Psychometric properties of the Barthel Index for evaluating physical function among Chinese oldest-old

Chi Zhang1, Xuxi Zhang2, Hao Zhang3, Ping Zeng4, Pengbin Yin4,5, Zhongquan Li6, Yali Zhao7* & Yao Yao2,8*

1The Key Laboratory of Geriatrics, Beijing Institute of Geriatrics, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing Hospital, National Center of Gerontology of National Health Commission, Beijing, China; 2Center for Healthy Aging and Development Studies, National School of Development, Peking University, Beijing, China; 3Department of healthcare policy and research, Weill Cornell Medicine, New York, NY, USA; 4Department of Orthopedics, The Fourth Medical Centre, Chinese PLA General Hospital, Beijing, China; 5National Clinical Research Center for Orthopedics, Sports Medicine & Rehabilitation, Beijing, China; 6School of Social and Behavioral Sciences, Nanjing University, Nanjing, China; 7Central Laboratory, Hainan Hospital of Chinese PLA General Hospital, Sanya, China; 8China Center for Health Development Studies, Peking University, Beijing, China

Abstract

Background The Barthel Index (BI) is among the most widely used instruments for evaluating physical function, yet its applicability has not been well studied in the oldest-old population.

Objective To test the psychometric properties of the BI for evaluating physical activities of daily living (ADL) in a large representative sample of oldest-old population in China.

Methods Participants were 1750 oldest-old adults (aged 80–116 years, 72.11% female) including 956 centenarians. ADL were assessed during face-to-face interviews. Multiple methodologies were applied to evaluate the reliability, validity and measurement invariance of the BI. An item response theory (IRT) framework was conducted to estimate the parameters of each item.

Results 48.91% participants had function dependence. Cronbach’s α coefficient of the BI was 0.902, but ‘Stair climbing’ impaired the overall internal consistency. The known-group validity of the BI was confirmed by significant differences in the BI score across age (P < 0.001), gender (P < 0.001), education (P < 0.001) and ethnicity (P = 0.038). The criterion-related validity was supported by significant correlations between the BI score with depression symptoms (r = −0.36, P < 0.001), subjective well-being (r = 0.23, P < 0.001) and self-report health status (r = 0.22, P < 0.001). Factor analysis yielded a two-factor structure (somatic function and physiological self-care) with appropriate invariance. Ten items showed acceptable discrimination parameters (1.80–5.87) and difficulty parameters (2.65–1.11) but had variant test information (1.73–10.22). ‘Bower control’ and ‘Bladder control’ were not conducive to the local independence.

Conclusions The BI has appropriate reliability, validity and measurement precision for community-based Chinese oldest-old and centenarians, but individual items have low quality. Somatic disability and incontinence are two latent categories of functional dependence in this population. Living environment needs to be taken into consideration for ADL instrument development and modification.

Keywords Barthel Index; Psychometric property; The oldest-old; Centenarians

Received: 11 November 2021; Accepted: 6 December 2021

*Correspondence to: Yao Yao, China Center for Health Development Studies, Peking University, Beijing, China. Email: yao.yao@bjmu.edu.cn; Yali Zhao, Central Laboratory, Hainan Hospital of Chinese PLA General Hospital, Sanya, China. Email: zhaoyl301@163.com
Introduction

The ability to perform activities of daily living (ADL) is an essential component of healthy aging. With the global population ageing, the oldest-old have been the fastest-growing segment and more susceptible to ADL decline and functional dependence. According to China’s General Program for Sustainable Development, China was projected to becoming a super-aged society by 2033 and having the largest population of the oldest-olds across the globe. Conveniently and accurately evaluating ADL impairments for the vulnerable oldest-old using appropriate tools requires considerable attention. The Barthel Index (BI), originally developed for stroke rehabilitation, has been widely used for evaluating ADL in healthcare and community settings. The BI can be used for a variety of clinical purposes, such as determining the degree of disability, evaluating treatment effects and identifying home care needs. In clinical practice, subjects with a total BI score of less than 90 or 95 are considered to be function dependent. Although substantial literature describing the BI’s applicability in general population is available, questions remain regarding certain properties among oldest-old population, especially in those from low- and middle-income countries (LMICs).

