Review Article

# Psychometric Properties of Instruments Assessing Intrinsic Capacity: A Systematic Review




## Abstract

**Introduction:** Intrinsic capacity (IC) is a multidimensional indicator proposed by the World Health Organization that encompasses mental and physical capacities associated with functional ability. With the help of IC, different pathways of aging can be better understood, and heterogeneity can be captured more effectively. Before IC can be clinically incorporated, it requires valid and usable instruments alongside a comprehensive evaluation of psychometric evidence. Therefore, the present systematic review critically appraised, compared, and summarized the measurement properties of existing IC instruments used by older people. **Methods:** Published studies were searched in seven databases: EMBASE, MEDLINE, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science, until August 2022. The measurement properties of the IC measures were evaluated using the COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN). **Results:** Of the 582 papers initially identified, 10 studies were eligible for inclusion. Seven instruments were classified as five-domain measures, and three as more than five-domain measures. No instrument assessed all nine criteria in the psychometric properties evaluation outlined by COSMIN. The most reported psychometric properties were construct validity ($n$ = 8), measurement invariance ($n$ = 8), and structural validity ($n$ = 7). There was underreporting of content validity, reliability, and measurement error. **Conclusion:** The present review indicated a general lack of psychometric assessments of existing IC instruments with independent studies as their evidence base. There is a need to explore further the associations of IC and its five domains of interaction, which express the ability of individuals to interact with the environment and affect their functional ability.


**Keywords:** *Aging, COSMIN, intrinsic capacity, older people, psychometrics*


**Yi-Jung Chen[1], Shikha Kukreti[2], Hsin-Lun Yang[1], Chien-Chih Liu[3], Ya-Chin Yeh[4], Xavier C. C. Fung[5], Chieh-Hsiu Liu[6,7], Li-Fan Liu[8], Mark D. Griffiths[9], Yi-Ching Yang[10,11], Chung-Ying Lin[1,12,13]**

*[1]Institute of Allied Health Sciences, College of Medicine, National Cheng Kung University, [2]Department of Nursing, College of Medicine, National Cheng Kung University, [3]Center for General Education, National Tainan Junior College of Nursing, [8]Institute of Gerontology, College of Medicine, National Cheng Kung University, [10]Department of Family Medicine, College of Medicine, Cheng Kung University Hospital, National Cheng Kung University, [11]Department of Family Medicine, College of Medicine, National Cheng Kung University, [12]Department of Occupational Therapy, College of Medicine, National Cheng Kung University, [13]Biostatistics Consulting Center, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, [4]Department of Occupational Therapy, Shu-Zen Junior College of Medicine and Management, Kaohsiung, [6]Department of*


## Introduction

Global aging is expected to lead to a shift from disease-centered to function-centered approaches.[1] The World Health Organization (WHO) has proposed a new concept called intrinsic capacity (IC), which encompasses individual's mental and physical capabilities, and determines functional ability combined with environmental factors.[2] It is possible for older adults to improve their quality of life in their later years, provided that they live in a suitable environment and have reached the peak of each health phase, reducing the burden on society. The usual care model for older populations focuses on predicting and responding to diseases based on specific disease markers. Research suggests that shifts from disease-centered care to IC have significant implications for nursing practice among older hospitalized adults.[3]

IC can provide a more holistic understanding of different aging pathways, thereby capturing heterogeneity, a hallmark of older populations. Likewise, it might serve as a positive parameter for assessing health and providing guidance to health professionals on how to improve the well-being of older adults. However, it is necessary to conduct more comprehensive research before the IC construct can be globally incorporated into practice across different aging populations (e.g. young-old populations with different income and societal structures).

IC is a strong predictor for health outcomes from the perspective of function.[3,4] There are several complex IC indicators mentioned in various studies.[4,5] Nevertheless, there is no consensus regarding a standard manner to operationalize an IC in research studies or in clinical settings. There are already psychometric scales that are widely used and which separately assess IC domains,[6,7]



**How to cite this article:** Chen YJ, Kukreti S, Yang HL, Liu CC, Yeh YC, Fung XC, *et al.* Psychometric properties of instruments assessing intrinsic capacity: A systematic review. Asian J Soc Health Behav 2023;6:141-55.

**Access this article online**

**Website:** www.healthandbehavior.com

**DOI:** 10.4103/shb.shb_343_23

**Quick Response Code:**

*Family Medicine, Taoyuan General Hospital, Ministry of Health and Welfare, [7]School of Medicine, National Tsing Hua University, Taiwan, [5]Department of Rehabilitation Sciences, Faculty of Health and Social Sciences, The Hong Kong Polytechnic University, Hong Kong, China, [9]Department of Psychology, International Gaming Research Unit, Nottingham Trent University, Nottingham, UK*

**Received:** 07 August, 2023.
**Revised:** 22 October, 2023.
**Accepted:** 26 November, 2023.
**Published:** 31 January, 2024

**ORCID:**
Chung-Ying Lin:
https://orcid.org/0000-0002-2129-4242
Li-Fan Liu:
https://orcid.org/ 0000-0001-6610-5604

***Address for correspondence:***
*Dr. Chung-Ying Lin,*
*Institute of Allied Health Sciences, College of Medicine, National Cheng Kung University, No. 1, University Road, Tainan 701401, Taiwan.*
*E-mail: cylin36933@gmail.com*
*Prof. Li-Fan Liu,*
*Institute of Gerontology,College of Medicine, National Cheng Kung University, No.1, University Road, Tainan 701401, Taiwan.*
*E-mail: lilian@mail.ncku.edu.tw*



but the way to transform functionality scores into a standard index of intrinsic behavior deserves further examination. Consequently, there is a need for a comprehensive study to understand which of these can be used specifically to evaluate the overall physical and mental state of an individual.

To facilitate IC for an aging population, there is a need for valid and usable instruments that can evaluate the effectiveness of interventions to achieve healthy aging.[8] It is possible to use information from such instruments to support better decision-making in developing new projects and to enhance the quality of life among older people in later years. In addition to providing standardized information that allows for comparisons between different environments, appropriate instruments can also offer insight into how environments can be better adapted to older adults' needs. Before widespread deployment can be recommended, psychometric instruments need to meet established reliability and validity criteria, as well as be easy to administer by users.[9] Despite the utility of the WHO framework for healthy aging based on IC, measures across the different domains remain unstandardized.[10] The method for calculating a comprehensive IC index based on scores from different domains has not been unanimously agreed on. Consequently, the present review was conducted to systematically identify and critically appraise the psychometric properties of instruments used to assess IC.

## Methods

### Protocol and registration

A psychometric systematic review was performed in August 2022 based on the guidelines of both the (i) Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement for systematic reviews[11] and (ii) COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) checklist.[12] The psychometric systematic review protocol was registered on the International prospective register of systematic reviews (OSF Registries, registration Doi: 10.17605/OSF.IO/FU8GS).

### Information sources and search strategy

A literature search was carried out using EMBASE, MEDLINE, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science (these seven databases were used because they are among the most commonly used databases in this field of psychology) to retrieve relevant studies published up to August 31, 2022, with the following search strategy: IC (All fields) OR ICOPE (All fields). Slight modifications were made to the search strategy to optimize the search within each database. The reference lists of all included studies were screened manually to identify potential papers that might have been missed from the database search.

### Eligibility criteria

The present review included all study designs as long as the primary peer-reviewed paper reported at least one psychometric property of instruments assessing IC (i.e., reliability, internal consistency, measurement error, criterion validity, hypotheses testing for construct validity, structural validity, content validity, cross-cultural validity, responsiveness) as defined by the COSMIN. The review only included English language studies of all types of research design. Gray literature such as conference proceedings, dissertations, or unpublished literature were excluded because these publications may not have been peer-reviewed and may have had insufficient information to assess methodological quality. Their existence may also be temporary.

