations were identified. The average I-CVI of 144 items was 0.54, with 31 items with I-CVIs greater than 0.80, 66 items with I-CVIs between 0.5 and 0.8, and 47 items with I-CVIs less than 0.50. Items with a high proportion of agreement (I-CVI >0.80) were considered highly relevant, while items three or more experts rated as 'not relevant' were considered irrelevant to the construct of interest (physical functioning) or inappropriate to the context of KNHANES. Items with low I-CVI values were revised or removed after having examined the frequency of items across existing physical functioning scales, evaluating their clarity, unambiguousness, comprehensibility, redundancy with other items in the same domain, and cultural appropriateness. As a consequence, 84 items were removed, and 60 items were retained, which included 25 items of functional limitation in the lower extremity, 13 items of functional limitation in the upper extremity, six ADL items, eight IADL items, and eight items of social function.

**Scale Development**

***Focus group interview:*** Two focus group interviews (FGIs) were carried out with 16 laypersons aged 50 to 89 years in September 2020. The FGI participants reviewed the pretest questionnaire and evaluated the extent to which the pretesting items reflect physical functioning in aging populations. Each group consisted of an equal number of men and women matched by their age-specific groups of 50s, 60s, 70s, and 80s. A semi-structured interview guide was utilized to direct FGIs. First, each of the pretesting questions was posed to participants, asking each person to assess whether the posed questions captured the important situation related to the physical function relevant to their age. Participants were asked to explain tasks or activities the posed question was asking, which allowed for assessing the participants'

understanding of their meaning and intent. Also, the participants were asked to articulate their choice of response options and to confirm that the response scale differentiated the degree of each individual's functional status. Questions that were confusing and unclear due to format and wording problems were also identified. During open-ended discussions, participants were asked to suggest additional questions that could reflect physical functioning in aging populations. Finally, a professional copyediting service reviewed the grammar and wording of the questionnaire. Based on the feedback from FGI, along with professional copyediting, we refined and modified the pretest survey questionnaire.

***Pretesting the items:*** Pretesting was conducted to obtain statistical information on pretesting items by examining the psychometric measurement properties of each item. Given that a sample of at least 500 respondents is recommended for IRT model estimations,[2)] the pilot testing questionnaire was administered to a representative sample of 508 from the target population (adults aged 50 years or older) in February 2021. A stratified random sampling by age, gender, and geographic region was used to recruit participants for the pilot testing. The pilot testing questionnaire consisted of 60 pretest items and questions on socio-demographic information, health conditions, symptoms of depression, and subjective cognitive decline. Pretest items were phrased, "How much difficulty do you have in carrying out a particular activity without the help of someone else or the use of assistive devices?" and rated on a 5-point Likert scale: 4 (none), 3 (a little), 2 (some), 1 (a lot), and 0 (cannot do). Pilot-testing data were collected using computer-assisted personal interviewing.

***Item reduction analysis:*** Both item response theory (IRT) and classical test theory approaches were used to evaluate item-level psychometric properties. Item analyses were conducted to draw the optimal set of items from the pretest questionnaire. Since functional limitations (FL) and disability are separate constructs, item analysis techniques, including item calibration, exploratory factor analysis, differential item functioning (DIF), and item-total correlations,

were conducted for each of the domains separately. First, we examined the corrected item-total correlations to detect items exhibiting a weak relationship with the construct of interest. All corrected item-total correlations exceeded the accepted cutoff of 0.3,[3] which is indicative of the association between the total score and the other items within the same domain.

Before conducting IRT models, two major assumptions of the IRT approach, unidimensionality and local independence, were tested. To test the unidimensionality of items for each of the domains, we conducted factor analyses using a principal axis factoring method. More than half (54.5%) of the total FL variance was explained by the first factor, and the ratio of the first to the second largest eigenvalues was 6.75. For the disability items, 45.2% of the total disability variance was accounted for by the first factor, and the ratio of the first to the second largest eigenvalues was 4.5. Using the ratio of the "first to second eigenvalues greater than four" rule as an index of unidimensionality,[4] the assumption of unidimensionality was met for each of the domains.

