BMJ Open

Measurement of the severity of disability in community-dwelling adults and older adults: interval-level measures for accurate comparisons in large survey data sets

José Buz,[1] María Cortés-Rodríguez[2]

To cite: Buz J, Cortés-Rodríguez M. Measurement of the severity of disability in community-dwelling adults and older adults: interval-level measures for accurate comparisons in large survey data sets. *BMJ Open* 2016;**6**: e011842. doi:10.1136/bmjopen-2016-011842

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-011842).

Received 10 March 2016
Revised 30 June 2016
Accepted 27 July 2016

[1]Department of Developmental Psychology, University of Salamanca, Salamanca, Spain
[2]Faculty of Sciences, Department of Statistics, University of Salamanca, Salamanca, Spain

Correspondence to
Dr José Buz; buz@usal.es

ABSTRACT

Objectives: To (1) create a single metric of disability using Rasch modelling to be used for comparing disability severity levels across groups and countries, (2) test whether the interval-level measures were invariant across countries, sociodemographic and health variables and (3) examine the gains in precision using interval-level measures relative to ordinal scores when discriminating between groups known to differ in disability.

Design: Cross-sectional, population-based study.

Setting/participants: Data were drawn from the Survey of Health, Ageing and Retirement in Europe (SHARE), including comparable data across 16 countries and involving 58 489 community-dwelling adults aged 50+.

Main outcome measures: A single metric of disability composed of self-care and instrumental activities of daily living (IADLs) and functional limitations. We examined the construct validity through the fit to the Rasch model and the know-groups method. Reliability was examined using person separation reliability.

Results: The single metric fulfilled the requirements of a strong hierarchical scale; was able to separate persons with different levels of disability; demonstrated invariance of the item hierarchy across countries; and was unbiased by age, gender and different health conditions. However, we found a blurred hierarchy of ADL and IADL tasks. Rasch-based measures yielded gains in relative precision (11–116%) in discriminating between groups with different medical conditions.

Conclusions: Equal-interval measures, with person-invariance and item-invariance properties, provide epidemiologists and researchers with the opportunity to gain better insight into the hierarchical structure of functional disability, and yield more reliable and accurate estimates of disability across groups and countries. Interval-level measures of disability allow parametric statistical analysis to confidently examine the relationship between disability and continuous measures so frequent in health sciences (eg, cholesterol, blood pressure, C reactive protein).

Strengths and limitations of this study

- This is the first study that provides a Rasch-based single metric of disability to be used for accurate comparisons of disability severity levels across groups/countries and their relationships with external variables.
- We empirically assess the reliability of scores using Rasch modelling to address the misuse of estimating reliability by means of Cronbach's α in highly skewed distributions with marked ceiling/floor effects.
- The measurement of disability with reliable interval-level measures is a cost-effective and efficient approach to gain comprehensive data on persons with disabilities, thus providing important keys regarding how and when to promote prevention programmes, modify interventions or develop enabling environments.
- The examination of differential item functioning (DIF) by medical conditions and physical symptoms is limited to three broad groups. The presence of DIF with more specific health conditions, as well as contextual and environmental variables, should be investigated in future studies.
- Despite the advantages of a Rasch-based single metric of disability over separate scales with summative scores, our metric should be improved by adding more items of difficult tasks to adequately measure the lowest disability levels in the general population.

INTRODUCTION

The measurement of the severity of disability is a critical element for studying the causes and consequences of ageing and for planning health programmes and services.[1] Until now, having valid and reliable measures of disability based on survey data remains a major challenge. Activities of daily living (ADLs) and instrumental activities of daily living (IADLs) scales have shown construct

under-representation, lack of sensitivity to change, low discriminative power, presence of bias, and striking floor and ceiling effects in community-dwelling populations.[2–8] To overcome some of these problems, aggregated measures of ADLs and IADLs have been constructed.[2 9–16] In general, these studies have supported a single underlying dimension,[2 10 12 14–16] but they have also underlined serious concerns regarding the purported hierarchy of functional disability, and evidences of differential item functioning (DIF) regarding age and gender.[9 11 12 15] When ADL and IADL scales have been combined, the age-related and gender-related measurement bias was significantly attenuated.[2 15] Moreover, conducting parametric statistics with summative scores from these scales violates the fundamental assumption of equal-interval scaling and increases the probability of type I and II errors.[2 13 17 18] It has also been observed that summative scores of ADLs/IADLs underestimate mean disability in cross-cultural studies.[2] Summative scores obtained from hierarchical scales wrongly assume that (1) all items are measuring the same disability continuum, (2) each item contributes equally to the final score and (3) scores are not dependent on samples and items.