The BI measures a person’s physical function through 10 basic daily tasks (grooming, bathing, feeding, toileting, stair climbing, dressing, bowel control, bladder control, transferring and ambulation) regarding the ability of performing ADL. Identifying the latent categories of functional dependence according to the BI’s factor structure is beneficial for targeting long-term care and clinical intervention. However, whether the BI is a single measurement or a multiple-dimensional instrument among oldest-old is still unclear. A number of validity researches demonstrated the unidimensionality of the 10-item BI, whereas a few studies from Asian countries including China obtained a two-factor structure. Notably, specific self-report ADL tasks (bathing, toileting and stair climbing) are closely related to the living environment and facility, which have been confirmed to be associated with functional disability among older adults in the United Kingdom and China. The interaction between living environments and specific ADL task reinforces the need for in-depth analyses of item quality in function assessment tools. Item response theory (IRT) analyses, which have many advantages over the classical test theory (CTT), are applicable for evaluating individual item’s contribution to a scale through parameter estimation. In the last two decades, this psychometric methodology has been increasingly applied in comprehensive geriatric assessment.

The BI has been validated by previous studies among hospitalized and community population. Shahram et al. confirmed the BI to be a reliable and valid questionnaire stroke clinical trials. In another study, the BI showed acceptable fit of a unidimensional structure among geriatric rehabilitation patients. However, the majority of previous validity studies of the BI generally focused on middle-aged or older populations from Western countries with only one study focused on the oldest-old yet with small sample size. As many cultural-related diversity exist regarding residential environments and culture gaps of older people between LMICs and Western world, whether previous results derived from developed countries could apply well to oldest-old adults in LMICs such as China remains unclear. To address these gaps, we comprehensively evaluated the reliability, validity and factorial invariance of the BI in a large sample of the oldest-old and centenarian Chinese population. We also employed the IRT framework to evaluate the quality of each item.

Materials and methods

Study populations

The data used in this study was collected from two-stage investigations in the China Hainan Centenarian Cohort Study (CHCCS) from June 2014 to December 2017. The goal of CHCCS is to assess the physical function, mental health, and social status of longevous adults and establish ageing indicators for healthy longevity. Details of sampling strategies, inclusion criteria, and age verification were described in our previously published profile. In total, 1794 old persons aged 80 or above by household register of the Department of Civil Affairs in Hainan were invited to participate, and 794 oldest-old participants (aged 80–99 years) and 956 centenarians were involved in the final study. For the centenarian participants, strict three-step age verification methods were adopted to ensure the age authenticity, and 58 subjects who failed to meet the validation were excluded. First round: housed register, identification card, census-derived list; Second round: claimed birthday and Chinese zodiac sign, claimed age when they experienced the specific social upheaval; Third round: milestone assessments such as marriage date, date of first born, subsequent birthdates of children, mother’s death. In the current study, 124 subjects who were unable to cooperate due to dementia or paralysis were excluded before the survey, and participants who failed to complete any question in the BI were also excluded. The flow chart of participant recruitment in this study was described in Figure 1.

The ethics committee of the Hainan Hospital of the Chinese People’s Liberation Army General Hospital approved the study protocol (serial no. 301hn11201601). All subjects (or guardians) provided written informed consent before participating in the survey. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.
Procedures

For the survey data used in this study, participants were interviewed at their residences or nearby health service centres by a well-trained interdisciplinary medical group. The face-to-face interview was performed by native nurses who were trained in interviewing older adults and able to speak the local dialect. For the ADL measurement, investigators checked the participants’ self-report function performance according to their performance during the clinical examinations. There were 3–21 missing values for the 10 items in the BI. Considering the subjects’ age and poor health condi-
tions, missing data of individual BI questions were coded as incapacity with a score of ‘0’ (dependence). Participants’ response and the corresponding missing values are described in Supplementary Table 1. Demographic characteristics including age, gender, ethnicity, education, residence, marital and birth status were collected. To test the criterion-related validity of the BI, related health outcomes (depressive symptoms, subjective well-being and self-report health status [SRH]) were also obtained. Subjects with good literacy and cognitive function completed the survey independently, whereas subjects who were illiterate answered questions with the cooperation of the investigator and their legal representative.

Measures

We used a Chinese BI version firstly translated by Leung and colleagues in 2007 using standard translation methods to assess the ADL of the participants. The BI consists of 10 items: grooming, bathing, feeding, toileting, stair climbing, dressing, bowel control, bladder control, ambulation and chair-bed transferring. Due to the generally low educational level and potential sensory impairment of oldest-old adults living in Hainan, research team further simplified the linguistic expressions of the BI’s questions and responses and tested the feasible and valid in a pilot survey in 2014. The scoring of BI was provided in Supplementary Table 2. In brief, participants’ physical performance was assessed on a 2–4 Likert classified response depending on the amount of assistance required for accomplishing the tasks in daily living. Items were scored on a scale of ‘0’, indicating dependence, to ‘5, 10 or 15’, indicating ‘needs help’ to ‘independent’. A higher score indicates higher levels of function, and the maximum summed score of 100 indicates complete independence. When the total score is ≥95, the subject is considered to have good functional ability or independent; participants with a total score of less than 95 were considered dependent.