Local dependence (LD) indicates excess covariance between items after controlling for the effect of the dominant factor, the underlying trait, and suggests that responses to these items may be explained by factors other than the dominant factor. Generalizations for polytomous responses of the LD statistic described by Chen & Thissen were used to identify locally dependent items.[5] A pair of items with the standardized local dependence chi-square (LD $X^2$) statistic greater than 10 was considered to fail to meet the local independence assumption.[6] There were eight pairs of FL items exhibiting local dependence, while no disability item pairs had LD $X^2$ values that exceeded 10. For each pair of items with LD, one of the items was removed. Based on the item's discriminating power and the amount of item information, a better-functioning item was selected. Consequently, eight FL items with LD were eliminated. Next, the S-$X^2$ statistic was used to identify misfit items exhibiting significant ($p<0.01$)

discrepancies between observed and expected responses.[7] Seven FL items and five disability items (12 items in total) did not fit the data well (Supplementary Tables S2 and S3).

Item discrimination parameters and item difficulty parameters were used for the item reduction process. Item parameters were calibrated for each domain, consisting of 38 FL items and 22 disability items, using a graded response model (GRM), which is an IRT model for ordinal polytomous items.[8] The discrimination parameter represents the extent to which an item correctly differentiates individuals of varying levels of physical functioning. To add to that, it is associated with the amount of information provided by an item, which determines the level of assessment precision of the item. Accordingly, the larger the discrimination parameter, the better the item differentiates between individuals with good physical functioning and those with poor physical functioning, but also with greater precision of the item measurement. The discrimination parameters for the 38 FL items ranged from 1.98 to 5.94 and the 22 disability items ranged from 1.72 to 6.88, indicating a very good discrimination power.[9]

Under the GRM framework, threshold/difficulty parameters represent the item's locations along the latent trait scale at which respondents have a 0.50 probability of endorsing a given response category and higher. The difficulty parameters for 38 FL items ranging from –4.19 to –0.15 appeared to be skewed to the left along the latent trait continuum. Likewise, the difficulty parameters for the 22 disability items ranging from –4.53 to –0.37 were located around –2.0 and –1.0 on the latent trait continuum. Six FL items and seven disability items that were located to the left end of the trait continuum were removed because these items were too easy for the target population. Furthermore, item information curves (IIC) were used in an effort to select the best set of items. Given that there is an inverse relationship between the amount of information and measurement errors, item information plays a significant role in the precision of the measurement. The area under item information curves was examined to evaluate the

amount of information the items provide for estimating a respondent's location on the latent continuum. Items that contributed little to information were removed.

Finally, DIF analyses were performed to identify items that differed in estimated item parameters by gender. Two items in the FL domain and one item in the disability domain were identified to have differential item functioning. That is, there were significant differences in estimated item difficulty parameters of men and women. Specifically, women tended to have more difficulty "lifting 5 kg" and "lifting 10 kg" than men did at the same level of latent ability, while men reported more difficulty "preparing meals" than women at the same level of latent ability.

**Analysis of IRT Assumptions**

In terms of assumption testing of factor analysis, the value of Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy was 0.91, and all KMO values for individual items were above 0.5, which means that the sample size is sufficient for factor analysis. The Bartlett Test of Sphericity revealed that correlations between items were sufficient to conduct PCA ($\chi^2$=3,707.81, *df*=300, p<0.0001). Next, the assumptions underlying the IRT framework were tested. Given that 41.8% of the total variance was explained by the first factor and that the ratio of the first and the second largest eigenvalues was 5.48, the LF scale items appeared to measure a single latent construct,[4] satisfying the unidimensionality assumption. All LD statistics were below 8, which indicates that no item pairs showed excess covariance after controlling for the underlying trait. The overall model-data fit was good ($M_2$=258.25, p=0.16; RMSEA=0.02), and the S-$X^2$ item-misfit statistics showed no misfit item at a significance value of 0.05. Lastly, the monotonicity assumption was met in that the probability of endorsing higher response options increased monotonically as the latent trait level increased.

**REFERENCES**

1. Nagi SZ. Disability concepts revisited: implications for prevention. In: Pope AM, Tarlov AR, editors. Disability in America: Toward a national agenda for prevention. Washington, DC: National Academy Press; 1991, p. 309-327.

2. Price LR. Psychometric methods: theory into practice. New York: The Guilford Press; 2017.

3. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw-Hill; 1994.

4. Lumsden J. The construction of unidimensional tests. Psychol Bull 1961; 58(2):122.

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

6. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. J Patient Rep Outcomes 2019;3:1-6.

7. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Meas 2000;24(1):50-64.

8. Samejima F. Graded response model. In van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. New York: Springer; 1997, p. 95-107.

9. Baker FB. The basics of item response theory. College Park: ERIC Clearinghouse on Assessment and Evaluation; 2001.