Frequently reported large floor and ceiling effects in ADL and IADL scales in relatively healthy populations also represent evident threats to validity and reliability but, surprisingly, their effects have been largely ignored. For example, the examination of the reliability of scores in ADLs, IADLs and mobility scales with more precise and more appropriate statistics than Cronbach's α has not been addressed.

A recognised advance in ensuring the quality of health-related instruments is the Rasch model, a parametric item response theory (IRT) model that transforms raw scores into interval-scaled measures, and allows the unequivocal confirmation of the formal item hierarchy.[10] According to the model, the probability of endorsing an item is a logistic function of the difference between the person's ability (latent trait, θ) and the item difficulty (δ). Thus, persons with low disability have a lower probability of being limited in easy activities (eg, eating), whereas more disabled persons have a higher probability of being limited in more difficult activities (eg, shopping). This is usually presented as follows:

$$P_i(X_{is} = 1) = e(\theta_{si} - \delta_{si})/[1 + (\theta_{si} - \delta_{si})]$$

X_{is} refers to a correct response (X=1) made by participant s to item i; θ_s refers to the trait level of participant s; δ_i refers to the difficulty of item i; e is the base of the natural logarithm (e=2.71828).

Persons and items are calibrated on a common interval-level scale (expressed in logits), so it is possible to assess how reliably persons and items can be hierarchically ordered from low to high levels of disability. A unique property of this model is *specific objectivity*, meaning that the estimation of item parameters is independent of the persons used (ie, person invariance), and that the estimation of the person parameters is independent of the particular items employed (ie, item invariance).[18] Finally, for the Rasch model, missing data do not cause bias or lower the precision of disability measurements.

The aim of this study is to provide a single metric of disability using Rasch modelling with data drawn from the Survey of Health, Ageing and Retirement in Europe (SHARE) to be used for disability severity comparisons across groups or countries. In addition to ADL and IADL items, we incorporate mobility tasks in order to expand the validity construct, based on the accumulative evidence suggesting that mobility limitations are a precursor of disability in ADLs and IADLs and that they are less affected by floor effects.[7 9 14 19] To the best of our knowledge, neither the precise severity level of aggregated ADL, IADL and mobility items has been estimated, nor has the ability of a single metric to separate persons with different levels of disability been established. We performed DIF to examine whether the measures were invariant across age, gender, medical conditions, symptomatology and self-rated health. Finally, we adopted the method of known-groups validity to examine the gains in precision using interval-level measures relative to ordinal scores for discriminating between groups known to differ in disability.

METHODS

Study design

Cross-sectional, population-based study.

Participants

Data were drawn from wave 4 (2010–2011) of SHARE including comparable data across 16 countries and involving 58 489 community-dwelling adults aged 50+. Representative samples from Austria, Belgium, the Czech Republic, Denmark, Estonia, France, Germany, Hungary, Italy, the Netherlands, Poland, Portugal, Slovenia, Spain, Sweden and Switzerland were obtained using probability samples. Methodological details of the survey are available elsewhere.[20 21] We excluded participants aged under 50 years (n=1254), with missing information across all ADL/IADL/mobility items (n=339), or institutionalised (n=368), which resulted in a final sample of 56 528 participants. Calibrated sampling weights were used to adjust for the complex sampling design.

Measures

Disability is measured in SHARE by asking respondents whether they had 'any difficulty' (yes=1, no=0), because of a physical, mental, emotional or memory problem, in carrying out daily activities (ADLs, six items; IADLs, seven items) and functional limitations (10 Nagi-based questions). ADLs included bathing, dressing, eating, getting into/out of bed, using the toilet and walking across a room. IADLs included making meals, shopping, doing work around the house/garden, making telephone calls,

using a map, medications and managing money. Mobility questions asked about kneeling, climbing one flight/several flights of stairs, walking 100 m, sitting for 2 hours, getting up from a chair, pulling large objects, lifting heavy weights, lifting hands above shoulders and picking up a small coin. The SHARE asked about any difficulty in physical functioning even with the help of assistive devices. No information about specific devices was gathered. Data were collected by the interviewer by means of Computer Assisted Personal Interviewing (CAPI). Showcards were used alongside CAPI.