Depressive symptoms were measured using the Geriatric Depression Scale (GDS-15), which consists of 15 dichotomous items that assess the presence of depressive symptoms in the last week. A higher score indicates more depressive symptoms (a possible range of 0–15; the observation range is 0–15), and as previously described, Chinese oldest-old with a score of 7 or higher were considered to have a depressive symptom. The Satisfaction with Life Scale (SWLS) consisting of five 7-point Likert items was used to assess participants’ subjective well-being aspect (a possible range of 0–35; the observation range is 0–35). SRH was measured using a Visual Analogue Scale (VAS). The VAS score records an individual’s self-rated health on a 20-cm, vertical visual analogue scale ranging from 0 to 100, with notes at both ends labelling ‘the worst health you can imagine’ (at 0) and ‘the best health you can imagine’ (at 100).

Statistical analysis

Continuous variables with a normal distribution were described as the mean ± standard deviation (m ± SD), continuous variables with a skewed distribution were presented as median (interquartile range), and categorical variables were described as numbers with percentages [n (%)]. As for the skewed distribution of the summed BI score, the Wilcoxon or Kruskal–Wallis rank sum test was used to compare the differences of the BI scores across different groups. Spearman correlation coefficient was used to analyse the correlation between the BI score and other measured outcomes. Internal consistency of the BI was evaluated using Cronbach’s α coefficient and item-total correlation. An α coefficient > 0.7 indicates good internal reliability. An exploratory factor analysis (EFA) using the principal component method was performed to explore the factor structure. The eigenvalue criterion (>1) of each factor was used to identify the number of meaningful factors. Confirmatory factor analyses (CFA) with robust weighted least squares estimation were conducted in Mplus (Version 7.4) in cases of violation of the multivariate normality assumption. The χ²/df, root mean square error of approximation (RMSEA), comparative fit index (CFI) and normed fit index (NFI) were used to evaluate fitness. The model is considered to have a good fit with a χ²/df of 5 or less, an RMSEA of 0.1 and a CFI and NFI greater than 0.90.\(^{23}\) Factorial invariance of the BI across age, gender, education, ethnicity and residence was tested by a multi-group confirmatory factor analysis (MGCFA) approach, which consisted of a series of nested confirmatory steps. Configural invariance (free parameters), metric invariance (constraints of equivalent factor loadings), scalar invariance (further constraints of the intercepts) and strict invariance (further constraints of residual variances) models were tested across subgroups. A non-significant \(\Delta \chi^2 (P > 0.05)\), a \(\Delta \text{CFI} < 0.01\) and a \(\Delta \text{RMSEA} < 0.015\) were used to compare the fit of nested models.\(^{24}\)

Since subjects in the current study would not rely guessing to answer questions about physical function, we estimated the discrimination (a) and difficulty (b) parameters of 10 items in the BI using the IRTPRO Version 4.2 software. As the previous research did, a two-parameter logistic (2PL) model was used for two binary items (‘grooming’ and ‘bathing’), and a graded response model was used for the other eight multigraded items.\(^{20,25}\) In clinical settings, the discrimination between \([-2.80, +2.80]\) is suitable, and the difficulty is acceptable at \([-3, +3]\).\(^{25}\) Primarily, we tested the assumptions of an item response analysis. The fit of each item under the particular models was tested using the S-\(\chi^2\) statistic suggested by Orlando and Thissen.\(^{26}\) A non-significant S-\(\chi^2\) statistic \((P > 0.05)\) indicates fitness to the model. The unidimensionality assumption was tested according to results from factor analyses. The local independence was confirmed using the \(\chi^2\)LD statistic recommended by Orlando.\(^{26}\) A \(\chi^2\)LD
statistic < 10 between any two items indicated an acceptable level of local independence, which means that subjects’ item responses are independent after controlling for the underlying trait. The item characteristic curve (ICC) was used to establish the relationship between subjects’ potential trait and their responses, and the item information curve (IIC) was used to evaluate the measurement precision through the test information function (TIF). The measurement precision of a scale is sufficient when the total test information is above 25.28

Results

Demographic characteristics

Among 1750 participants, 54.63% were centenarians, 72.11% were female, 84.46% were illiterate, 89.14% were Han ethnic, and 63.26% were rural residents (Table 1). The ADL score was 95 (75, 100), and 48.91% had function dependence. The GDS-15 score was 4 (3, 7), and 25.49% was screened as having depressive symptoms. The mean SWLS score was 95 (75, 100), and 48.91% had function dependence. The mean VAS score was 61.92 ± 15.26. Participants who were older, female, illiterate, ethnic, had sensory impairment, and had depressive symptoms showed lower summed BI scores (P < 0.05).

