

Development of a Short Form of the Life Functioning Scale for Older Adults in the Korea National Health and Nutrition Examination Survey

Eunsaem Kim

Ajou University

Jinhee Kim

Ajou University

Yunhwan Lee

yhlee@ajou.ac.kr

Ajou University

Jihye Yun

Korea National Institute of Health

Jaewon Choi

Ajou University

Miji Kim

Kyung Hee University

Kyungwon Oh

Korea Disease Control and Prevention Agency

Yangha Kim

Korea Disease Control and Prevention Agency

Research Article

Keywords: Short form, Physical functioning, Item response theory, Psychometrics, Older adults, KNHANES

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11201544/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

Lengthy national health surveys impose a substantial respondent burden on older adults. This study aimed to develop and validate a short form of the 25-item Life Functioning (LF-25) scale to reduce this burden.

Methods

Two candidate short forms were generated by combining item response theory (IRT)-based item selection with qualitative content validation. Following baseline administration to 519 adults aged 50 years and older, a 10-item scale (LF-10) was selected. The psychometric properties of the LF-10 were subsequently tested in an independent sample of 263 older adults. Predictive validity was evaluated by regressing six-month health outcomes on baseline LF-10 scores in a follow-up subsample.

Results

A graded response model showed acceptable fit (RMSEA = 0.07), and confirmatory factor analysis supported a unidimensional structure. The LF-10 correlated strongly with both the original LF-25 and a criterion Physical Functioning scale. Internal consistency was excellent (McDonald's ω = 0.93), and test–retest reliability was excellent (intraclass correlation coefficient = 0.94). Baseline LF-10 scores significantly predicted physical functioning and self-rated health at the six-month follow-up.

Conclusions

The LF-10 is a valid, reliable, and efficient measure of physical functioning in older adults that halves administration time relative to the LF-25. It has been adopted by the Korean Frailty and Aging Cohort Study (2023) and the Korea National Health and Nutrition Examination Survey (2024), supporting its feasibility for national health surveillance.

Introduction

As life expectancy has increased over recent decades, healthy aging has become a key public health priority. Yet, an extended lifespan is frequently accompanied by physical decline, frailty, and disability, which substantially diminish quality of life in older adults [1]. The World Health Organization [2] defines healthy aging as the process of developing and maintaining functional ability—essentially, the capacity to perform daily activities and maintain independence. Central to this definition is the concept of intrinsic capacity—the composite of all physical and mental reserves an individual can draw upon—which, together with a person's physical and social environments, determines functional ability and quality of life. Physical functioning is particularly important because it directly reflects an older adult's overall health status and capacity to perform daily activities independently within both home and community settings. Because physical functioning is a key determinant of care needs [3], its accurate measurement is vital for monitoring health status, evaluating functional capacity, and planning long-term care in older populations.

Existing measures offer limited scope and precision for comprehensively evaluating physical function in community-dwelling older adults. Most focus on functional disability, with particular emphasis on basic and instrumental activities of daily living (ADLs and IADLs). When administered to community-dwelling older populations, ADL- and IADL-based instruments show pronounced ceiling effects [4], as their item content is

confined to low-demand self-care and everyday activities. The absence of higher-demand physical tasks (e.g., lifting heavy objects and stair climbing)—early markers that precede deterioration in basic ADLs [5]—leaves physical functioning insufficiently represented across its full range. Existing instruments also inadequately address social and leisure activities, which are central to maintaining engagement with life and well-being. Instruments with broader content coverage are therefore needed to represent physical functioning as older adults actually experience it.

To address this gap, we previously developed a physical functioning measure for adults aged 50 years and older [6] for use in the Korea National Health and Nutrition Examination Survey (KNHANES). Grounded in Nagi's disablement model [7], the original scale (LF-25) consists of 25 items: 10 items on lower body function, six on upper extremity function, six on IADLs and ADLs, and three on social functioning (Supplementary Table 2). The LF-25 showed strong psychometric properties and a unidimensional structure [6], but it is too long for KNHANES, which administers approximately 400 items within a limited timeframe [8]. Lengthy national health surveys increase response burden and cognitive fatigue, which can lead to lower response rates, greater item nonresponse, and ultimately higher measurement error and reduced data quality [9–11]. For this reason, item reduction is widely used in large national surveys. Although short forms inevitably sacrifice some measurement precision and construct breadth, a well-designed short form preserves the core construct and its validity while meaningfully reducing respondent burden [12, 13]. Several self-reported physical function measures for older adults have been shortened using item response theory (IRT), including the computer-adaptive Late-Life Function and Disability Instrument [14, 15], the Mobility Assessment Tool—short form [16], and the PROMIS Physical Function short forms [17, 18].

This study aimed to develop a short form of the LF-25 that maintains the psychometric quality of the full-length scale while substantially reducing response burden and administration time. Specifically, we sought to (1) derive candidate short forms by integrating IRT-based item selection with expert content review and cognitive interviewing; (2) select the final short form on the basis of measurement precision and administrative feasibility; and (3) evaluate its structural, concurrent, known-groups, and predictive validity, together with its internal consistency and test–retest reliability, in independent samples of Korean adults aged 50 years and older.

Methods

This study drew on five distinct samples collected between 2021 and 2023. Supplementary Fig. 1 shows the flow of participants, exclusions, and analyses across item reduction, the three validation phases, and the external replication in the 2023 Korean Frailty and Aging Cohort Study (KFACS). The Phase 2 validation sample and the Phase 3 follow-up sample were independent.