Demographic and health variables: We included the following variables: (1) age, gender and years of education, using the UNESCO International Classification of Educational Degrees (ISCED-97); (2) self-reported illness diagnosed by a general practitioner (heart disease, hypertension, hypercholesterolaemia, stroke, diabetes, lung disease, asthma, arthritis, osteoporosis, cancer, ulcer, Parkinson disease, cataracts, hip fracture, other fractures, Alzheimer disease and benign tumour); (3) presence of long-term health problems that affect daily routines (yes/no); (4) self-reported physical symptoms (pain, angina or chest pain, breathlessness, persistent cough, swollen legs, sleeping problems, falling over and fear of falling, dizziness, stomach or intestine problems, incontinence and fatigue); and (5) self-rated health using a single question with answer categories ranging from 1=poor to 5=excellent.

Data analyses

Descriptive data

Demographic and health variables were examined using descriptive statistics. For subsequent analyses, we randomly split the sample into two subsamples: one for multigroup confirmatory factor analyses (MGCFA; n=28 788), and the other for Rasch-based analyses (n=27 740).

Multigroup confirmatory factor analysis

Before Rasch analysis was conducted, as recommended,[22] tests of measurement invariance were performed to establish whether the general factor structure (configural invariance) and the factor loadings (metric invariance) were the same across countries. Once we tested that the goodness of fit of the unidimensional model in each country was adequate, we conducted two hierarchically nested invariance models with increasingly restrictive constraints. To estimate the parameters, we used the diagonally weighted least squares and the asymptotic covariance matrix. Model fit can be considered good with root mean square error of approximation (RMSEA) $\leq$0.05 and comparative fit index (CFI) >0.90. The comparison for nested models was based on ΔCFI$\leq$0.01.[23] High floor/ceiling effects in categorical data can produce attenuated estimates of the correlation among indicators, lead to 'pseudofactors' that are artefacts of extremeness, and produce incorrect test statistics and SEs. Therefore, we carried out the analysis excluding extreme scores. The final sample included 15 325 participants.

Rasch analysis

We adopted a parametric model (Rasch modelling for dichotomous responses) for this work because it was appropriate for our purposes and had several advantages: (1) person-free and item-free invariant parameters can be estimated, (2) interval-level measures that show how much (more or less) ability or difficulty exists between persons or items are provided and (3) the estimates of person and item parameters can be represented graphically on a common metric to easily examine the scale targeting, construct validity and predictive validity.

Fit to the Rasch model was evaluated by the mean square fit statistics (infit MnSq and outfit MnSq) and Rasch residual-based principal components analysis (PCA). Mean square fit statistics indicate how much misfit is revealed in the actual data. Infit is a weighted fit statistic in which relatively more impact is given to unexpected responses close to a person's or item's measure. Outfit is an unweighted statistic that gives more impact to unexpected responses far from a person's or item's measure. The expected value for MnSq is close to 1.0 with an accepted range of 0.6–1.4 for surveys. Values $\geq$2.0 indicate a severe misfit.[24] In PCA, a strong measurement dimension for unidimensionality is achieved when the variance explained is >40%, and the eigenvalue of the first component of residuals is <2.0.[25]

Reliability was estimated with the Rasch-based person reliability (PR) and the person separation (Gp). PR is more precise and less misleading than Cronbach's α (KR-20) because (1) it provides a more detailed picture of the precision of measures, (2) statistics are estimated from linear measures and (3) it is not affected by extreme scores where error variance is the largest. Gp represents the scale's ability to separate the sample into different strata of disability (strata=(4Gp+1)/3). We also examined how precise the scale was at various ranges of the disability continuum to determine appropriate cut-off points by plotting the test information function (TIF) according to persons' ability. TIF is defined as the reciprocal of the precision with which a parameter is estimated. Score accuracy is high where SEs are low. PR$\geq$0.70 (for group comparisons), Gp$\geq$1.5, TIF$\geq$4 and SE around 0.5 are desirable values.[22 26]

The invariance of the item hierarchy across countries was evaluated by (1) intraclass correlation coefficients (ICCs) that indicated the overall agreement across the 16 countries and (2) a matrix of Spearman correlation coefficients that revealed the consistency between countries in the rank order of the item calibrations. Coefficients can be interpreted as follows: 0.6 or higher indicates moderate agreement; 0.7–0.8 indicates strong agreement and >0.8 indicates almost perfect agreement.[27]