Reliability, validity and factor structure of the BI

The summed BI score was significantly correlated with GDS-15 (r = –0.36, P < 0.001), SWLS (r = 0.23, P < 0.001) and VAS (r = 0.22, P < 0.001). As Table 2 showed, the overall α coefficient of the BI was 0.903 and increased only when ‘stair climbing’ was deleted. The item-total correlation coefficients ranged from 0.508–0.789 (Ps<0.001). Two factors explaining 73.64% of the total variance were retained in the EFA, and all item loadings on the corresponding factor were greater than 0.6. Factor 1 consisting of eight items (grooming, bathing, toileting, dressing, feeding, stair climbing, transfer and ambulation) was defined as ‘somatic function’; Factor 2 consisting of two items (bowel control and bladder control) was defined as ‘physiological self-care’.

Considering previous studies obtained a unidimensional structure of the BI and the item ‘stair climbing’ impaired the overall internal consistency in the current study, we involved four competition models to explore the optimal structure of the BI in CFAs. Model A: two factors with 10 items; Model B: delete ‘stair climbing’ from Model A; Model C: one factor with 10 items; Model D: delete ‘stair climbing’ from Model C. As described in Table 3, all the four CFA models showed acceptable RMSEA (<0.1), CFI(>0.9) and NFI (>0.9), whereas the Model A had a smaller χ²/df (5.538) and RMSEA (0.051). Next, we tested the factorial invariance of the two-factor BI model (Model A) across age, gender, education, residence, and ethnicity using MGCFA (Table 4). The configural invariance model (free parameters) was used as a basic model, and three restrictive models (restrict loading, intercept and residual sequentially) were tested in a stepwise manner. The invariance of metric, scalar, and residual restricted models was confirmed (△CFI < 0.01, △NFI < 0.01, △RMSEA<0.015) across residence and ethnicity subgroups. For gender and education subgroups, the invariance of metric and scalar restricted models was confirmed, but the residual restricted model was not (△CFI > 0.01, △NFI > 0.01). For age subgroups, the metric restricted model met the invariance criteria (△CFI = 0.010, △NFI = 0.009, △RMSE = 0.011).

Table 1 Demographic characteristics and physical function of 1750 oldest-old in Hainan

| Characteristics            | n (%)     | ADL Median | Interquartile | Z      | P-value |
|---------------------------|-----------|------------|---------------|--------|---------|
| Age group                 |           |            |               |        |         |
| 80–99                     | 794(45.37)| 100        | 95,100        | 22.633 | <0.001  |
| 100+                      | 956(54.63)| 85         | 60,95         |        |         |
| Gender                    |           |            |               |        |         |
| Male                      | 488(27.89)| 100        | 90,100        | 10.119 | <0.001  |
| Female                    | 1262(72.11)| 90       | 70,100        |        |         |
| Education                 |           |            |               |        |         |
| Illiterate                | 1478(84.46)| 90      | 75,100        | 8.127  | <0.001  |
| Literate                  | 272(15.54)| 100        | 90,100        |        |         |
| Residence                 |           |            |               |        |         |
| Rural                     | 1107(63.26)| 95      | 80,100        | 0.322  | 0.748   |
| Urban                     | 643(36.74)| 95         | 70,100        |        |         |
| Ethnicity                 |           |            |               |        |         |
| Han                       | 1560(89.14)| 90      | 75,100        | 2.071  | 0.038   |
| Minority                  | 190(10.86)| 95         | 80,100        |        |         |
| Visual impairment         |           |            |               |        |         |
| Yes (with 86 missing)     | 1280(73.14)| 90      | 70,100        | 6.294  | <0.001  |
| No                        | 470(26.86)| 95         | 85,100        |        |         |
| Auditory impairment       |           |            |               |        |         |
| Yes (with 96 missing)     | 604(34.51)| 95         | 85,100        | 8.634  | <0.001  |
| No                        | 1146(65.49)| 90      | 70,100        |        |         |
| Self-reported health      |           |            |               |        |         |
| Good                      | 242(13.83)| 95         | 80,100        | 50.109 | <0.001  |
| General (with 11 missing) | 1194(68.23)| 95     | 85,100        |        |         |
| Bad                       | 314(17.94)| 80         | 50,90         |        |         |
| Depressive symptoms       |           |            |               |        |         |
| Yes                       | 446(25.49)| 80         | 50,95         | 13.751 | <0.001  |
| No                        | 1304(74.51)| 95       | 85,100        |        |         |