Item Reduction

For the quantitative reduction, we analyzed secondary data from the original LF-25 validation study conducted in 2021 ($n = 259$). Based on preliminary item statistics (e.g., standard deviations, factor loadings, and missing data rates), we identified redundant or underperforming items. Item selection was then guided primarily by an IRT framework [19]. To maximize the information yielded by the shortened scale, we prioritized items with high discrimination parameters and a broad, evenly spaced range of difficulty thresholds, to retain measurement precision across the latent trait (θ) continuum.

To preserve the conceptual breadth of the original LF-25, we added two qualitative steps: expert content evaluation and cognitive interviewing. Nine experts from the original development panel rated each item, and items below the

recommended thresholds for a nine-member panel (content validity index [CVI] 0.78; content validity ratio [CVR] 0.62) were removed [20]. Cognitive interviews with eight older adults then identified items that were confusing or lacked real-world relevance, and these were also removed (Supplementary Tables 3 and 4). Finally, to balance psychometric precision against comprehensive content validity, we constructed two candidate short forms. The 10-item LF scale (LF-10) prioritized psychometric strength, retaining the items that provided the most information across ability levels. The 12-item LF scale (LF-12) retained all LF-10 items and added two items recommended by the expert panel for content coverage: "reaching behind back" and "engaging in indoor leisure activities". Both candidate forms are listed in Supplementary Table 5.

Validation of Short-Form LF Scales

Validation comprised three phases: short-form selection (Phase 1), psychometric validation (Phase 2), and predictive validity testing (Phase 3).

Phase 1: Short-Form Selection

To evaluate the two candidate short forms and identify the version most suitable for KNHANES, a baseline survey was administered in April–May 2022 to 519 adults aged 50 years and older. All participants first completed the full-length LF-25 and were then randomly assigned to complete either the LF-12 (n = 259) or the LF-10 (n = 260). Following psychometric evaluation of both candidates, the LF-10 was selected as the final short form because it offered the best balance of measurement precision and administrative feasibility. Item analysis results for both candidates and the rationale for this decision are given in the Supplementary Methods. The LF-10 is provided in English and Korean in Supplementary Tables 1 and 10.

Phase 2: Psychometric Validation

An independent validation survey was conducted in July–August 2022 with 263 adults aged 50 years and older. Participants were recruited by stratified random sampling on age, sex, and geographic region, and data were collected by face-to-face computer-assisted personal interviewing (CAPI). Because confirmatory factor analysis (CFA) requires an independent sample, the factor structure found in this sample was tested in the Phase 3 follow-up sample (n = 263). All follow-up respondents completed the LF-10, regardless of which short form they had received at baseline. To examine structural validity in an older population, the CFA was replicated in an independent sample from the 2023 KFACS (n = 879). Finally, to evaluate test–retest reliability, the LF-10 was re-administered to a subsample of validation survey participants (n = 115) 10 to 14 days after their initial interview.

Phase 3: Predictive Validity

A follow-up assessment was conducted in October 2022, five to six months after baseline, with 263 participants drawn from the Phase 1 baseline sample. At follow-up, participants completed both the LF-25 and the LF-10, allowing the correlation between the two scales to be examined.

Measures and Ethical Considerations

LF-10 items ask: "How much difficulty do you have carrying out this activity without help from another person or the use of assistive devices?" Responses are rated on a 5-point scale: none (4), a little (3), some (2), a lot (1), and cannot do (0). Total scores are computed by summing item scores and multiplying by 2.5, yielding a 0–100 metric on which higher scores indicate better physical functioning. Sociodemographic and health data were also collected. Physical health was assessed by self-rated health and the number of physician-diagnosed chronic conditions. Cognitive function was assessed with a single item on subjective cognitive decline, and mental health was assessed using the Korean version of the Patient Health Questionnaire-9 (PHQ-9) [21]. Concurrent validity was examined against the Physical Functioning (PF) scale (10 items on mobility and self-care, scored 0–100) [22], the criterion measure used in the original LF-25 validation study [6]. The CAPI system automatically recorded completion time for the LF-10 in Phase 2 and for the LF-25 in Phase 3. This study was approved by the Institutional Review Board of Ajou University Hospital (AJIRB-SBR-SUR-21-677), and written informed consent was obtained from all participants before data collection.

Statistical Analysis

Data were first screened for outliers and missing values. Cases with extreme total scores ($|z| \geq 4.5$) were identified as outliers and excluded. Structural validity was examined with a recommended two-step factor-analytic strategy in independent samples [23]. First, an exploratory factor analysis (EFA) examined the underlying structure of the LF-10 in the Phase 2 sample. The number of factors was determined using Kaiser's eigenvalue criterion (eigenvalue ≥ 1) and scree plot inspection. Next, CFA tested this structure in the Phase 3 and KFACS samples. Because LF-10 items are ordinal and subject to ceiling effects, all factor analyses used the weighted least squares mean- and variance-adjusted (WLSMV) estimator with pairwise deletion. Model fit was evaluated using the comparative fit index (CFI), Tucker–Lewis index (TLI), standardized root mean square residual (SRMR), and root mean square error of approximation (RMSEA). Good model fit was defined as CFI > 0.95 , TLI > 0.95 , SRMR < 0.08 , and RMSEA < 0.06 [24].