### Selection of sources of evidence

The results of all searches were entered into the EndNote Team, 2013, Philadelphia, U.S., Clarivate, Endnote X20.0.1

software program for systematic reviews. This systematic review used the PRISMA 2020 flow chart assessment for the study selection process. The process followed to identify the studies to include in this systematic review is described in Figure 1.

## Data charting process

Two independent reviewers (Chien-Chih Liu, Yi-Jung Chen), who are experienced researchers and actively involved in psychometric research activities, screened the titles and abstracts to identify potential studies for full-text screening. Any disagreements were discussed between the two reviewers. A third reviewer (Ya-Ching Yeh) was consulted if necessary. Full texts of all potential papers were then retrieved and screened using the same procedure. The findings were then verified by another independent investigator (Chung-Ying Lin) for the final review.

## Critical appraisal of individual sources of evidence

The methodological quality of the included studies was assessed using the COSMIN Risk of Bias checklist (http://www.cosmin.nl). The COSMIN checklist was selected due to its comprehensive assessment of all domains of psychometric properties compared to other risk-of-bias assessment tools that focus on only a few aspects. Moreover, it has been utilized in systematic reviews evaluating the methodological quality of studies involving performance-based outcomes with generic items designed for multiple applications. More specifically, the COSMIN checklist contains ten boxes (patient-reported outcome measures development, reliability, internal consistency, measurement error, criterion validity, hypotheses testing [construct validity], structural validity, content validity, cross-cultural validity, responsiveness) evaluating the methodological standards of a study in

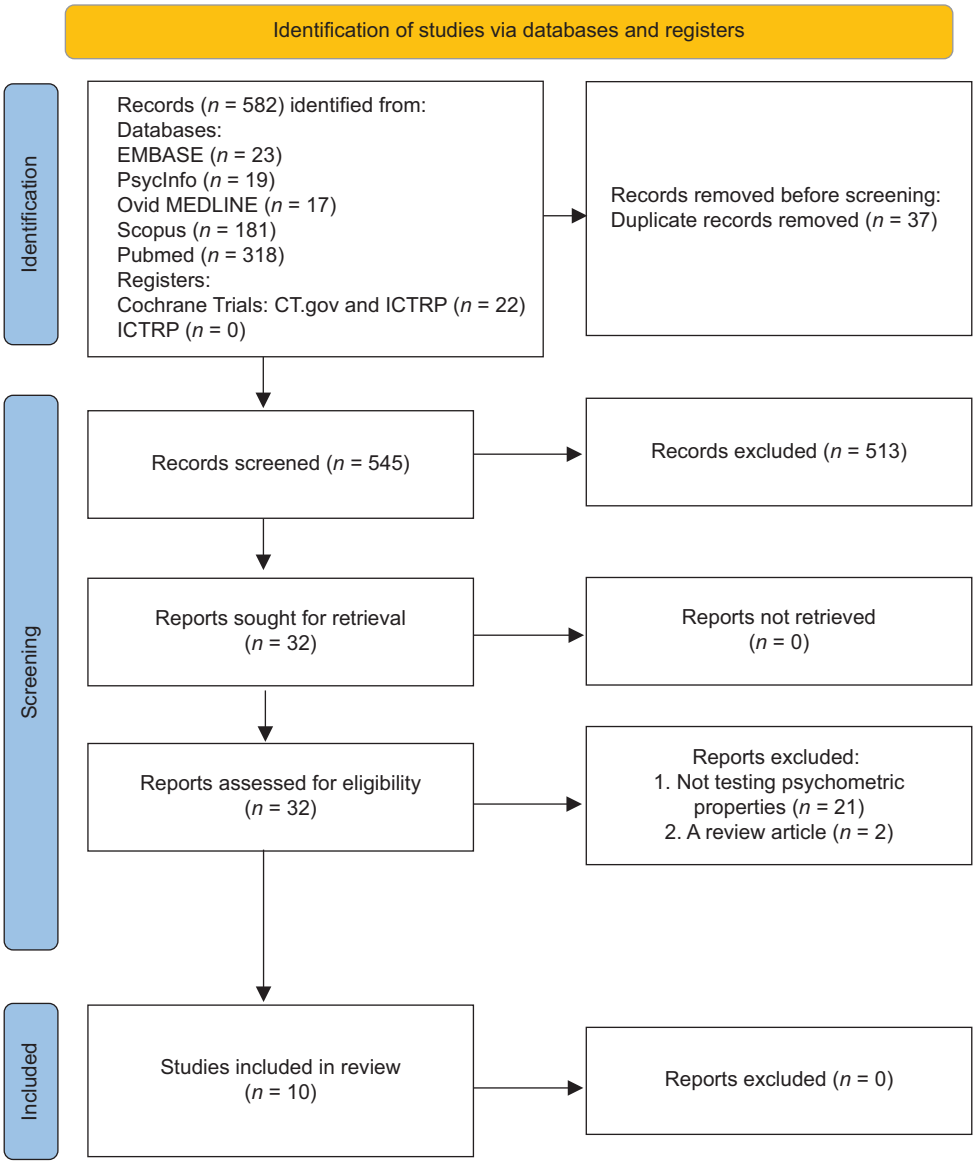


**Figure 1: Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 flow diagram of the study selection process**

terms of its design and statistical approach. Each item is scored on a four-point rating scale: very good, adequate, doubtful, and inadequate. The overall quality score of a measurement property was graded based on the lowest rating of any item within that measurement property box (i.e. the "worst score counts" method). The checklist includes statistical quality evaluation criteria for various psychometric properties [Appendix S]. Each criterion was rated as positive (+), negative (−), or indeterminate (?) depending on the study results. For example, a positive rating for reliability is given if the intraclass correlation coefficient (ICC) ≥0.70, whereas a negative rating is given if ICC ≤0.70. Indeterminate (?) rating is given if no ICC is reported.

### Data items

To evaluate the quality of psychometric properties of instruments involves examining various properties, including content validity, structural validity, cross-cultural validity, internal consistency, reliability, measurement error, and construct validity.

### Synthesis of results

An assessment was conducted of the statistical outcomes of each measurement property from the instruments on IC in comparison with the updated criteria for adequate measurement properties specified by Terwee *et al*.[13] and Prinsen *et al*.[14] The results were classified as satisfactory (+), inadequate (−), or inconclusive (?), and all evaluations are presented in the results table along with their corresponding criterion.

## Results

### Study selection

Of 582 initially identified outputs, 37 were duplicates [Figure 1]. Ten papers from eight projects were included. Those populations were recruited from the general community ($n$ = 8), superior quality senior community ($n$ = 1), and hospital ($n$ = 1). The characteristics and sociodemographics of the population included in each study are summarized in Table 1. The COSMIN grading, statistical findings [Tables 2 and 3], and levels of evidence [Table 4] for various measurement properties of the instruments on IC in different projects are also reported.

### Structural validity

Of the 10 studies, seven[5,15-20] demonstrated very good methodological quality and had a positive rating for the quality of statistical findings based on a reflective model [Table 2]. Six studies refer to the model fit to confirm a bi-factor model for an instrument with five subscales.[5,15,17-20] Another study displayed a two-parameter logistic item response theory model on a scale with 41 items.[16] The best-evidence synthesis showed strong positive evidence for the structural validity of the instruments on IC with cognition, sensory, locomotor, vitality, and psychological [Table 4].