ADL, activities of daily living.
Item characteristics of the BI

As described in Supplementary Table 3, no statistically significant $S \chi^2$ was found for the 10 items (ranged from 10.05 to 41.41, $Ps > 0.05$), which indicated that the parametric models were fit to the data. In the EFA, the ratio of the first eigenvalue (6.014) to the second eigenvalue (1.351) was greater than 3 (Table 2), and the one-factor Model C (CFI = 0.988, RMSEA = 0.057) showed acceptable fitness. Thus, the unidimensional hypothesis was supported by Hambleton's criteria. As Supplementary Table 4 described, the $\chi^2$/LD statistic between ‘bowel control’ and ‘bladder control’ was 63.7, which was the highest, and the $\chi^2$/LD statistic between any other two items (0.1–8.4) were all less than the criterion of 10. The ICC and IIC of 10 items in the BI were summarized in Figure 2, and estimated parameter values were shown in Table 5. The discrimination parameter of 10 items (ranged from 10.05 to 13.97) was far greater than 3 ($>3$), thus the unidimensional hypothesis was supported by the unidimensional hypothesis. In the CFA, the model with one factor showed the best fit to the data, as $\chi^2$ and $\chi^2$/df were both < 3. The goodness of fit of the BI was 0.989 and RMSEA was 0.057. The BI was shown to be a unidimensional scale with a total test information of 43.85, which corresponded to the latent trait level of 0.94. When ‘stair climbing’, ‘bowel control’ and ‘bladder control’ were deleted from the BI, there was no obvious loss of total information function.

Discussion

This is the first study to evaluate the psychometric properties of the BI among oldest-old and centenarian populations with a considerable sample size. We observed significant variances of the participants’ ADL score across age, gender, education and health status subgroups, which confirmed the known-group validity of the BI. Significant correlations between the BI score with measurements of GDS-15, SWLS, and LAS supported the BI’s criterion-related validity with was reported in most previous validity studies. The BI had a satisfactory internal consistency coefficient of 0.903, which was generally consistent with that reported in previous studies ranging from 0.69 to 0.96. The dimensionality of
the BI is not consistent across all the studies. EFA and CFA showed that the two-factor structure was an optimal solution in Chinese oldest-old population. Studies from Iran, Sri Lanka and China obtained the same two-dimensional factorial structure in older adults. However, a number of studies supported the comparability and robust property of the BI. For instance, despite concerns that demographic differences exist between the oldest-old and centenarians, the age invariance indicated that subjects across the two subgroups responded to the scale with the same underlying framework.

In IRT analyses, the highest $\chi^2/df$ statistic (63.7) between the BI factor of ‘bowel control’ and ‘bladder control’ indicated that they seriously destroyed the BI’s local independence. Local independence of a valid scale means that the subject’s latent trait is the only factor affecting their response. In terms of psychometric properties, the two-factor structure showed acceptable invariance across age, gender, education, residence, and ethnicity subgroups, which supported the comparability and robust property of the BI. For instance, despite concerns that demographic differences exist between the oldest-old and centenarians, the age invariance indicated that subjects across the two subgroups responded to the scale with the same underlying framework.

In IRT analyses, the highest $\chi^2/df$ statistic (63.7) between ‘bowel control’ and ‘bladder control’ indicated that they seriously destroyed the BI’s local independence. Local independence of a valid scale means that the subject’s latent trait is the only factor affecting their response. In terms of psychometric properties, the two-factor structure showed acceptable invariance across age, gender, education, residence, and ethnicity subgroups, which supported the comparability and robust property of the BI. For instance, despite concerns that demographic differences exist between the oldest-old and centenarians, the age invariance indicated that subjects across the two subgroups responded to the scale with the same underlying framework.
measure a different concept of the other items.\textsuperscript{37} Hartingsveld \textit{et al.}\textsuperscript{38} also found that incontinence items impeded the BI’s one-dimensional property. Although ‘bowel control’ and ‘bladder control’ provided less information than other somatic items (Table 5), these two incontinence items are necessary to identify potential categories of disability when using the BI. All the 10 BI items showed sufficient discrimination parameters (>1). Another IRT study conducted by Liu and colleagues found that most items in the BI showed similar good discrimination, but two items (‘stair climbing’ and ‘feeding’) had very low discrimination parameters among older persons from long-term care facilities.\textsuperscript{20} The common low difficulty parameter found in the current study (–2.65 to 1.11) was derived from the ceiling effect of the BI, which has been frequently reported in previous validation studies.\textsuperscript{39,40} The BI had sufficient measurement precision with a total test information of 43.85. In Liu’s study,\textsuperscript{20} the BI’s information function was as sufficient as ours (TIF > 40), and the corresponding location was also relatively low (\(\theta < 0\)). Notably, the TIF of the 10 individual items ranged from 1.37 to 10.22, which meant that the contribution of individual items to the scale was quite different. Particularly, the TFI of ‘stair climbing’ was the lowest, and this item impairs the overall internal consistency. The majority of partic-