The invariance of the item hierarchy across subgroups was examined with DIF analyses in five different groups:

age (<75 vs 75+), gender (male vs female), medical conditions (none vs 1+; ≤1 vs 2+), physical symptoms (none vs 1+; ≤1 vs 2+) and self-rated health (excellent/very good/good vs fair/poor). We used the Mantel-Haenszel model (MH) and the DIF CONTRAST estimate that calculates the difference between the estimators of the item parameter of difficulty for each group. In large samples, differences higher than 0.64 and 0.50 logits for MH and DIF CONTRAST, respectively, and statistically significant (with Bonferroni correction), are considered substantial.[24 28] To detect whether DIF may cause bias, we assessed its impact on the scale measures by examining differential test functioning.[22 29] We estimated a Rasch model for each group separately and the expected score was plotted against the measured disability dimension using test characteristic curves (TCCs). The area between the curves reveals the magnitude of bias.[15 30]

Relative precision

The relative precision (RP) method was used to compare the best performance between interval-level measures and summative scores for distinguishing disability severity levels among persons with different medical conditions. RP indicates how much more or less precise Rasch-based scores are relative to the ordinal scores. RP is calculated as the ratio of pairwise F statistics (the interval-level measure F statistics divided by the ordinal score F statistic).

Descriptive analyses and general linear models were conducted with SPSS V.21, MGCFA with LISREL V.8.80 and Rasch analyses with WINSTEPS V.3.70.

RESULTS

Demographic data

Table 1 shows the basic characteristics of participants in each country. The average age ranged from 64.5 to 69.2 years, with women representing ~55% of the sample within each country. Although in the majority of the countries more than half of the respondents reported having long-term illness and approximately two chronic conditions and physical symptoms, their self-rated health was good.

MGCFA analyses

As shown in table 2, the unidimensional solution showed a good model fit (RMSEA from 0.039 to 0.057) in all countries. All factor loadings were statistically significant ($p<0.01$) and salient. The subsequent configural and metric models showed good fit to the data and the restrictions imposed did not result in a significant drop in model fit.

Rasch analyses

Fit of persons and items to the Rasch model: As recommended,[31] the most misfitting persons (outfit MnSq>2.0) were removed because their inclusion distorted the person parameter estimates. We followed an iterative process by first removing the individuals with the highest outfit (MnSq=9.90, mainly as a result of unexpected responses by low and high disabled persons), and then by examining person estimates in each step. Separation and person reliability reached their highest values after excluding 1258 respondents. We did not find a pattern in the sociodemographic variables, health variables or across countries for those persons with idiosyncratic responses. The final sample included 26 482 respondents, including a low percentage of misfitting persons (2.8% with outfit MnSq ranging from 2.0 to 3.77). Statistics indicated a good model data fit for persons (mean infit MnSq=1.00, SD=0.31; mean outfit MnSq=0.71, SD=0.42) and for items (mean infit MnSq=0.98, SD=0.14; mean outfit MnSq=0.74, SD=0.42). The infit and outfit statistics for all the items were in an appropriate range. The low outfit MnSq (<0.60) statistics in ADL/IADL items indicated that they were too predictable. This overfit had no practical implications, except in situations of shortening scales, because these items did not degrade the measure. The PCA showed that the scale met the criterion for essential unidimensionality (44% of explained variance and eigenvalue of 1.7). Logits were transformed into more meaningful values from 0 (no disability) to 100 (highest disability; table 3).

Person–item targeting and item hierarchy: The item locations ranged from 3.06 logits for the easiest task (taking medicines) to −3.56 logits for the most challenging tasks (stooping, kneeling, crouching), indicating an adequate spread of disability levels (see table 4). The mean level of disability among participants ($\theta=-2.77$ logits) was lower than the average level of item difficulty ($\delta=0$), indicating that the scale was 'slightly off target' $2<|\theta-\delta|<3$ from the sample.[18] Thus, items that spread outside the range of persons did not contribute much to the measurement. The person–item map (figure 1) showed that the easiest tasks (eg, eating, taking medicines) were off-target even for persons located at or close to the average level of persons. This indicated that better targeted items at the lower end of the scale were appropriate for adequately measuring persons with the lowest disability levels. The addition of mobility tasks to ADLs and IADLs in a single metric yielded a lower percentage of persons with zero scores (floor effect=48.5%) than that resulting from separate scales (see table 1).