### Internal consistency

Four studies evaluated the internal consistency of the instruments on IC and had very good methodological quality.[5,15,18,20] Two studies showed a positive rating for the quality of statistical findings, indicating that the total score predominantly reflected a single factor.[13,14] Two studies demonstrated negative quality criteria of the hierarchical (ωH) index attributed to the multidimensionality caused by the subdomain factors.[18,20] The best-evidence synthesis showed limited positive evidence for the internal consistency of the instruments on IC, using the same items.

### Measurement invariance/Cross-cultural validity

Eight studies examined the measurement invariance of the instruments on IC across different factors [Table 2]. The measurement invariance across age, gender, education, marital status, and multimorbidity was fully supported in these studies.[5,15,17-22] Furthermore, one study examined the differential item functioning contrast across multiple countries and IC using Rasch analysis; t-scores were lower in each older group, and males and/or individuals with higher educational level, greater wealth, and never smoking had higher t-scores.[16]

### Validity

#### *Criterion validity (predictive and concurrent validity)*

According to the COSMIN checklist, two studies had very good methodological quality for predictive validity[22,23] and two studies had inadequate methodological quality for predictive validity.[16,17] Two studies[22,23] demonstrated positive quality criteria and reported strong correlations between the instruments on IC and three functional assessment tools: the Katz activities of daily living index,[22] the risk of incident fall,[22] and the five-item Cardiovascular Health Study frailty phenotype.[23] In addition, one study[21] demonstrated very good methodological quality in evaluating the concurrent validity of the instruments on IC. Similarly, the instruments on IC significantly correlated with the Barthel index, the instrumental activities of daily living (IADL), the Fried phenotype, and the Strength, Assistance with Walking, Rising from a Chair, Climbing Stairs, and Falls. The best-evidence synthesis yielded an unknown level of evidence for predictive validity and concurrent validity [Table 4].

#### *Construct validity (convergent validity)*

Eight studies[5,15-21] had very good methodological quality. All the cross-sectional and longitudinal cohort studies displayed weak to moderate correlation coefficients (<0.5) between the instruments on IC and the activities of daily living (β = 0.23–0.52)[12,14] (odds ratio [OR] =1.72)[19] or the IADL (β = 0.32–0.48)[12,14] ($r$ = 0.45)[20] (OR = 1.95).[19] One

**Table 1: The characteristics and sociodemographic of the populations in the included studies**

| Author/region | Study design (project name) | Sample size (percentage of females) | Age (years) | Number of items | The COSMIN measurements addressed | Other notes |
|---|---|---|---|---|---|---|
| Beard *et al*. (2019)/UK[5] | Longitudinal (English Longitudinal Study on Ageing) | 2352 (55.3) (nationally representative sample) | ≥50 | 10 | Structural validity<br>Internal consistency<br>Measurement invariance<br>Convergent validity<br>Discriminative validity | Biomarkers and self-reported measures<br>Home visiting<br>12 year follow-up |
| Daskalopoulou *et al*. (2019)/ Latin American [15] | Cross-sectional (10/66 Dementia research) | 12,865 (64) | ≥65 | 26 | Structural validity<br>Internal consistency<br>Measurement invariance<br>Convergent validity | Low-and middle income country<br>Items were collected between 2003 and 2010<br>Community screening |
| Ma *et al*. (2020)/ China[21] | Cross-sectional (NA) | 376 (40.43) | 68.65±11.41 | 7 | Cross-cultural validity<br>Criterion validity<br>Convergent validity | Hospital-based screening |
| Sanchez-Niubo *et al*. (2020)/ Australia, China, Europe, UK, Spain, Japan, Korea, India, Mexican, Irish[16] | Cross-sectional (NA) | 343,915 (55) | 60 (18–114) | 41 | Structural validity<br>Cross-cultural validity<br>Reliability<br>Convergent validity | ATHLOS project, data from 16 international cohorts were harmonized<br>Using IRT models |
| Liu *et al*. (2021)/ China[22] | Longitudinal (NA) | 212 (59.4) | 83.8±4.4 | 12 | Cross-cultural validity<br>Criterion validity | Superior quality senior community<br>2 years follow-up<br>Hospital-based screening |
| Yu *et al*. (2021)/ Hong Kong[17] | Longitudinal (Mr. OS and Ms. OS study†) | 3736 (49.7) | 72.22 | 38 | Structural validity<br>Measurement invariance<br>Discriminative validity<br>Predicative validity | Cohort study on osteoporosis and general health<br>7 years follow-up with complete data on IADL<br>Biomarkers and self-reported measures |
| Beard *et al*. (2021)/China[18] | Longitudinal (China Health and Retirement Longitudinal Study) | 7643 (Nationally representative sample) | ≥60 | 37 | Structural validity<br>Internal consistency<br>Measurement invariance<br>Convergent validity<br>Discriminative validity | 2 years follow-up |
| Yu *et al*. (2022)/ Hong Kong[23] | Longitudinal (Mr. OS and Ms. OS study†) | 3018 (50 baseline) | 72.5±5.2 | 38 | Measurement invariance<br>Criterion validity<br>Convergent validity<br>Discriminative validity | Cohort study on osteoporosis and general health<br>4 years follow-up with complete data on frailty<br>Biomarkers and self-reported measures |
| Aliberti *et al*. (2022)/Brazil[19] | Cross-sectional (Brazilian Longitudinal study of Aging) | 7175 (53.1) (Nationally representative sample) | 62.4±9.3 | 15 | Structural validity<br>Measurement invariance<br>Convergent validity<br>Discriminative validity | Self-reported and physical performance measures |

*Contd...*

**Table 1: Contd...**

| Author/region | Study design (project name) | Sample size (percentage of females) | Age (years) | Number of items | The COSMIN measurements addressed | Other notes |
|---|---|---|---|---|---|---|
| Gao *et al*. (2022)/China[20] | Longitudinal (China Health and Retirement Longitudinal Study) | 13,233 (52.48) | 50–60 (34.55) | 37 | Structural validity<br>Internal consistency<br>Measurement invariance<br>Convergent validity<br>Discriminative validity | 2 years follow-up |

†Mr. OS and Ms. OS (Hong Kong) is a cohort study to examine the determinants of osteoporotic fractures in older Chinese men and women. NA: Not available, IRT: Item response theory, ATHLOS: Ageing Trajectories of Health-Longitudinal Opportunities and Synergies, IADL: Instrumental activities of daily living, COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments



study showed strong correlation coefficients (>0.5) between the instruments on IC and the healthy life expectancy at birth and the gross domestic product per capita.[16]

## Measurement of intrinsic capacity by domains

### *Locomotion*

Eight studies included the Short Physical Performance Battery (SPPB) in their assessments, either in its entirety[5,18] [Table 5] or in part.[15,17,19,21-23] The SPPB comprises three components: chair-stand, gait speed, and standing balance. However, there were variations in the methods used to assess chair-stand and gait speed across the studies. For instance, while the duration required to walk 2.44 m was recorded in the gait speed test, Daskalopoulou *et al*.[15] calculated gait speed test by the time taken to walk 10 m. One study employed self-reported measures for locomotion but also included other measures such as running, jogging walking, getting up, climbing, stooping, kneeling, and crouching [Table 5].[20]

### *Vitality*

Handgrip strength was assessed as an indicator of vitality in five studies,[5,17-19,23] and two studies examined lung function through the forced expiratory volume and peak flow test.[5,18] In addition to these, vitality was assessed in three studies by inquiring about unintended weight loss and appetite.[19,21,22] Other self-reported assessments included impairment in activities of daily living,[15,20] experiences of pain,[16] energy levels,[16] urine incontinence,[16,20] exhaustion,[19] and endurance.[19] Finally, two of the studies included the measurement of biomarkers as an assessment of vitality, specifically dehydroepiandrosterone, hemoglobin level, and insulin-like growth factor [Table 5].[5,18]