**Figure 2** Item characteristic curves and item information curves of 10 items in the Barthel Index. The item characteristic curve (ICC) is represented by a solid line, and the item information curve (IIC) is represented by a dotted line.

**Table 5** Item analyses of the Barthel Index based on item response theory

| Item | Contents       | Parameters estimate | Information (\(\theta\)) |
|------|----------------|---------------------|--------------------------|
| 1 Grooming | 5.17 0.58 –1.04 0.05 | —— —— —— —— | 6.61 (–1.01) |
| 2 Bathing | 5.87 0.67 –0.85 0.05 | —— —— —— —— | 8.46 (–0.94) |
| 3 Toileting | 4.43 0.34 –1.16 0.06 | —— —— —— —— | 6.76 (–1.46) |
| 4 Bowel control | 1.99 0.19 –2.46 0.16 | —— —— —— —— | 2.19 (–2.07) |
| 5 Bladder control | 1.80 0.17 –2.65 0.19 | —— —— —— —— | 2.56 (–2.14) |
| 6 Dressing | 5.79 0.54 –1.35 0.06 | —— —— —— —— | 10.22 (–1.63) |
| 7 Feeding | 3.21 0.24 –1.74 0.08 | —— —— —— —— | 3.49 (–1.95) |
| 8 Stair climbing | 1.62 0.12 –0.08 0.06 | —— —— —— —— | 1.73 (–0.33) |
| 9 Transferring | 3.06 0.20 –1.50 0.07 | —— —— —— —— | 3.69 (–1.56) |
| 10 Ambulation | 3.37 0.22 –1.28 0.06 | —— —— —— —— | 4.09 (–1.48) |

S.E., standard error.
\*Discrimination parameters of items.
\*Difficulty parameters of items for response category ‘1’.
\*Difficulty parameters of items for response category ‘2’.
\*Difficulty parameters of items for response category ‘3’.
Participants in the current study lived in bungalows and did not need to climb stairs. Therefore, ‘stair climbing’ seemed to be a redundant task in population from current study. Although this findings might not be completely applicable to other older groups, we argued that living conditions were to precise measurement of ADL for a specific population. Similar viewpoints were reported in Tomoko et al.’s study, which focused on individuals that lived in long-term healthcare facilities that had either an elevator or no stairs. In this study, we provided psychometric evidence and methodology suggestion for the development and modification of function assessment item banks. Comparability of ADL measurement should be linked to the diversity of living facilities for older people in different areas. For conveniently measurement purpose, researchers have suggested that the standard BI versions should be modified to accommodate different populations with varied living settings. A recent systematic review found that six simplified versions of BI containing three to eight items were also used well in patients with stroke.

One strength of this study is the unique study population with a considerable sample size. Another strength is the combination of psychometrical methodologies of CTT and IRT. This study has some limitations. Firstly, the sample used in this study originated from one province in China, and generalizability to older people with varied background needs to be further verified. In any case, we have provided commonly practical methodological suggestions for assessing and optimizing ADL assessment tools. Secondly, due to the subjects’ advanced age and cognitive impairment, information bias was inevitable. However, we have adopted strict quality assurance approaches to ensure the authenticity of subjects’ ages and ADL measurements during the investigation. Thirdly, clinical factors such as chronic diseases, geriatric syndromes and polypharmacy were not included, which might influence the psychometric evaluation of the BI. Fourthly, we could not analyse the test–retest and inter-rater reliability of the BI due to the lack of repeated measurement data. Future studies are justified to investigate this issue.

**Conclusion**

The BI has appropriate reliability, validity and measurement precision for community-based Chinese oldest-old and centenarians. Somatic disability and incontinence are two latent categories of functional dependence in this population. Living environment needs to be taken into consideration for ADL instrument development and modification.