Regarding the hierarchy of functional decline, mobility tasks were, as expected, more challenging than IADLs and ADLs. However, IADLs were not clearly more challenging than ADLs. Specifically, some ADLs were more challenging (eg, 'dressing' or 'bathing') than some IADLs (eg, 'managing money' or 'preparing a hot meal'). Similarly, item location estimates for apparently similar activities (eg, 'walking 100 m' and 'walking across a room') were markedly different (−1.05 and 2.21 logits, respectively).

The rank ordering of the item difficulties was similar for all countries (Spearman correlation coefficients

Table 1 Demographic and health variables of participants aged 50+ in SHARE wave 4 (2010/11) by country

Variables	Austria	Germany	Sweden	The Netherlands	Spain	Italy	France	Denmark
Sample size, N	5044	1543	1919	2689	3477	3510	5515	2190
Age, mean (SE)	65.5 (0.15)	67.3 (0.28)	69.2 (0.26)	67.5 (0.21)	67.7 (0.26)	65.9 (0.33)	66.8 (0.19)	66.2 (0.24)
Range (IQR)	14 (50–97)	13 (50–100)	13 (50–99)	13 (50–98)	18 (50–102)	14 (50–100)	17 (50–104)	16 (50–100)
Sex (% men)	42.6	46.2	45	43.9	44	46.5	43.3	44.6
Marital status (%)								
Married	63.6	55.5	56.7	68.4	68.3	72.5	65.2	65.7
Single	8.5	2.2	2.2	4.4	8.1	7.5	8.5	7
Divorced	13	10.2	9.9	10.2	4.5	2.9	10.1	11.7
Widowed	14.9	32	31.2	17	19.1	17.1	16.2	15.6
Educational level (% ISCED)								
Low (0–2)	22.8	12.7	36.4	46	79.4	67.4	43.4	10
Medium (3–4)	50.8	60	36.7	25.6	10.5	26.7	36.3	34.9
High (5–6)	26.4	27.3	26.8	27.5	10	5.9	20.3	55.1
Multimorbidity, mean (SE)	1.7 (0.02)	1.7 (0.04)	1.4 (0.03)	1.4 (0.03)	2.0 (0.04)	1.5 (0.04)	1.6 (0.03)	1.5 (0.03)
Symptomatology, mean (SE)	1.7 (0.03)	2.1 (0.05)	1.7 (0.05)	1.6 (0.04)	2.1 (0.05)	1.7 (0.06)	2.2 (0.04)	1.6 (0.04)
Long-term illness (% yes)	46.9	66.6	55.2	52.2	54.4	39.5	49	48.2
Self-rated health (%)								
Excellent	9.4	3.6	16.3	10.6	4.3	8	6.1	19.4
Very good	25.9	12.9	23.9	16.5	14.4	17	14.6	33.7
Good	34.6	40.9	27.5	42	34.6	36.8	43	24.6
Fair	23.8	32.5	23.9	26.2	30.3	23.6	25.2	17.3
Poor	6.4	10.1	8.5	4.7	16.4	11.9	11	5.1
ADL (%)								
No limitations	90	96.7	89.2	93.3	85.5	89.1	88.4	93
IADL (%)								
No limitations	82.3	83.9	86.2	85.3	79.3	83.7	84.5	88.4
Mobility (%)								
No limitations	51.1	41.8	53	59.6	46.4	46	50.2	65.2

Variables	Switzerland	Belgium	Czechia	Poland	Hungary	Portugal	Slovenia	Estonia
Sample size (N)	3612	5053	5845	1704	2971	1991	2700	6667
Age, mean (SD)	66.3 (0.18)	66.2 (0.22)	65.6 (0.18)	67.7 (0.25)	65.6 (0.37)	64.5 (0.44)	65.9 (0.22)	67.1 (0.14)
Range (IQR)	15 (50–101)	17 (50–101)	14 (50–99)	14 (50–104)	13 (50–101)	15 (50–95)	16 (50–99)	16 (50–101)
Sex (% men)	45	46.4	41.7	41.6	39.8	44.1	42.1	36.8
Marital status (%)								
Married	64.2	65.9	67	43.4	57	80.6	62.7	50.5
Single	8.2	6.5	3.1	1.5	4	2.7	7	9.6
Divorced	15	12.8	12.6	3.6	10.8	4.1	6.1	14.7
Widowed	12.5	14.8	17.3	51.5	28.2	12.6	24.2	25.2
Educational level (% ISCED)								
Low (0–2)	16.3	43.6	42.2	20.5	34.7	54.1	35.8	32.1
Medium (3–4)	66.3	27.5	46.4	58.1	52.2	7.7	47.3	47.1
High (5–6)	17.3	29	11.4	21.4	13.1	38.2	16.9	20.8
Multimorbidity, mean (SE)	1.3 (0.02)	1.9 (0.03)	1.7 (0.03)	1.8 (0.04)	2.3 (0.07)	1.7 (0.09)	1.6 (0.03)	2.1 (0.02)
Symptomatology, mean (SE)	1.4 (0.03)	2.0 (0.04)	2.2 (0.04)	2.5 (0.06)	3.0 (0.12)	2.0 (0.11)	1.8 (0.04)	2.4 (0.03)