### *Cognition*

Verbal fluency,[5,15,16,19] time orientation,[15,16,19-22] and delayed recall[15,16,20] were included in the majority of studies' assessments of cognition. Other methods were also used including attention,[5] long memory test,[15] the praxis-fold a piece of paper,[15] story recall difficulty,[15] processing speed,[16] episodic memory,[18] semantic memory,[19] and numeracy.[16,20] Two studies used measures such as the 30-item Mini-Mental State Examination (MMSE),[17,23] and one study used some components from the Telephone Interview of the Cognitive Status battery [Table 5].[18]

### *Psychological*

The majority of studies used self-reported measures and tools tailored specifically for older adults to assess depressive symptoms, including the Geriatric Depression Scale (GDS),[17,23] the Center for Epidemiological Studies Depression (CES-D) scale,[18] five-item CES-D,[5] and eight-item CES-D.[19] Other studies assessed psychological-related variables such as sleep disturbance,[5,15,16,18,19] the presence of depressed or hopeless feelings,[20-22] and the experience of no interest or pleasure [Table 5].[15,20-22]

### *Sensory*

Two studies utilized performance-based measures such as the whisper test and audiometry test.[16,21] The Snellen Eye Test and the Frisby Stereo Test were used in two studies as performance tests to evaluate vision.[17,23] Eight studies assessed vision and hearing for the sensory domain using self-report questionnaires to evaluate either vision or hearing.[5,15,16,18-22] Some of the questions related to vision concerned the participants' capability to see distant objects, read, and the interference of poor eyesight in daily activities.[16,19,20] Some assessments included questions concerning their general hearing [Table 5].[19,20]

## Discussion

The present review was conducted to systematically examine studies evaluating the psychometric properties of instruments with regards to IC to investigate the methodological quality of these studies, to evaluate the quality of psychometric properties, and to grade the existing evidence. Ten studies from eight projects were included in the evaluation. The review found that most instruments had been tested for validity but seldom for reliability. The quality of the psychometric properties evaluated using the criteria for good measurement properties included the following: content validity, structural validity, cross-cultural validity, internal consistency, reliability, measurement error, and construct validity. A total of eight

**Table 2: Summary of structural validity, internal consistency, and measurement invariance of measures on intrinsic capacity**

| Author/region | Structural validity (Box 3) | COSMIN score/quality score[#] | Internal consistency (Box 4) | COSMIN score/quality score[#] | Measurement invariance (Box 5) | COSMIN score/quality score[#] |
|---|---|---|---|---|---|---|
| Beard *et al*., (2019)/UK[7] | Bi-factor EFA: $\chi^2$=71.2 (df=39), RMSEA=0.012 (90% CI 0.011–0.024)<br>CFI=0.99 and TLI=0.99<br>Bi-factor CFA: $\chi^2$=1180.6 (df=89), RMSEA=0.035 (90% CI 0.033–0.037), CFI=0.98 and TLI=0.97 | Very good/+ | The ωH value for the general factor was 0.78, and the subscore values for specific factors were 0.79, 0.80, 0.81, 0.82 and 0.83 for cognition, sensory, locomotor, vitality, and psychological | Very good/+ | Regression coefficient (95% CI) for age−0.052 (−0.054–−0.046); female−0.322 (−0.358–−0.286); higher education 0.779 (0.735–0.823); highest wealth 0.616 (0.553–0.678); 3 or more multimorbidity −0.764 (−0.816–−0.712) | Very good/+ |
| Daskalopoulou *et al*., (2019)/ Latin American[15] | EFA: $\chi^2$=786.05, df=227, RMSEA=0.025; 90% CI=0.023–0.027, CFI=0.991<br>CFA: Bi-factor: CFI=0.972, RMSEA=0.041<br>Second-order: CFI=0.962, RMSEA=0.045 | Very good/+ | A comparison of ωH with ω (0.84/0.96=0.88), and the subscore values for specific factors were 0.06, 0.02, 0.03 and 0.02, respectively | Very good/+ | Bifactor model had acceptable fit across countries and gender (RMSEA values range from 0.030 to 0.052; CFI values range from 0.923 to 0.976) | Very good/+ |
| Ma *et al*. (2020)/ China[21] | NA | NA | NA | NA | IC score decreased with increasing age, from 5.32±0.79 at age 50–59 years to 4.01±1.56 at age 80 and older. There was no difference observed in IC between men and women | Very good/+ |
| Sanchez-Niubo *et al*. (2020)/ Australia, China, Europe, UK, Spain, Japan, Korea, India, Mexican, Irish (Global)[16] | IRT model converged successfully with an excellent fit (RMSEA=0.03, TLI=0.99 and CFI=0.99) | Very good/+ | NA | NA | NA | NA |
| Liu *et al*. (2021)/ China[22] | NA | NA | NA | NA | There were no differences in gender, marital status, and educational level between functional decline and nonfunctional decline group. And no differences were observed regarding gender, marital status, educational level and CCI between fall and nonfall group | Very good/+ |
| Yu *et al*. (2021)/ Hong Kong[17] | Bi-factor CFA: RMSEA=0.031 (90% CI=0.028–0.035)<br>5-factor CFA: RMSEA=0.055 (90% CI=0.053–0.058) | Very good/+ | NA | NA | Women had a lower IC score compared to man ($P$<0.0001)<br>Lower IC scores were also found in participants who had lower levels of education ($P$<0.0001), lower subjective social status ($P$<0.001), reported more chronic diseases ($P$<0.0001), or had a higher number of IADL limitations ($P$<0.0001) | Very good/+ |

*Contd...*

**Table 2: Contd...**

| Author/region | Structural validity (Box 3) | COSMIN score/quality score[#] | Internal consistency (Box 4) | COSMIN score/quality score[#] | Measurement invariance (Box 5) | COSMIN score/quality score[#] |
|---|---|---|---|---|---|---|
| Beard *et al*. (2021)/China[18] | Bi-factor EFA: $\chi^2$=52.4 (df=39), RMSEA=0.007 (90% CI 0.000–0.011), CFI=0.999, and TLI=0.998<br>Bi-factor CFA: $\chi^2$=625.9 (df=88), RMSEA=0.028 (90% CI 0.026–0.030), CFI=0.97, and TLI=0.95 | Very good/+ | A comparison of ωH with ω (0.67/0.85=0.79)<br>The ωh value for the general factor was 0.67, and the subscore values for specific factors were 0.33, 0.53, 0.33, 0.13, and 0.56 for cognition, sensory, locomotor, vitality, and psychological | Very good/− | Regression coefficient (95% CI) for age−0.022 (−0.024–−0.02); female−0.349 (−0.372–−0.325); higher education 0.664 (0.614–0.715); lowest wealth−0.191 (−0.231–−0.152); 3 or more multimorbidity−0.157 (−0.193–−emale | Very good/+ |
| Yu *et al*. (2022)/Hong Kong[23] | NA | NA | NA | NA | NA | NA |
| Aliberti *et al*. (2022)/Brazil[19] | Bi-factor CFA: $\chi^2$=239.9, $P$<0.001, CFI=0.984, RMSEA=0.020, SRMR=0.015 | Adequate/+ | NA | NA | Higher levels of IC were associated with preserved ADL and IADL, younger age, male sex, white race, having a partner, living in urban areas, higher education, fewer chronic diseases, and reporting smoking and alcohol consumption | Very good/+ |
| Gao *et al*. (2022)/China[20] | 5 factor EFA: CFI=0.948; RMSR=0.03; RMSEA=0.049; 95% CI=0.049–0.050<br>Second-order CFA: $\chi^2$=1007.8; df=30; $P$<0.001) | Very good/+ | A comparison of ωH with ω (0.69/0.90=0.77)<br>The ωh value for the general factor was 0.69, and the subscore values for specific factors were 0.42, 0.34, 0.26, 0.61, and 0.45 for sensory functions, cognition, mobility, activities of daily living, and psychology symptoms | Very good/− | Higher levels of healthy aging scale were associated with female (ARC=−2.75; 95% CI=−3.17–−2.32), younger age (50–60, ARC=−1.47, 95% CI=−2.03–−0.92; 60–70, ARC=−3.16; 95% CI=−3.75–−2.56; 70-80, ARC=−6.44; 95% CI=−7.19–−5.69; >80, ARC=−12.12; 95% CI=−13.4–−10.85), divorced/separated (ARC=−0.95, 95% CI=−1.83–−0.06), widowed/never married (ARC=−2.29, 95% CI=−2.97–−1.61), higher education (high school, ARC=9.58, 95% CI=8.73–10.44; vocational school, ARC=12.37, 95% CI=11.05–13.68; college and above, ARC=12.39, 95% CI=10.99–13.79), better self-rated health (very good, ARC=−2.39, 95% CI=−3.31–−1.47; good, ARC=−7.33, 95% CI=−8.18–−6.48; fair, ARC=−17.29, 95% CI=−18.22–−16.36; poor, ARC=−27.67, 95% CI=−28.94–−26.41), and fewer chronic diseases | Very good/+ |