Measurement quality was evaluated using a graded response model (GRM). Because most participants chose the top categories, the lowest categories (e.g., 'cannot do', 'a lot') had very few or no responses (Supplementary Table 6). Response categories endorsed by fewer than 2% of respondents were collapsed with the adjacent category [25], because sparse categories produce unstable threshold estimates, inflated standard errors, and convergence problems [26, 27]. This ensured a minimum frequency of roughly five responses in the lowest available category for every item. Collapsing was applied only to GRM calibration; all other analyses—descriptive statistics, total score computation, factor analyses, reliability estimation, and validity testing—used the original five-category responses. IRT parameters were estimated by marginal maximum likelihood using the Bock–Aitkin EM algorithm, which draws on all available responses. Overall GRM fit was assessed with the limited-information M_2 statistic. Because M_2 is sensitive to sample size and often rejects models with only trivial misspecification, fit was judged mainly by its associated RMSEA [28]. Item fit was assessed with the $S-X^2$ statistic [29], and local independence was evaluated using the standardized LD χ^2 statistic, with values ≥ 10 taken to indicate a possible violation [30]. Finally, differential item functioning (DIF) by sex was tested using the Wald χ^2 statistic.

Concurrent validity was assessed using Pearson correlation coefficients to examine the association of the LF-10 with the PF scale. Known-groups validity was evaluated with independent-samples t -tests and one-way ANOVAs comparing LF-10 scores across sociodemographic (age, sex, educational attainment, marital status, household income) and health groups (self-rated health, subjective cognitive decline, depression, multimorbidity). Reliability was estimated via internal consistency (McDonald's ω and Cronbach's α) and test–retest reliability (weighted kappa, percentage agreement, and intraclass correlation coefficient [ICC]). Predictive validity was tested by regressing 6-month PF scores (linear regression) and 6-month self-rated health (SRH; poor/very poor, fair,

good/very good; ordinal logistic regression) on baseline LF-10 scores. Both models adjusted for age, sex, education, household income, multimorbidity (≥ 2 chronic conditions), and the baseline value of the outcome.

Descriptive statistics, correlation and group comparisons, reliability estimation, and regression analyses were conducted in SPSS version 30.0 (IBM Corp., Armonk, NY, USA). IRT analyses were performed in IRTPRO version 6.0 (Vector Psychometric Group, Chapel Hill, NC, USA). EFA and CFA were conducted using Mplus version 8.11 (Muthén & Muthén, Los Angeles, CA, USA). All reported p -values were two-sided, with statistical significance set at $p < 0.05$. For item fit ($S-X^2$) and DIF tests, a more stringent threshold of $p < 0.01$ was applied to account for multiple comparisons.

Results

Descriptive Statistics

Of 263 individuals who completed the Phase 2 validation survey, 260 were retained for analysis after the exclusion of three outliers. Fourteen cases (5.4%) had at least one missing value, with no apparent systematic pattern. Participants ranged in age from 50 to 91 years (mean = 63.7, SD = 8.6), and 51.9% ($n = 135$) were female. Other characteristics are shown in Table 1. The mean LF-10 total score was 94.2 (SD = 12.3), ranging from 15 to 100. In total, 58.5% ($n = 152$) scored at the ceiling (100). By comparison, the KFACS sample had a mean LF-10 score of 86.2 (SD = 16.8), and 25.1% ($n = 221$) scored at the ceiling. Phase 2 response distributions are shown in Supplementary Table 6. Mean completion time was 11.68 minutes (SD = 4.35) for the LF-25 in the Phase 3 sample ($n = 263$) and 5.86 minutes (SD = 1.70) for the LF-10 in the Phase 2 validation sample ($n = 263$), a reduction of approximately 50%. Because these times came from separate samples, the difference is an approximate between-sample comparison, not a within-person comparison.

Table 1
Sample characteristic and LF-10 scores by sample characteristics ($n = 260$)

	Categories	n (%)	LF-10 score (SD)
Gender	Men	125 (48.1)	95.66 (10.91)
	Women	135 (51.9)	92.78 (13.28)
Age (y)	50–59	102 (39.2)	98.14 (5.95)
	60–69	90 (34.6)	95.58 (9.61)
	≥ 70	68 (26.2)	86.32 (17.76)
Educational level	Middle school or lower	83 (31.9)	87.35 (17.67)
	High school	129 (49.6)	97.36 (6.54)
	College or higher	48 (18.5)	97.34 (6.79)
Marital status	Married	220 (86.3)	95.09 (11.67)
	Widowed	23 (9.0)	84.89 (14.31)
	Divorced	12 (4.7)	95.21(13.55)
Monthly household income (in thousand KRW)	< 2,000	47 (18.1)	84.41 (18.67)
	2,000–3,999	90 (34.6)	95.52 (11.25)
	≥ 4,000	101 (40.8)	97.88 (5.70)
	Don't know/refused	17 (6.5)	90.74 (10.96)
Self-rated health	Very good or good	158 (60.8)	98.78 (3.88)
	Fair	85 (32.7)	91.35 (9.48)
	Poor	17 (6.5)	65.29 (25.96)
Subjective cognitive decline	Yes	30 (11.5)	85.67 (15.31)
	No	230 (88.5)	95.27 (11.38)
Depressive symptoms	Minimal or no depression ^{a)}	247 (95)	96.47 (9.24)
	Mild to moderate ^{b)}	13 (5)	69.42 (28.16)
Multimorbidity	0	119 (45.8)	97.92 (5.54)
	1	70 (26.9)	94.00(11.99)
	≥ 2	71 (27.3)	88.03 (17.33)
Note. ^{a)} PHQ-9 ≤ 4; ^{b)} PHQ-9 ≥ 5; LF-10: 10-item Life Functioning scale; SD = standard deviation; PHQ-9 = Nine-item Patient Health Questionnaire.			