Continued

Table 1 Continued

Variables	Belgium	Czechia	Poland	Hungary	Portugal	Slovenia	Estonia	Switzerland
Long-term illness (% yes)	32.6	48	53.3	66.6	72.8	37.4	47.7	74.2
Self-rated health (%)								
Excellent	11.9	7.4	2.5	0.8	3.4	3.2	6.6	1.5
Very good	29.7	22.1	15.9	7.1	10.6	10.8	12.6	3.9
Good	39.6	42.3	40	33.9	23.5	29	36.6	22.9
Fair	15.5	22.1	29.1	34.5	34.5	38.2	28.4	48.8
Poor	3.2	6.2	12.5	23.8	27.8	18.8	15.8	22.8
ADL (%)								
No limitations	94.1	84.9	90.4	82.2	87	83	89.7	83
IADL (%)								
No limitations	91.2	80	81.3	79.1	72.2	78.5	82.8	76.9
Mobility (%)								
No limitations	65.1	45.4	46.9	39.7	35.4	40.3	43.7	42.1

All data are weighted, unless otherwise indicated. SE was calculated to adjust for the complex design; non-weighted ISCED.
ADL, activities of daily living; IADL, instrumental activities of daily living; ISCED, International Classification of Educational Degrees; SHARE, Survey of Health, Ageing and Retirement in Europe.

ranged from 0.88 to 0.99; table 5). The ICC for agreement in item hierarchy across all countries was high (ICC=0.94, 95% CI 0.90 to 0.97, p<0.001). Therefore, the scale demonstrated strong invariance of item hierarchy despite the environmental and cultural differences across countries.

Additionally, *specific objectivity* (generalisability) was empirically tested by randomly splitting the sample (n=13 870), calculating the difficulty estimates of the items, and conducting a linear regression analysis between the measures. The expected values for a perfect fit are 1, 0 and 1 for the correlation value, the intercept and the slope estimate, respectively. We found values of 0.997, 0.024 and 0.991, respectively, thus confirming objective specificity.

Reliability: As is shown in table 6, the reliability of the person ability estimates is 0.74 (person separation=1.70). Therefore, the scale was able to separate persons in two (nearly three) levels of disability.[24] This corroborates, in part, the aforementioned targeting problem regarding the person–item map. Visual analysis of TIF (figure 2) revealed that the score precision drops substantively as the scores approach the higher and lower ends. Thus, a cut-off of 11 (raw score) was the most appropriate to distinguish among disabled persons with low or high disability. Tentatively, cut-offs of 8 and 15 (raw score) could be used for low (1–8), moderate (9–14) and high (15+) levels of disability (see also figure 1).

In contrast, ADL and IADL scores from separate scales showed an insufficient reliability; person reliability, person separation and TIF indicated that these scores were not able to separate two distinct strata of persons with disability. SE revealed that the precision of scores was twice the desired value of 0.5. Gp≤1, and person reliability <0.50, imply that more than 50% of the differences between measures are due to measurement error.[24] Mobility scores showed slightly better results. From an epidemiological point of view, this finding suggests that, statistically, cut-off scores such as ADL 1+and IADL 1+ represent adequately the boundary between 'non-disabled' and 'disabled' persons, but additional cut-off scores are not appropriate.

Differential item functioning: DIF was found in four items as a function of age. Difficulty estimates were significantly greater for the younger respondents compared with the older respondents (75+) on 'sitting 2 hours' and 'getting in/out of bed', while 'shopping' and 'managing money' were more difficult for the older respondents compared with the younger respondents. Across gender, 'lifting over 5 kilos' showed a higher difficulty estimate for males, while 'dressing' and 'preparing hot meals' showed a higher difficulty estimate for females. No further DIF was found. TCCs for age and gender groups revealed that their expected and observed scores matched almost perfectly, indicating that items displaying DIF were not causing bias.