*Contd...*

**Table 2: Contd...**

| Author/region | Structural validity (Box 3) | COSMIN score/quality score[#] | Internal consistency (Box 4) | COSMIN score/quality score[#] | Measurement invariance (Box 5) | COSMIN score/quality score[#] |
|---|---|---|---|---|---|---|
| | | | | | (1, ARC=−1.79, 95% CI=−2.3–−1.29; ≥2, ARC=−4.66, 95% CI=−5.17–−4.15) | |

[#]Quality score of the measurement property: (+) positive measurement property, (−) negative measurement property, (?) indeterminate. CFI: Comparative fit index, RMSEA: Root-mean-square error of approximation, SRMR: Standardized root mean square residual, IRT: Item response theory, CCI: Charlson comorbidity index, CI: Confidence interval, ARC: Adjusted regression coefficients, ADL: Activities of daily living, IC: Intrinsic capacity, IADL: Instrumental ADL, NA: Not available, *r*: Pearson's correlations coefficients, COSMIN: COnsensus-based Standards for the selection of health Measurement Instruments, CFA: Confirmatory factor analysis, EFA: Exploratory factor analysis, TLI: Tucker–Lewis index

**Table 3: Summary of criterion validity and construct validity of measures on intrinsic capacity**

| Author/region | Criterion validity (Box 8) | COSMIN score/quality score[#] | Construct validity (Box 9) | COSMIN score/quality score[#] |
|---|---|---|---|---|
| Beard *et al*. (2019)/UK[5] | NA | NA | IC/ADL: β=−0.52, $R^2$=0.20<br>IC/IADL: β=−0.48, $R^2$=0.21 | Very good/− |
| Daskalopoulou *et al*. (2019)/ Latin American[15] | NA | NA | IC/self-rated health: Standardized estimate−0.373; bootstrap 95% CI=0.352–0.394, *P*<0.001, $\chi^2$=8238.22, df=348, RMSEA=0.050; 90% CI=0.049–0.051, CFI=0.922 | Very good/− |
| Ma *et al*. (2020)/China[21] | AUC-ROC for the IC versus fried phenotype, FRAIL, ADL disability, IADL disability, and SARC-F were 0.817, 0.843, 0.954, 0.912, and 0.909, respectively | Very good/+ | IC was significantly positively correlated with the resilience score (*r*=0.316, *P*<0.001), and MMSE score (*r*=0.358, *P*<0.001), while it was negatively correlated with IADL score (*r*=−0.446, *P*<0.001), Fried frailty score (*r*=−0.398, *P*<0.001), FRAIL score (*r*=−0.365, *P*<0.001), SARC-F score (*r*=−0.347, *P*<0.001), GDS score (*r*=−0.552, *P*<0.001), physical fatigue (*r*=−0.278, *P*<0.001), and mental fatigue (*r*=−0.195, *P*=0.001) | Very good/− |
| Sanchez-Niubo *et al*. (2020)/ Australia, China, Europe, UK, Spain, Japan, Korea, India, Mexican, Irish (Global)[16] | The group with obesity, arterial hypertension, depression, physical diseases and loneliness were associated with lower IC<br>The group with the lowest IC had a 50% survival probability in 10 years and for the other groups it was in at least 20 years | Inadequate/− | Correlations between IC by country and ecological country indicators were 0.81 with HALE and 0.58 with GDP | Very good/+ |
| Liu *et al*. (2021)/China[22] | The AUC for IC for predicting functional decline was 0.814 (95% CI: 0.756–0.871)<br>The AUC for IC for predicting falls was 0.806 (95% CI: 0.744–0.868) | Very good/+ | Concerning IC, the univariable logistic regression analysis illustrated that the impaired chair rise test, weight loss, appetite loss, vision impairment, orientation and memory impairment, feeling hopeless, and interest loss increased the risk of functional decline (*P*<0.05) | NA |
| Yu *et al*. (2021)/Hong Kong[17] | IC had a direct effect in predicting incident IADL limitations at the 7 years follow-up (β=−0.21, *P*<0.001) | Inadequate/− | Female sex also had a direct effect on IC (β=−0.58, *P*<0.001), although its effect on the number of chronic diseases was not significant<br>Higher subjective social status had a direct effect on the number of chronic diseases (β=−0.05, *P*<0.05) and IC (β=0.05, *P*<0.05). The model explained 7.4% of the variance in incident IADL limitations | Very good/− |

*Contd...*

**Table 3: Contd...**

| Author/region | Criterion validity (Box 8) | COSMIN score/ quality score[#] | Construct validity (Box 9) | COSMIN score/ quality score[#] |
|---|---|---|---|---|
| Beard *et al*. (2021)/China[18] | NA | NA | IC predicted declining performance in both IADLs (β=−0.324, *P*<0.001) and ADLs (β=−0.227, *P*<0.001) | Very good/− |
| Yu *et al*. (2022)/Hong Kong[23] | Combination of vitality and sensory for men (year 4, OR=0.03, 95% CI=0.004–0.22, AUC=0.798) and with the combination of vitality and locomotor for women (year 2, OR=0.16, 95% CI=0.07–0.34, AUC=0.754) with incident frailty | Very good/+ | Vitality was the domain most strongly associated with incident frailty at each follow-up (year 2, OR=0.33, 95% CI=0.24–0.45; year 4, OR=0.33, 95% CI=0.23–0.46) | NA |
| Aliberti *et al*. (2022)/Brazil[19] | NA | NA | IC/older age *r*=−0.29, 95% CI=−0.32–−0.27 IC composite score was associated with almost twice the odds of preserved ADL (OR=1.72; 95% CI=1.54–1.93), preserved IADL (OR=1.95; 95% CI=1.77–2.16), and high performance in AADL (OR=1.79; 95% CI=1.59–2.00) | Very good/− |
| Gao *et al*. (2022)/China[20] | NA | NA | Lowest score quartile: The second (AOR, 0.78; 95% CI=0.67–0.92), third (AOR, 0.70; 95% CI=0.58–0.85), and fourth (AOR, 0.52; 95% CI=0.41–0.66) score quartiles had lower adjusted OR of times of inpatient service | NA |