Validity of the LF-10 Scale

IRT assumptions were checked before GRM calibration and were met (Supplementary Methods); all standardized LD χ^2 values were below 10. Although the M_2 statistic was significant ($M_2 = 496.59$, $df = 204$, $p < 0.001$), the RMSEA

of 0.07 indicated acceptable fit, within the 0.05–0.08 range for mediocre fit [31]. Item discrimination parameters ranged from 1.78 to 5.97, all exceeding 1.50, which indicates high discrimination [32]. The threshold parameters ranged from – 2.39 to -0.60, covering the lower-to-middle range of the latent trait (θ). The item characteristic curves (Supplementary Fig. 2) and test information curve (Supplementary Fig. 3) likewise indicated the greatest precision at the lower end of the physical functioning continuum. After adjusting the significance threshold for multiple testing ($p < 0.01$), all items showed adequate fit to the GRM. No item showed significant DIF by sex (all $p > 0.01$). Women had a lower latent mean than men ($\mu = -0.49$ vs. 0.00), indicating a true group difference rather than item bias. Table 2 summarizes the GRM and DIF results.

Table 2
Item parameters and fit statistics for the 10-item Life Functioning scale ($n = 260$)

	GRM item parameters					Item Fit ^{c)}		Uniform DIF	
	$a^a)$	$b_1^b)$	b_2	b_3	Overall Difficulty	S- χ^2	p	$\chi^2_{c/a}$ (df)	p
Q1. Standing up from an armless chair	5.97 (1.39)	-2.10 (0.22)	-1.32 (0.13)	—	-1.72	11.03	0.01	1(2)	0.59
Q2. Standing on tiptoes	5.56 (1.14)	-1.95 (0.20)	-1.54 (0.15)	-0.97 (0.11)	-1.49	24.54	0.01	3.2(3)	0.36
Q3. Bending/squatting/kneeling	3.84 (0.64)	-2.17 (0.24)	-1.47 (0.15)	-0.74 (0.10)	-1.46	16.67	0.12	1.4(3)	0.72
Q4. Walking 400m	4.52 (0.90)	-1.97 (0.21)	-1.78 (0.19)	-1.16 (0.12)	-1.64	16.77	0.05	0.9(3)	0.82
Q5. Climbing 1 floor	4.38 (0.77)	-1.91 (0.20)	-1.28 (0.13)	-0.60 (0.09)	-1.27	10.51	0.40	5.3(3)	0.15
Q6. Grasping small objects	2.81 (0.74)	-2.05 (0.27)	—	—	-2.06	2.05	0.56	0.2(1)	0.66
Q7. Lifting 5kg	2.61 (0.42)	-2.39 (0.29)	-1.95 (0.22)	-1.00 (0.12)	-1.79	19.87	0.10	4.3(3)	0.23
Q8. Bathing/showering	4.41 (1.39)	-1.91 (0.22)	—	—	-1.92	3.76	0.05	0.1(1)	0.70
Q9. Using public transportation	4.72 (1.04)	-2.25 (0.25)	-1.36 (0.14)	—	-1.81	11.6	0.04	0.3(2)	0.86
Q10. Participating in social activities	1.78 (0.42)	-2.00 (0.31)	—	—	-2.01	8.36	0.30	0(1)	0.84
Note. ^{a)}item discrimination parameter; ^{b)}threshold parameters (b_1, b_2, and b_3 represent the location or the value of theta (latent trait) that corresponds to a 0.5 probability of responding at or above that location on an item); ^{c)}Item fit was evaluated using the Orlando and Thissen S-χ^2 statistic; GRM = Graded Response Model; () = standard errors; DIF = differential item functioning by sex.									

Based on the scree plot (Supplementary Fig. 4) and fit statistics, a single-factor solution best represented the data, with the first factor accounting for 81.5% of the total variance. The EFA demonstrated excellent fit: RMSEA = 0.056 (90% CI: 0.033, 0.077), SRMR = 0.046, CFI = 0.996, TLI = 0.995, and χ^2 (35) = 63.08 ($p = 0.003$). All items loaded strongly on the single factor, from 0.68 (“participating in social activities”) to 0.97 (“standing up from an armless

chair"; Supplementary Table 7). In the independent Phase 3 sample ($n = 256$ after excluding seven outliers), CFA fit was acceptable on most indices: TLI = 0.993, CFI = 0.994, SRMR = 0.041, RMSEA = 0.083 (90% CI: 0.064, 0.103), and $\chi^2 (35) = 96.67, p < 0.001$. Standardized factor loadings ranged from 0.86 ("bathing or showering") to 0.97 ("using public transportation"), all significant at $p < 0.001$ (Fig. 1). A further CFA in the 2023 KFACS sample ($n = 879$; age range 75.5–90.6 years, mean = 83.1, SD = 3.5) replicated the structure in a substantially older population with acceptable model-data fit: $\chi^2 (35) = 292.94, p < 0.001$; TLI = 0.976; CFI = 0.982; SRMR = 0.048; RMSEA = 0.092 (90% CI: 0.082, 0.101) (Supplementary Fig. 5). In the Phase 3 sample, where participants completed both instruments separately, LF-10 scores correlated strongly with LF-25 scores ($r = 0.90, p < 0.001$). The LF-10 also correlated strongly with the PF scale in the Phase 2 validation sample ($r = 0.96, p < 0.001$). As expected, LF-10 scores were lower among participants who were older, female, less educated, widowed, or lower-income, and among those with poorer SRH, cognitive decline, depressive symptoms, or multimorbidity (Supplementary Table 8).