Relative precision: As can be seen in table 7, interval measures produced gains in RP in all of the medical

Table 2 Goodness of fit indices for measurement invariance model comparisons across 16 countries

	N	RMSEA 90% CI	CFI	SRMR	GFI	ΔCFI
Unidimensional model						
Austria	1285	0.039 (0.036 to 0.043)	1	0.07	0.99	
Germany	460	0.040 (0.038 to 0.045)	1	0.08	0.98	
Sweden	475	0.053 (0.051 to 0.060)	0.99	0.08	0.97	
The Netherlands	614	0.050 (0.044 to 0.056)	0.99	0.10	1	
Spain	942	0.056 (0.053 to 0.066)	0.98	0.07	0.97	
Italy	965	0.057 (0.054 to 0.066)	0.99	0.10	0.97	
France	1467	0.052 (0.049 to 0.058)	0.99	0.09	0.99	
Denmark	400	0.051 (0.048 to 0.064)	1	0.13	0.98	
Switzerland	691	0.048 (0.044 to 0.053)	0.99	0.10	1	
Belgium	1478	0.050 (0.046 to 0.053)	0.99	0.07	0.98	
Czechia	1650	0.045 (0.042 to 0.048)	1	0.07	0.98	
Poland	497	0.049 (0.039 to 0.051)	1	0.07	0.98	
Hungary	961	0.055 (0.051 to 0.060)	0.98	0.11	0.97	
Portugal	628	0.052 (0.047 to 0.058)	0.99	0.08	0.97	
Slovenia	769	0.053 (0.049 to 0.063)	0.99	0.07	0.98	
Estonia	2043	0.047 (0.044 to 0.050)	1	0.07	0.98	
Configural		0.044 (0.036 to 0.049)	0.99			
Metric		0.061 (0.058 to 0.065)	0.99			≤0.01

Factor loadings in the metric invariance model were: walking 100 m (λ=0.75), sitting 2 hours (λ=0.42), getting up from a chair (λ=0.62), climbing stairs (λ=0.64), climbing a stair (λ=0.70), kneeling (λ=0.64), reaching arms (λ=0.53), pulling large objects (λ=0.74), lifting heavy weights (λ=0.68), picking a small coin (λ=0.51), dressing (λ=0.71), walking across a room (λ=0.87), bathing (λ=0.86), eating (λ=0.77), getting into/out of bed (λ=0.77), toileting (λ=0.88), using a map (λ=0.59), preparing a hot meal (λ=0.88), shopping (λ=0.87), telephone calls (λ=0.79), taking medications (λ=0.86), housework (λ=0.75), managing money (λ=0.74).
CFI, comparative fit index; GFI, goodness of fit index; RMSEA, root mean square error of approximation; SRMR, standardised root mean square residual; ΔCIF, comparative fit index difference test.

conditions (above 50% in 9 out of 16 comparisons). Specifically, Rasch-based measures were two times more effective than summative scores for detecting differences in disability in persons 'diagnosed vs non-diagnosed' as having osteoporosis or benign tumour. Interval measures were also ~70% better at discriminating between diagnosed and non-diagnosed hypertension, cholesterol, asthma or arthritis. Low gains were observed for medical conditions such as Alzheimer disease, Parkinson and hip fracture.

DISCUSSION

Principal findings

Our study presents a hierarchical scale with equal-interval measures and person-invariant and item-invariant properties to measure disability severity in community-dwelling adults and older adults. We provide strong evidence regarding the hierarchical structure of functional disability, independent of country, age, gender, medical conditions, symptomatology and self-rated health.

Fit statistics, PCA and invariance analyses showed that the single metric of disability achieved the requirements of a strong hierarchical scale. Our findings support previous studies suggesting that ADL, IADL and mobility items contributed a unidimensional construct of disability.[14 15 32] In addition to this, the property of specific objectivity facilitates the generalisability of results. As regards, we aim to address the most recent claims resulting from public health studies[33] for the need to create composite measures of disability that permit accurate comparisons of functional status across and within countries.