**Funding**

This work was supported by grants from National Key R&D Program of China [Grant No. 2018YFC2000400], Natural Science Foundation of China [Grant Nos. 81903392 and 81941021] and the Fundamental Research Funds for the
Author contributions

CZ, YZ, and YY proposed the concept and design, analysed and interpreted the data, and wrote the manuscript; CZ, XZ, and YY interpreted the data, drafted and edited the manuscript, supervised the study, and obtained funding; HZ, PZ, PY and ZL drafted and edited the manuscript. All authors read and approved the final version of the manuscript. YZ and YZ are guarantors.

Acknowledgements

We are grateful to the CHCCS study, which provided the data in this research. We also thank for the support from the Healthy Aging Consortium of the China Cohort Consortium (see http://chinacohort bjmu.edu.cn/). The authors certify that they comply with the ethical guidelines for publishing in the Journal of Cachexia, Sarcopenia and Muscle: update 2017.

References

1. Katz S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. J Am Geriatr Soc 1983;31:272–277.

2. Lunney JR, Lynn J, Foley DJ, Lipson S, Guralnik JM. Patterns of functional decline at the end of life. JAMA 2003;289:2387–2392.

3. Luo Y, Su B, Zheng X. Trends and challenges for population and health during population aging — China, 2015-2050. China CDC Weekly 2021;3:593–599.

4. Corrao S, Argano C, Natoli G, Nobili A, Coraza GR, Mannucci PM, et al. Disability, and not diabetes, is a strong predictor of mortality in oldest old patients hospitalized with pneumonia. Eur J Intern Med 2018;54:53–59.

5. Tettamanti M, Marcon G. Cohort profile: ‘Centenari a Trieste’ (CaT), a study of the health status of centenarians in a small defined area of Italy. BMJ Open 2018;8:e019250.

6. Quinn TJ, Quinn TJ, McArthur K, Ellis G, Stott DJ. Functional assessment in older people. BMJ 2011;343:1–7.

7. de Wit I, Putman K, Devos H, Brinkmann N, Dejaeger E, de Weerdt W, et al. Long-term prediction of functional outcome after stroke using single items of the Barthel index at discharge from rehabilitation centre. Disabil Rehabil 2014;36:353–358.

8. Shah S. In praise of the biometric and psychometric qualities of the Barthel Index. Physiotherapy 1994;80:769–771.

9. Laake K, Laake K, Laake P, Ranhoff AH, Sween U, Wyller TB, et al. The Barthel ADL index: factor structure depends upon the category of patient. Age Ageing 1995;24:393–397.

10. Leung SO, Cc C, Shah S. Development of a Chinese version of the modified Barthel index — validity and reliability. Clin Rehabil 2007;21:912–922.

11. Hormozi S, Alizadeh-Khoei M, Sharifi F, Taati F, Aminaliroya R, Fadase S, et al. Iranian version of Barthel index: validity and reliability in outpatient’s elderly. Int J Prev Med 2019;1:130.

12. Liu C, Fong JH, Hoh JWT, Wu H, Dong Y, Gu D, et al. Bathing and toileting difficulties of older adults in rural China: the role of environment. BMC Geriatr 2020;20:533.

13. Jagger C, Arthur AJ, Spiers NA, Clarke M. Patterns of onset of disability in activities of daily living with age. J Am Geriatr Soc 2001;49:404–409.

14. Chakravarty EF, Jb B, Fries JF. Improving patient reported outcomes using item response theory and computerized adaptive testing. J Rheumatol 2007;34:1426–1431.

15. Gill TM. Assessment of function and disability in longitudinal studies. J Am Geriatr Soc 2010;58:S308–S312.

16. Jette AM, Haley SM, Ni P, Olarsch S, Moed R. Creating a computer adaptive test version of the late-life function and disability instrument. J Gerontol A Biol Sci Med Sci 2008;63:1246–1256.

17. Oveisgharan S, Shirani S, Ghobani A, Soltanzade A, Baghaei A, Hosseini S, et al. Barthel index in a Middle-East country: translation, validity and reliability. Cerebrovasc Dis 2006;22:350–354.

18. Bouwstra H, Smit EB, Wattel EM, van der Wouden JC, Hertogh C, Terluin B, et al. Measurement properties of the Barthel index in geriatric rehabilitation. J Am Med Dir Assoc 2018;20:420–425.

19. Hartigan I. A comparative review of the Katz ADL and the Barthel index in assessing the activities of daily living of older people. Int J Older People Nurs 2007;2:204–212.

20. Liu W, Unick J, Galik E, Resnick B. Barthel Index of activities of daily living: item response theory analysis of ratings for long-term care residents. Nurs Res 2015;64:88–99.