[#]Quality score of the measurement property: (+) positive measurement property; (−) negative measurement property; (?) indeterminate. *r*: Pearson's correlations coefficients, β: Standardized coefficient, OR: Odds ratio, CI: Confidence interval, CFI: Comparative fit index, RMSEA: Root-mean-square error of approximation, AUC: Area under the curve, ROC: Receiver operating characteristic curve, AOR: Adjusted OR, ADL: Activities of daily living, IC: Intrinsic capacity, IADLs: Instrumental ADLs, AADL: Advanced ADL, SARC-F: The Strength, Assistance with walking, Rising from chair, Climbing stairs, and Falls questionnaire was used to assess sarcopenia, with higher scores indicating more severe, MMSE: Mini-mental state examination, GDS: Geriatric Depression Scale, HALE: Healthy life expectancy, GDP: Gross domestic product, NA: Not available, COSMIN: COnsensus-based Standards for the selection of health Measurement INstruments



studies were identified in which cross-cultural validity was tested,[5,15,17-22] and three studies[21,23] tested criterion validity. While a couple of rapid reviews have documented IC's association with health outcomes,[10,24] a collective body of evidence is lacking. A recent critical literature review by Gonzalez-Bautista *et al*.[10] also examined the assessment of IC, and the tools used to assess the domains. However, the present review went much further than this paper by additionally examining the measurement invariance of the instruments on IC across age, gender, education, marital status, and multimorbidity.

A measurement instrument's content validity is arguably the most important psychometric property to consider.[25] However, none of the included studies was rated very good or adequate for the methodological quality of their content validity. Among the included studies, the majority were longitudinal studies using data collected in community settings. The most commonly used tools for assessing IC were the (i) Gait speed test, walking speed, and chair stand test (locomotion), (ii) Grip-strength and weight loss (vitality), (iii) MMSE (cognition), (iv) GDS or CES-D scale (psychological), and (v) Self-reported vision and health questionnaires (sensory). However, the analysis found heterogeneity and low concordance in the operationalization of some of the domain measurements, particularly the vitality and psychological domains, which make cross-study comparisons difficult. A similar observation was reported in a systematic review by George *et al*.[24] Therefore, it is critical to clarify concepts regarding psychological domains and vitality to reach a consensus on their appropriate measurement and weight. Moreover, only one study[16] showed a strong correlation coefficient between the instruments on IC and the healthy life expectancy at birth and the gross domestic product per capita.

It is important to consider both validity and reliability when selecting instruments for assessing health outcomes.[9] The present review found that most of the instruments used in the included studies were psychometrically validated. The locomotion domain was assessed using performance-based tests, and the vitality domain was assessed using blood biomarkers. Compared to other domains, psychological and sensory measures were mostly self-reported, which may lead to social desirability bias and recall bias.[26] By using assessment tools that are less susceptible to bias and using appropriate weightings for the different IC domains, IC

**Table 4: Levels of evidence of various measurement properties of intrinsic capacity in different projects**

| Project | Number of studies (total participants) | Content validity | Structural validity | Internal consistency | Reliability | Measurement error | Criterion validity | Construct validity | | | Responsiveness |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Known-group | Convergent | Discriminant | |
| ELSA | 1 (2352) | Limited (−) | Strong (+) | Strong (+) | NA | NA | NA | Limited (−) | Strong (+) | NA | NA |
| 10/66 Dementia research | 1 (12,865) | Limited (−) | Strong (+) | Strong (+) | NA | NA | NA | Limited (−) | Strong (+) | NA | NA |
| Department of Geriatrics in Xuanwu Hospital Capital Medical University | 1 (376) | Limited (−) | NA | NA | NA | NA | Strong (+) | Strong (+) | Strong (+) | Strong (+) | NA |
| ATHLOS project | 1 (343,915) | Limited (−) | Strong (+) | NA | NA | NA | Limited (−) | Strong (+) | Strong (+) | Limited (−) | NA |
| CCRC | 1 (212) | Limited (−) | NA | NA | NA | NA | Strong (+) | NA | NA | Strong (+) | NA |
| Mr. OS and Ms. OS study | 2 (4000) | Limited (−) | Strong (+) | NA | NA | NA | Strong (+) | Strong (+) | Strong (+) | Strong (+) | NA |
| CHARLS | 2 (13,233) | Limited (−) | Strong (+) | Strong (+) | NA | NA | NA | NA | Strong (+) | NA | NA |
| ELSI-Brazil | 1 (7175) | Limited (−) | Strong (−) | NA | NA | Limited (+) | NA | NA | Strong (+) | Strong (+) | NA |

Mr. OS and Ms. OS (Hong Kong) is a cohort study to examine the determinants of osteoporotic fractures in older Chinese men and women, (+) Positive result rating, (−) Negative result rating, (±) Conflicting, (?) unknown. ELSA: English Longitudinal Study on Ageing, ATHLOS: Ageing Trajectories of Health-Longitudinal Opportunities and Synergies, CCRC: Taikang Yanyuan continuing care retirement community in China, CHARLS: China Health and Retirement Longitudinal Study, ELSI-Brazil: Brazilian Longitudinal study of Aging, NA: Not available

composite scores can also be made more valid and reliable. The variety of instruments currently available requires an in-depth understanding of their measurement properties to make an informed decision about which tool to select and how to assess IC in an aging population.

To the best of the present authors' knowledge, only one rapid review[24] has previously provided an assessment based on psychometric properties recommended in the COSMIN guidelines. However, their search strategy was lacking because most of their included studies were retrospective in nature. However, the present systematic review included studies that were mostly prospective; therefore, the present review provides a better evaluation of the psychometric properties of instruments concerning IC. The review here might aid emerging studies attempting to assess IC at a population scale when designing their research instruments. The strength of the present review is the detailed and systematic electronic database search strategy used, which was based on the application of the COSMIN as well as the use of the most up-to-date methodology, whereby quality assessments were performed by using both the COSMIN checklist and applying the quality criteria for good psychometric properties.

### Limitations

Despite these strengths, the present systematic review has some limitations. First, the review did not search for gray literature (i.e., those from the Google or Google Scholar search engines). Therefore, the coverage of the included studies in the systematic review might be somewhat restricted. However, the present review aimed to evaluate the psychometric properties of IC instruments. Therefore, it is important to analyze the studies receiving rigorous peer review for the present systematic review. Accordingly, it can be tentatively concluded that the lack of gray literature in the present review might not have any severe biases on the findings. Second, diverse IC instruments were identified in the present systematic review. However, it was unable to compare them because each of the IC instruments did not have much psychometric evidence on them. Therefore, the present findings could not provide a strong recommendation regarding which IC instrument is most preferred. Third, the review was unable to conduct a meta-analysis due to the diversity of the IC instruments. Therefore, the evaluation of the psychometric properties of these IC instruments was based on qualitative synthesis rather than a more quantitative one. In the future, meta-analyses might be needed to assess the overall psychometric properties of each specific IC instrument when the evidence becomes sufficient. Finally, this systematic review only included studies published in English, which may have excluded relevant research in other languages.