Reliability of the LF-10 Scale

The LF-10 showed strong internal consistency, with a McDonald's omega (ω) of 0.93 and a Cronbach's alpha (α) of 0.91. Of the 115 retest participants, 114 were analyzed after one outlier was excluded. Based on a two-way mixed-effects model with absolute agreement, the intraclass correlation coefficient (ICC) was 0.94 (95% CI: 0.92, 0.96), indicating excellent test–retest reliability. At the item level, weighted kappa values ranged from 0.42 to 0.81, reflecting moderate to substantial agreement [33]. Supplementary Table 9 presents the full reliability results.

Predictive Validity

Of the 519 baseline participants, 263 (50.7%) completed the follow-up assessment six months later. After excluding seven outliers, 256 were analyzed. Baseline LF-10 scores significantly predicted six-month follow-up PF scores ($\beta = 0.465, p < 0.001$) after adjustment for covariates and baseline PF (Table 3). In the adjusted ordinal model (Table 4), the assumption of proportional odds was met ($\chi^2 (8) = 9.03, p = 0.34$). Baseline LF-10 scores were prospectively associated with SRH at six months (Wald = 3.937, $p < 0.05$), although the association was modest (estimate = 0.054, 95% CI: 0.001, 0.107). Along with baseline LF-10 scores, better baseline SRH and male sex were also significant predictors of better SRH at follow-up.

Table 3

Results of regression of Physical Functioning (PF) scores at a six-month follow-up on baseline LF-10 scores ($n = 256$)

Variable	Model 1 (Unadjusted)			Model 2 (Adjusted)		
	<i>B</i>	<i>SE B</i>	<i>β</i>	<i>B</i>	<i>SE B</i>	<i>β</i>
Baseline LF-10 scores	0.627	0.042	0.693 ^{***}	0.420	0.130	0.465 ^{***}
Age				-0.061	0.088	-0.056
Sex				-1.702	1.076	-0.075
Educational attainment				1.326	0.967	0.081
Household income				0.533	0.404	0.082
PF scores at baseline				0.065	0.138	0.066
Multimorbidity				-3.189	1.421	– .120 [*]
R ²	0.481			0.535		
ΔR ²				0.054 ^{***}		
Note. B = unstandardized coefficients; β = standardized coefficients; SE = standard error; SRH = self-rated health status.						
*** <i>p</i> < 0.001, * <i>p</i> < 0.05						

Table 4
Results of ordinal regression of health conditions at a six-month follow-up on baseline LF-10 scores ($n = 256$)

Variable	Model 1 (Unadjusted)			Model 2 (Adjusted)		
	Estimate	SE	Wald	Estimate	SE	Wald
Baseline LF-10 scores	0.145	0.019	60.690***	0.054	0.027	3.937*
Age				-0.060	0.031	3.719
Sex						
Male				0.796	0.367	4.700*
Female				Ref	-	-
Household income				-0.127	0.138	0.841
Educational attainment						
≤ Middle school				-0.143	0.523	0.075
High school				0.129	0.374	0.119
≥ College				Ref	-	-
SRH at baseline				1.787	0.347	26.462***
Multimorbidity						
≤ 1				0.516	0.431	1.436
≥ 2				Ref	-	-
Note. SE = standard error; SRH = self-rated health status; Ref = reference category.						
*** $p < 0.001$, * $p < 0.05$						

Discussion

Accurate measurement of physical functioning in older adults is essential for assessing health status and planning long-term care. Many existing measures, however, focus on functional disability with particular emphasis on ADLs and IADLs, and their item content is largely confined to low-demand self-care and everyday activities. To address this gap, we previously developed the 25-item LF scale for KNHANES participants aged 50 years and older [6]. However, older respondents to large-scale national surveys frequently experience cognitive fatigue and response burden when completing lengthy questionnaires [34]. This study therefore developed and validated a short form of the LF-25 that maintains measurement quality while maximizing administrative feasibility for KNHANES.

By retaining 40% of the original items, the LF-10 approximately halved administration time. Despite this reduction, it retained the unidimensional structure of the original instrument, and its strong correlation with the LF-25 ($r = 0.90$) supports construct equivalence [19]. The LF-10 also correlated strongly with the PF scale ($r = 0.96$) and showed excellent internal consistency ($\omega = 0.93$) and test–retest reliability (ICC = 0.94). In the Phase 2 sample, 58.5% of participants scored at the ceiling and 81.2% scored 90 or above. This is unsurprising. Because the LF-10 was developed for adults aged 50 years and older, most of this sample (57.7%) were aged 50–64 years, younger than

the KHNHANES target age of 65 years and older, and difficulty with the assessed tasks is uncommon at these ages. However, the distribution differed substantially in the 2023 KFACS sample, whose participants were on average two decades older (mean age = 83.1 years). In this sample, the ceiling proportion fell to 25.1%, and scores were more widely spread, with 48.7% scoring below 90. These findings suggest that the observed ceiling effect reflects the functional status of the population sampled rather than a fixed property of the instrument. The LF-10 therefore appears able to distinguish meaningfully among older adults in the age range covered by KHNHANES. Nevertheless, a 25.1% ceiling among adults in their eighties shows that it has limited capacity to distinguish among the most highly functioning older adults.

Although CFI, TLI, and SRMR indicated good fit, the RMSEA values of both CFA models (0.083 and 0.092) exceeded the conventional threshold of 0.06. Uniformly high factor loadings are known to inflate RMSEA [35], and RMSEA can also be unreliable in models with relatively few degrees of freedom [36]. Given the very high loadings of the LF-10 items (0.86–0.97), the elevated RMSEA likely reflects these model characteristics rather than true misfit.