21. He Y, Zhao Y, Yao Y, Yang S, Li J, Liu M, et al. Cohort profile: The China Hainan centenarian cohort study (CHCCS). Int J Epidemiol 2018;47:694–695.

22. Tang WK, Chan SS, Chiu HF, Wong KS, Kwok TC, Mok V, et al. Can the geriatric depression Scale detect poststroke depression in Chinese elderly? J Affect Disord 2004;81:153–156.

23. Bollen AK. Structural equations with latent variable. New York: Wiley; 1989.

24. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Modeling 2002;9:233–255.

Online supplementary material

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Three-steps age validation process of centenarians in CHCCS.

Table S1. 1750 participants’ response ratio of 10 items in the Barthel Index

Table S2. Rating scale of activities of daily living by the Barthel Index

Table S3. $\chi^2$ statistics diagnostic statistics of 10 items in the Barthel Index

Table S4. The $\chi^2$LD statistic matrix of 10 items in the Barthel Index

Conflict of interest

None declared.
25. Thomas ML. The value of item response theory in clinical assessment: a review. Assessment 2011;18:291–307.

26. Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Measur 2000;24:50–64.

27. Smith EV Jr. Effect of item redundancy on Rasch item and person estimates. J Appl Meas 2005;6:147–163.

28. Lai JS, Cella D, Chang CH, Bode RK, Heinemann AW. Item banking to improve, shorten and computerize self-reported fatigue: an illustration of steps to create a core item bank from the FACIT-Fatigue Scale. Qual Life Res 2003;12:485–501.

29. Hambleton RK. Emergence of item response modeling in instrument development and data analysis. Med Care 2000;38:i60–i65.

30. Isahaya K, Takao N, Tsuchihashi Y, Akiyama H, Hasegawa Y. Validity of the remote evaluation of NIH stroke scale, modified Rankin scale and Barthel index using a tablet device. Jpn J Stroke 2019;41:368–374.

31. Unnanuntana A, Jarusriwanna A, Nepal S. Validity and responsiveness of Barthel index for measuring functional recovery after hemiarthroplasty for femoral neck fracture. Arch Orthop Trauma Surg 2018;138:1671–1677.

32. Morley D, Selai C, Thompson A. The self-report Barthel Index: preliminary validation in people with Parkinson’s disease. Eur J Neurol 2012;19:927–929.

33. Maclsaac RL, Ali M, Taylor-Rowan M, Rodgers H, Lees KR, Quinn TJ. Use of a 3-item short-form version of the Barthel index for use in stroke: systematic review and external validation. Stroke 2017;48:618–623.

34. Sherwood S, Morris J, Mor V. Compendium of measures for describing and assessing long-term care populations. Boston, MA: Hebrew Rehabilitation Center for the Aged; 1977.

35. Lekamwasam S, Karunatilake K, Lekamwasam V. Physical dependency of elderly and physically disabled; measurement concordance between 10-item Barthel index and 5-item shorter version. Ceylon Med J 2011;56:114–118.

36. Ohura T, Higash T, Ishizaki T, Nakayama T. Assessment of the validity and internal consistency of a performance evaluation tool based on the Japanese version of the modified Barthel index for elderly people living at home. J Phys Ther Sci 2014;26:1971–1974.

37. Organization WH. International classification of functioning, disability and health. Geneva: WHO; 2001.

38. de Morton NA, Ji K, Davidson M. Rasch analysis of the Barthel index in the assessment of hospitalized older patients after admission for an acute medical condition. Arch Phys Med Rehabil 2008;89:641–647.

39. Van Hartingsveld F, Lucas C, Kwakkel G, Lindeboom R. Improved interpretation of stroke trial results using empirical Barthel item weights. Stroke 2006;37:162–166.

40. Saito T, Izawa KP, Matsui N, Arai K, Ando M, Morimoto K, et al. Comparison of the measurement properties of the functional independence and difficulty scale with the Barthel index in community-dwelling elderly people in Japan. Aging Clin Exp Res 2017;29:273–281.

41. Vikman T, Fielding P, Lindmark B, Fredrikson S. Effects of inpatient rehabilitation in multiple sclerosis patients with moderate disability. Adv Physiother 2008;10:58–65.

42. Ohura T, Hase K, Nakajima Y, Nakayama T. Validity and reliability of a performance evaluation tool based on the modified Barthel index for stroke patients. BMC Med Res Methodol 2017;17:131–131.

43. von Haehling S, Morley JE, Coats AJS, Anker SD. Ethical guidelines for publishing in the journal of cachexia, sarcopenia and muscle: update 2017. J Cachexia Sarcopenia Muscle 2017;8:1081–1083.