## Conclusion

Several measures of IC have been used in the context

**Table 5: Measurement tools and methods used for intrinsic capacity domains**

| Author/region | Locomotion | Vitality | Cognition | Psychological | Sensory |
|---|---|---|---|---|---|
| Beard *et al*. (2019)/ UK[5] | Walk 8 feet (2.4 m) at their usual walking pace<br>Chair-stand test<br>Side-by-side, semi-tandem, and full tandem of static balance | Handgrip strength<br>Forced expiratory volume<br>DHEA(S) levels<br>Hemoglobin level<br>Insulin-like growth factor 1 | Verbal fluency<br>Delayed verbal memory<br>Attention | Five of the eight CES-D items (i.e., felt depressed, was happy, felt lonely, enjoyed life, felt sad)<br>Sleep disturbance | Hearing impairments were measured using self-reported<br>Vision impairments were measured using self-reported |
| Daskalopoulou *et al*. (2019)/ Latin American[15] | Walking a km difficulty<br>Time in seconds taken to walk 10 m | Washing whole body difficulty<br>Using the toilet difficulty | Learn test<br>Delayed recall<br>Long memory test<br>Immediate recall<br>Verbal fluency<br>Time orientation<br>Praxis-fold a piece of paper<br>Story recall difficulty | Sleep trouble or recent change in pattern<br>Feeling of not coping properly with everyday routine<br>Gets worn out or exhausted during daytime or evening | Hearing problem<br>Eye problem |
| Ma *et al*. (2020)/ China[21] | Chair rises within 14 s | Weight loss (>3 kg over the previous 3 months)<br>Appetite loss | Orientation in time and space<br>Recall the three words | Feeling down, feeling depressed or hopeless over the past 2 weeks<br>Having little interest or pleasure in doing things over the past 2 weeks | Whisper test<br>Vision impairments were measured using self-reported |
| Sanchez-Niubo *et al*. (2020)/ Australia, China, Europe, UK, Spain, Japan, Korea, India, Mexican, Irish Global[16] | Stooping, kneeling or crouching<br>Lifting or carrying weights<br>Climbing stairs<br>Getting up from sitting down<br>Walking by yourself and without any equipment<br>Pulling or pushing large objects<br>Sitting for long periods<br>Reaching or extending arms<br>Walking speed<br>Dizziness when walking on a level surface<br>Picking up things with fingers | Experiences in some degree of pain<br>Having high level of energy<br>Urine incontinence | Memory<br>Immediate recall<br>Delayed recall<br>Verbal fluency<br>Orientation in time<br>Processing speed<br>Numeracy | Sleeping | Near vision<br>Far vision<br>Eyesight using glasses or lens as usual<br>Hearing in general<br>Hearing in a conversation |
| Liu *et al*. (2021)/ China[22] | Chair rises within 14 s | Weight loss (>3 kg over the previous 3 months)<br>Appetite loss | Orientation in time and space<br>Recall the three words | Feeling down, feeling depressed, or hopeless over the past 2 weeks<br>Having little interest or pleasure in doing things over the past 2 weeks | Hearing impairments were measured using self-reported<br>Vision impairments were measured using self-reported |
| Yu *et al*. (2021)/ Hong Kong[17] | Walking speed<br>Assessed the time required to rise from a chair to a full standing position five times with arms folded across the chest<br>Dynamic balance | Grip strength<br>Appendicular skeletal muscle mass | 30-item MMSE | 15-item GDS | Frisby Stereo test<br>Snellen "Tumbling E" chart |

*Contd...*

**Table 5: Contd...**

| Author/region | Locomotion | Vitality | Cognition | Psychological | Sensory |
|---|---|---|---|---|---|
| Beard *et al*. (2021)/ China[18] | Walk 8 feet (2.4 m) at their usual walking pace<br>Chair-stand test<br>Side-by-side, semi-tandem, and full tandem of static balance | Handgrip strength<br>Forced expiratory volume<br>DHEA(S) levels<br>hemoglobin level<br>Insulin-like growth factor 1 | Episodic memory<br>Some components of the TICS battery | 10-item CES-D scale<br>Sleep hours at night; nap minutes at noon; and sleep quality | Hearing impairments were measured using self-reported<br>Vision impairments were measured using self-reported |
| Yu *et al*. (2022)/ Hong Kong[23] | Walking speed<br>Assessed the time required to rise from a chair to a full standing position five times with arms folded across the chest<br>Dynamic balance | Handgrip strength<br>Appendicular skeletal muscle mass | 30-item MMSE | 15-item GDS | Frisby stereo test<br>Snellen "Tumbling E" chart |
| Aliberti *et al*. (2022)/ Brazil[19] | Gait speed was calculated by measuring the time to walk three meters at the usual pace<br>The balance test from the SPPB was applied | Handgrip strength<br>Weight loss<br>Self-report exhaustion<br>Poor endurance | Temporal orientation<br>Episodic memory<br>Semantic memory semantic verbal fluency task (executive functioning, vocabulary size, and lexical access speed) | 8-item CES-D scale<br>How would you evaluate the quality of your sleep?<br>During the last month, have you taken any sleeping pill? | "How do you evaluate your hearing?"<br>How good is your eyesight for seeing things at a distance, like recognizing a friend across the street?"<br>How good is your eyesight for seeing things up close like reading ordinary newspaper print?" |
| Gao *et al*. (2022)/ China[20] | Running or jogging walking, getting up climbing stooping, kneeling, or crouching were measured using self-reported | Reaching or extending arms lifting or carrying weights<br>Picking up<br>Dressing<br>Bathing or showering<br>Eating<br>Getting into or out of bed<br>Using the toilet<br>Controlling urination and defecation<br>Doing household chores<br>Preparing hot meals<br>Shopping<br>Managing money<br>Taking medications | Numeracy<br>Orientation in time<br>Immediate recall<br>Delayed recall | Bothering<br>Attention<br>Depressed<br>Energy<br>Hopefulness<br>Fearfulness<br>Restless<br>Happiness<br>Loneliness<br>Hopelessness | How do you evaluate your hearing?<br>How good is your eyesight for seeing things at a distance, like recognizing a friend across the street?<br>How good is your eyesight for seeing things up close like reading ordinary newspaper print?<br>How would you rate your memory at the present time? |

SPPB: Short physical performance battery, DHEA(S): Dehydroepiandrosterone, TICS: Telephone interview of cognitive status, CES-D: Center for epidemiological studies-depression, GDS: Geriatric Depression Scale, MMSE: Mini-mental state examination

of healthy aging, including self-reported questionnaires, performance-based tests, and laboratory studies. Currently, there is heterogeneity in the measurement process used for assessing IC domains, particularly in the vitality and psychological domains. To date, there is no standard IC score for clinical or community-based settings. Obtaining in-depth knowledge regarding IC is essential to understanding how it is built over the lifespan and how

cost-effective population-wide interventions could enhance IC in future generations and delay functional decline in current aging cohorts.


### Acknowledgment

This study was supported by a fund from the National Health Research Institutes, Taiwan (NHRI-11A1-CG-CO-04-2225-1).

### Financial support and sponsorship

This study was supported by a fund from the National Health Research Institutes, Taiwan (NHRI-11A1-CG-CO-04-2225-1).


### Conflicts of interest

Chung-Ying Lin is the co-editor-in-chief of the *Asian Journal of Social Health and Behavior*; this paper went through rigorous peer review and revision.