A key methodological challenge in this study was the handling of sparse response categories. In IRT, sparse categories can make threshold parameters unstable or impossible to estimate [26]. We therefore collapsed adjacent lower categories, following the common guideline of at least 2% of the total responses per category [25]. This sparseness is expected in community-dwelling older adults, the population for which the LF-10 is intended. However, collapsing may reduce measurement precision at the lower end of the functioning range, where distinguishing among frail individuals matters most. Because collapsing was applied only to GRM calibration, the LF-10 retains its original five-category format in KHNHANES. Future studies should test the full response scale in clinical or frail samples, where the lower categories are more fully used.

A key strength of this study is its integrated methodological approach. While several shortened measures of physical functioning exist, few were explicitly designed for community-dwelling older adults in national health surveys. For example, the widely used 12-item Short-Form Health Survey (SF-12) selected items using Classical Test Theory (CTT), which prioritizes the reproduction of summary scores over measurement precision across the full ability continuum [37]. CTT-based reduction methods often measure a narrower construct than originally intended because they preferentially select highly correlated items, inadvertently excluding valuable but statistically redundant content [38]. The LF-10 addressed these limitations by combining quantitative IRT-based item selection with qualitative content validation. Consistent with best-practice guidelines for patient-reported outcomes [39], qualitative procedures—including expert panel reviews (CVI/CVR) and cognitive interviewing—helped ensure that the retained items reflect the real-world functional experiences of older adults. This combination allowed the LF-10 to preserve clinically important content that statistical reduction alone might have eliminated. In addition, unlike the SF-12, which requires regression-based scoring algorithms, the LF-10 yields a simple 0–100 total score that policymakers and survey administrators can interpret directly.

Limitations and Future Directions

This study has several limitations. First, item thresholds were concentrated below the mean (-2.39 to -0.60), so the LF-10 measures less precisely at higher functioning levels ($\theta > 0$); future revisions should consider adding higher-demand physical tasks to extend the measurement range. Second, predictive validity was tested with only two outcomes over six months. While the LF-10 successfully predicted future self-rated health and PF scores, future studies should extend this timeline to examine whether the LF-10 predicts clinically meaningful outcomes such as disability onset, frailty, healthcare utilization, and mortality. Third, interpretation thresholds, including the

minimal clinically important difference (MCID), have not yet been established; anchor-based estimation will require large longitudinal samples and is a priority for future research. Finally, the LF-10 was developed and validated only in Korean samples; cross-cultural adaptation and validation are needed before use in other populations.

Conclusion

This study developed and validated the LF-10, an efficient short-form measure of physical functioning designed for KNHANES. By approximately halving administration time while maintaining psychometric quality comparable to the original 25-item scale, the LF-10 substantially reduces respondent burden for older adults. It has already been adopted by two major national studies: the longitudinal KFACS in 2023 and the nationally representative, cross-sectional KNHANES in 2024, where it is administered annually to community-dwelling adults aged 65 years and older [8]. Its adoption by studies with such different designs demonstrates its feasibility for large-scale health surveillance. As a reliable, easily administered population-level measure, the LF-10 equips public health practitioners and policymakers with the evidence needed to support targeted healthy aging initiatives in a rapidly aging society.

Declarations

Corresponding authors

Yunhwan Lee

Competing Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics approval

This study was performed in line with the principles of the Declaration of Helsinki. This study was approved by the Institutional Review Board of Ajou University Hospital (No. AJIRB-SBR-SUR-21-677).

Consent to participate

Written informed consent was obtained from all individual participants prior to data collection.

Consent to publish

Not Applicable

Funding

This research was supported by Korea Disease Control and Prevention Agency (KDCA) research project (Project No. 2022-11-008), and the Korea National Institute of Health (KNIH) research project (Grant No. 2021-ER0605-00,

Author Contribution

All authors contributed to the study's conception and design. Research design was led by ESK and YHL, and study goals were defined by YHL. Data curation was managed by JHK and MJK. Formal analysis and result validation were performed by ESK, JHK, and JWC. YHL, MJK, and JHY conducted the investigation, while YHL created the study model. Resources were provided by MJK, KWO, JHY, and YHK. Project administration was handled by JHY and YHK. MJK and KWO provided oversight, and KWO provided mentorship. Funding and financial support were secured by YHL and KWO. ESK and JHK prepared the original draft of the manuscript, and ESK created the data visualizations. YHL and KWO reviewed and edited the text. All authors read and approved the final manuscript.

Data availability

The datasets employed in this study are not yet publicly available.

References

1. Thiagarajan, A. J., Mikton, C., Harwood, R. H., Gichu, M., Gaigbe-Togbe, V., Jhamba, T., & Diaz, T. (2022). The UN Decade of healthy ageing: Strengthening measurement for monitoring health and wellbeing of older people. *Age and Ageing*, 51(7), afac147. <https://doi.org/10.1093/ageing/afac147>
2. World Health Organization (2021). Decade of healthy ageing: Baseline report. <https://www.who.int/publications/i/item/9789240017900>
3. Bierman, A. S. (2001). Functional status: The sixth vital sign. *Journal of General Internal Medicine*, 16(11), 785–786. <https://doi.org/10.1111/j.1525-1497.2001.10918.x>
4. Hopman-Rock, M., Van Hirtum, H., De Vreede, P., & Freiburger, E. (2019). Activities of daily living in older community-dwelling persons: A systematic review of psychometric properties of instruments. *Aging Clinical and Experimental Research*, 31(7), 917–925. <https://doi.org/10.1007/s40520-018-1034-6>
5. Weiss, C. O., Fried, L. P., & Bandeen-Roche, K. (2007). Exploring the hierarchy of mobility performance in high-functioning older women. *Journal of Gerontology: Series A Biomedical Sciences and Medical Sciences*, 62(2), 167–173. <https://doi.org/10.1093/gerona/62.2.167>
6. Lee, Y., Kim, E., Yun, J., Choi, J., Kim, J., Won, C. W., & Kim, J. (2024). The Life Functioning Scale: A measurement tool developed to assess the physical functioning abilities of community-dwelling adults aged 50 years or older. *Annals of Geriatric Medicine and Research*, 28(4), 410–418. <https://doi.org/10.4235/agmr.24.0087>
7. Nagi, S. Z. (1991). Disability concepts revisited: Implications for preventions. *Disability in America: Toward a national agenda for prevention*, 309–327.
8. Kim, S.-J., Park, S., Kim, S., Park, S., Kim, Y., Choi, Y., & Oh, K. (2026). Plan and operations of the 10th Korea National Health and Nutrition Examination Survey (2025–2027). *Epidemiology and Health*, 48, e2026001. <https://doi.org/10.4178/epih.e2026001>
9. Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349–360. <https://doi.org/10.1093/poq/nfp031>

10. Peytchev, A., & Peytcheva, E. (2017). Reduction of measurement error due to survey length: Evaluation of the split questionnaire design approach. *Survey Research Methods*, 11(4), 361–368.
<https://doi.org/10.18148/srm/2017.v11i4.7145>
11. Zablotsky, B., Lessem, S. E., Gindi, R. M., Maitland, A. K., Dahlhamer, J. M., & Blumberg, S. J. (2023). Overview of the 2019 national health interview survey questionnaire redesign. *American Journal of Public Health*, 113(4), 408–415. <https://doi.org/10.2105/AJPH.2022.307197>
12. Kočar, H. (2020). Development of a short form: Methods, examinations and recommendations. *Journal of Measurement and Evaluation in Education and Psychology*, 11(3), 301–310.
<http://doi.org/10.21031/epod.739548>
13. Stanton, J. M., Sinar, E. F., Balzer, W. K., & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology*, 55(1), 167–194. <https://doi.org/10.1111/j.1744-6570.2002.tb00108.x>
14. Jette, A. M., Haley, S. M., Ni, P., Olarsch, S., & Moed, R. (2008). Creating a computer adaptive test version of the late-life function and disability instrument. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 63(11), 1246–1256. <https://doi.org/10.1093/gerona/63.11.1246>
15. McDonough, C. M., Tian, F., Ni, P., Kopits, I. M., Moed, R., Pardasaney, P. K., & Jette, A. M. (2012). Development of the computer-adaptive version of the Late-Life Function and Disability Instrument. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 67(12), 1427–1438.
<https://doi.org/10.1093/gerona/gls108>
16. Rejeski, W. J., Marsh, A. P., Anton, S., Chen, S. H., Church, T., Gill, T. M., & LIFE Research Group. (2013). The MAT-sf: Clinical relevance and validity. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, 68(12), 1567–1574. <https://doi.org/10.1093/gerona/glt068>
17. Fox, G. C., Rodriguez, S., Rivera-Reyes, L., Loo, G., Hazan, A., & Hwang, U. (2020). PROMIS physical function 10-item short form for older adults in an emergency setting. *The Journals of Gerontology: Series A: Biological Sciences and Medical Sciences*, 75(7), 1418–1423. <https://doi.org/10.1093/gerona/glz163>
18. Smit, E. B., Bouwstra, H., Van Der Wouden, J. C., Hertogh, C. M. P. M., Wattel, E. M., Roorda, L. D., & Terwee, C. B. (2020). Development of a Patient-Reported Outcomes Measurement Information System (PROMIS®) short form for measuring physical function in geriatric rehabilitation patients. *Quality of Life Research*, 29(9), 2563–2572. <https://doi.org/10.1007/s11136-020-02506-5>
19. Goetz, C., Coste, J., Lemetayer, F., Rat, A. C., Montel, S., Recchia, S., & Guillemin, F. (2013). Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology*, 66(7), 710–718.
<https://doi.org/10.1016/j.jclinepi.2012.12.015>
20. Zamanzadeh, V., Rassouli, M., Abbaszadeh, A., Majd, H. A., Nikanfar, A., & Ghahramanian, A. (2014). Details of content validity and objectifying it in instrument development. *Nursing Practice Today*, 1(3), 163–171.
21. Park, S.-J., Choi, H.-R., Choi, J.-H., Kim, K.-W., & Hong, J.-P. (2010). Reliability and validity of the Korean version of the Patient Health Questionnaire-9 (PHQ-9). *Korean Academy of Anxiety Disorders*, 6(2), 119–124.
22. Lee, K. J., Han, G. S., Yoon, S. J., Lee, Y. K., Kim, C. H., Kim, J. L., & Lee, Y. H. (2002). The development of physical functioning scale for community-dwelling older persons. *Journal of Preventive Medicine and Public Health*, 35(4), 359–374.
23. Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, 49(2), 78–88. <https://doi.org/10.1037/cbs0000069>

24. Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). Routledge.
25. DiStefano, C., Shi, D., & Morgan, G. B. (2021). Collapsing categories is often more advantageous than modeling sparse data: Investigations in the CFA framework. *Structural Equation Modeling*, 28(2), 237–249. <https://doi.org/10.1080/10705511.2020.1803073>
26. Savalei, V. (2011). What to do about zero frequency cells when estimating polychoric correlations. *Structural Equation Modeling*, 18(1), 253–273.
27. Quan, Y., & Wang, C. (2025). Collapsing or not? A practical guide to handling sparse responses for polytomous items. *Methodology*, 21(1), e14303. <https://doi.org/10.5964/meth.14303>
28. Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research and Perspectives*, 11(3), 71–101. <https://doi.org/10.1080/15366367.2013.831680>
29. Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64. <https://doi.org/10.1177/01466216000241003>
30. Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265–289. <https://doi.org/10.3102/10769986022003265>
31. MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological methods*, 1(2), 130–149. <https://doi.org/10.1037/1082-989X.1.2.130>
32. Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer.
33. Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374. <https://doi.org/10.2307/2529786>
34. Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, 14(8), 1101–1108. <https://doi.org/10.1016/j.jval.2011.06.003>
35. Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, 71(2), 306–324. <https://doi.org/10.1177/0013164410384856>
36. Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507. <https://doi.org/10.1177/0049124114543236>
37. Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3), 220–233.
38. Kleka, P., & Soroko, E. (2018). How to abbreviate questionnaires and avoid the sins? *Survey Research Methods*, 12(2), 147–160. <https://doi.org/10.18148/srm/2018.v12i2.7224>
39. Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology*, 63(7), 737–745. <https://doi.org/10.1016/j.jclinepi.2010.02.006>

Figures

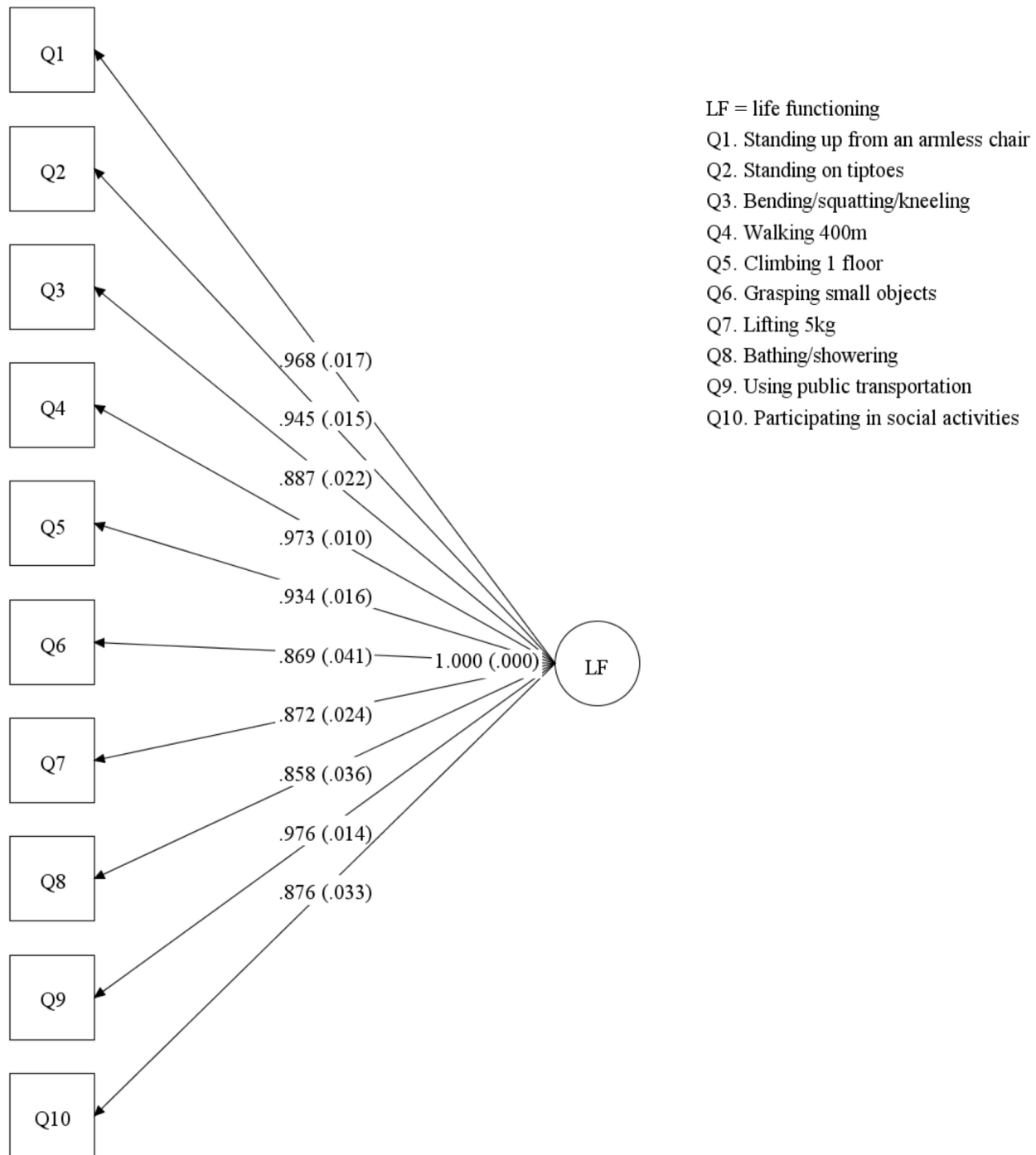


Figure 1

Results of Confirmatory Factor Analysis (CFA) ($n = 256$)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [LF10ManuscriptSupplementaryMaterialsQOLR.docx](#)