## References


1. Zhou Y, Ma L. Intrinsic capacity in older adults: Recent advances. Aging Dis 2022;13:353-9.
2. World Health Organization. World Report on Ageing and Health. Geneva: World Health Organization; 2015.
3. Wang J, Boehm L, Mion LC. Intrinsic capacity in older hospitalized adults: Implications for nursing practice. Geriatr Nurs 2017;38:359-61.
4. Cesari M, Araujo de Carvalho I, Amuthavalli Thiyagarajan J, Cooper C, Martin FC, Reginster JY, *et al.* Evidence for the domains supporting the construct of intrinsic capacity. J Gerontol A Biol Sci Med Sci 2018;73:1653-60.
5. Beard JR, Jotheeswaran AT, Cesari M, Araujo de Carvalho I. The structure and predictive value of intrinsic capacity in a longitudinal study of ageing. BMJ Open 2019;9:e026119.
6. Guralnik JM, Simonsick EM, Ferrucci L, Glynn RJ, Berkman LF, Blazer DG, *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 1994;49:M85-94.
7. Vellas B, Guigoz Y, Garry PJ, Nourhashemi F, Bennahum D, Lauque S, *et al.* The mini nutritional assessment (MNA) and its use in grading the nutritional state of elderly patients. Nutrition 1999;15:116-22.
8. Belloni G, Cesari M. Frailty and intrinsic capacity: Two distinct but related constructs. Front Med (Lausanne) 2019;6:133.
9. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med 2006;119:166.e7-16.
10. Gonzalez-Bautista E, Andrieu S, Gutiérrez-Robledo LM, García-Chanes RE, de Souto Barreto P. In the quest of a standard index of intrinsic capacity. A critical literature review. J Nutr Health Aging 2020;24:959-65.
11. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Ann Intern Med 2009;151:264-9, W64.
12. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
13. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC. Rating the methodological quality in systematic reviews of studies on measurement properties: A scoring system for the COSMIN checklist. Qual Life Res 2012;21:651-7.
14. Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, *et al.* How to select outcome measurement instruments for outcomes included in a "core outcome set" – A practical guideline. Trials 2016;17:449.
15. Daskalopoulou C, Chua KC, Koukounari A, Caballero FF, Prince M, Prina AM. Development of a healthy ageing index in Latin American countries – A 10/66 dementia research group population-based study. BMC Med Res Methodol 2019;19:226.
16. Sanchez-Niubo A, Forero CG, Wu YT, Giné-Vázquez I, Prina M, De La Fuente J, *et al.* Development of a common scale for measuring healthy ageing across the world: Results from the ATHLOS consortium. Int J Epidemiol 2021;50:880-92.
17. Yu R, Amuthavalli Thiyagarajan J, Leung J, Lu Z, Kwok T, Woo J. Validation of the construct of intrinsic capacity in a longitudinal Chinese cohort. J Nutr Health Aging 2021;25:808-15.
18. Beard JR, Si Y, Liu Z, Chenoweth L, Hanewald K. Intrinsic capacity: Validation of a new WHO concept for healthy aging in a longitudinal Chinese study. J Gerontol A Biol Sci Med Sci 2022;77:94-100.
19. Aliberti MJ, Bertola L, Szlejf C, Oliveira D, Piovezan RD, Cesari M, *et al.* Validating intrinsic capacity to measure healthy aging in an upper middle-income country: Findings from the ELSI-Brazil. Lancet Reg Health Am 2022;12:100284.
20. Gao J, Xu J, Chen Y, Wang Y, Ye B, Fu H. Development and validation of a multidimensional population-based healthy aging scale: Results from the china health and retirement longitudinal study. Front Med (Lausanne) 2022;9:853759.
21. Ma L, Chhetri JK, Zhang Y, Liu P, Chen Y, Li Y, *et al.* Integrated care for older people screening tool for measuring intrinsic capacity: Preliminary findings from ICOPE pilot in China. Front Med (Lausanne) 2020;7:576079.
22. Liu S, Yu X, Wang X, Li J, Jiang S, Kang L, *et al.* Intrinsic capacity predicts adverse outcomes using integrated care for older people screening tool in a senior community in Beijing. Arch Gerontol Geriatr 2021;94:104358.
23. Yu R, Leung J, Leung G, Woo J. Towards healthy ageing: Using the concept of intrinsic capacity in frailty prevention. J Nutr Health Aging 2022;26:30-6.
24. George PP, Lun P, Ong SP, Lim WS. A rapid review of the measurement of intrinsic capacity in older adults. J Nutr Health Aging 2021;25:774-82.
25. Terwee CB, Prinsen CA, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, *et al.* COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. Qual Life Res 2018;27:1159-70.
26. Biemer PP, Groves RM, Lyberg LE, Mathiowetz NA, Sudman S. Measurement Errors in Surveys. New Jersey: John Wiley &Sons; 2013.

**Appendix S: Quality criteria for rating measurement properties**

| Measurement property | Rating | Rating quality criteria |
|---|---|---|
| Reliability | | |
| Internal consistency | + | Cronbach α between 0.70 and 0.95 OR KR-20 between 0.70 and 0.90 |
| | − | Cronbach α <0.70 OR KR-20 <0.70 |
| | ? | Cronbach α not reported |
| Reliability | + | ICC >0.70 OR weighted κ>0.70 OR Pearson *r*≥0.80 |
| | − | ICC ≤0.70 OR weighted κ≤0.70 OR Pearson *r*<0.80 |
| | ? | Neither ICC, weighted κ, nor Pearson *r* determined |
| Measurement error | + | MIC >SDC OR MID >SDC OR MIC outside LoA |
| | − | MIC ≤SDC OR MID ≤SDC OR MIC equals or inside LoA |
| | ? | MIC not defined |
| Validity | | |
| Content validity | + | Target population considers all items in the questionnaire to be relevant AND considers the questionnaire to be complete |
| | − | Target population considers items in the questionnaire to be irrelevant OR considers the questionnaire to be incomplete |
| | ? | No target population involved |
| Structural validity | + | Factors should explain ≥50% of the variance |
| | − | Factors should explain <50% of the variance |
| | ? | Explained variance not mentioned |
| Construct validity/ hypothesis testing | + | (Correlation with an instrument assessing the same construct ≥0.50 OR ≥75% of the results were in accordance with the hypotheses) AND correlation with related constructs was higher than with unrelated constructs |
| | − | Correlation with an instrument assessing the same construct <0.50 OR <75% of the results were in accordance with the hypotheses OR correlation with related constructs was lower than with unrelated constructs |
| | ? | Sole correlations determined with unrelated constructs |
| Cross-cultural validity | + | (Original factor structure confirmed OR no important differential item functioning between language versions) AND the correlation between the translated or culturally adapted version and the original versions was ≥0.70 |
| | − | Original factor structure not confirmed OR important differential item functioning found between language versions OR the correlation between the translated or culturally adapted version and the original versions was <0.70 |
| | ? | Confirmatory factor analysis not applied AND differential item functioning not assessed |
| Criterion validity (predictive/ concurrent) | + | Correlation with standard was ≥0.70 OR AUC ≥0.70 OR no statistically significant differences between the walking test and the criterion standard were found OR sensitivity and specificity ≥0.70 |
| | − | Correlation with standard was <0.70 OR AUC<0.70 OR no statistically significantly differences between the walking test and the criterion standard were found OR sensitivity and specificity <0.70 |
| | ? | No convincing arguments that criterion standard is actually the best standard OR doubtful design or method |
| Responsiveness | + | (Correlation with an instrument assessing the same construct ≥0.50 OR ≥75% of the results were in accordance with the hypotheses OR AUC ≥0.70 OR sensitivity and specificity ≥0.70) AND correlation with related constructs was higher than with unrelated constructs |
| | − | Correlation with an instrument assessing the same construct <0.50 OR <75% of the results were in accordance with the hypotheses OR AUC <0.70 OR sensitivity and specificity ≤0.70 OR correlation with related constructs was higher than with unrelated constructs |
| | ? | Sole correlations determined with unrelated constructs |

AUC: Area under the receiver operating characteristics curve, ICC: Intraclass correlation coefficient, KR-20: Kuder-Richardson formula(s), LoA: Limit of agreement, MIC: Minimal important change, MID: Minimal important difference, SDC: Significant detectable change, N: No, Y: Yes, +: Measurement property evident, −: No measurement property evident, ?: Indeterminate