Differential item functioning

Our findings coincide with research showing DIF by age and gender.[9 12] However, we did not find evidence of bias.[15] It is important to note that our results are not completely comparable to previous studies that examined the 'need for help' instead of the 'difficulty with' daily activities. Plausibly, the 'need for help' is more dependent on social network availability, gender roles and culture, among other variables; hence, the existence of DIF can be expected. Furthermore, we also demonstrated that the scale was not biased by medical conditions, symptomatology and self-reported health. Therefore, researchers can use it confidently for comparisons of disability in adults and older adults with a wide variety of health conditions. This is an important contribution because previous studies have only focused on age and gender, and the impact of health-related variables has not been addressed.

We examined DIF in heterogeneous groups according to the number, but not the type, of self-reported diseases and symptomatology, and therefore did not explore the risk of bias associated with specific diseases or symptoms when performing different activities. Previously, a cross-cultural adaptation of the Functional Independence

Table 3 Normative measures for the disability scale across countries

Raw score (0–23)	Measure	Rescaled measure (0–100)	SE
0*	−5.86	0	1.88
1	−4.51	11	1.10
2	−3.62	19	0.84
3	−3.01	24	0.73
4	−2.52	28	0.67
5	−2.10	32	0.63
6	−1.73	35	0.60
7	−1.38	38	0.58
8	−1.05	41	0.56
9	−0.74	44	0.56
10	−0.43	46	0.55
11	−0.13	49	0.55
12	0.17	52	0.55
13	0.47	54	0.55
14	0.78	57	0.56
15	1.09	60	0.57
16	1.42	62	0.58
17	1.76	65	0.59
18	2.13	68	0.62
19	2.53	72	0.65
20	2.99	76	0.71
21	3.56	81	0.81
22	4.41	88	1.07
23*	5.72	100	1.87

Transformation of the raw scores to interval-level scores on a logit scale.
*Extreme scores measure inaccurate estimate of disability.
A higher raw score indicates a higher disability. As can be seen, the 1-point difference between a raw score of 15 and 16 points is 2 on the interval scale, whereas the 1-point difference between a raw score of 1 and 2 points is 8 on the interval scale.

Measure (FIM) for patients with stroke showed that different calibrations for several items were necessary.[34] Thus, future cross-cultural studies could assess DIF across subpopulations with specific medical conditions and settings in order to ensure the comparability of disability measures.

Contextual and environmental factors can also affect the calibration of items and distort outcome measures. As has been previously stated,[35] differences in the estimates of disability are caused by theoretical perspectives, methodological issues (eg, wording or response categories) and environmental factors. The 'difficulty with' or the 'need for help' with specific activities may be largely mediated/moderated by environmental variables. In practical terms, it is possible that calibrations (δ) for some daily activities can change in different geographical or cultural contexts (eg, 'dressing' is probably more challenging in Finland than in Bora Bora). Other factors affecting the estimates of disability are related to the availability of personal and social resources (income, spouse, education, etc), or even the use of assistive devices,[35] which is an issue that should be investigated in the future. Additionally, the analysis of DIF within IRT is a useful mechanism to evaluate the impact of these factors on disability estimates and make the appropriate adjustments (ie, different calibrations).

Relative precision

We demonstrated gains in RP for comparisons of disability severity using interval measures (averaged gained 58%) in all of the medical conditions. These gains have occurred mainly through greater differences between groups in scores at the lower extreme of the distribution, where the relationship between raw scores and Rasch measures is non-linear (as at the upper end). This is an important issue because large survey population studies have to face the challenge of comparing groups/countries with low and/or similar disability levels. Rasch measures and summative scores showed similar precision when comparing diagnosed and non-diagnosed participants with Alzheimer, Parkinson disease and hip fracture. Neuropsychological diseases and fractures produce severe disability levels involving instrumental and self-care activities of daily living. Mean scores of disability of participants diagnosed with these medical conditions indicate that they are located near to the middle of the distribution (eg, mean=10.89 for Alzheimer disease), where the relationship between raw scores and Rasch measures is linear. In these conditions, parametric analyses conducted with raw scores may yield an accurate comparison of groups. Although we have not measured change in scores over time, the advantages in the precision of Rasch measures are also applicable in longitudinal design studies.[36]

Scale targeting and hierarchical structure

Despite the aforementioned positive findings, there are some issues that cause concern. The first one is related to construct under-representation. The item–person map revealed that the scale is better targeted at more disabled people than those less disabled. Paradoxically, epidemiological studies attempt to target relatively healthy respondents (at the low end of the distribution) in order to better plan health, social and long-term care services. Off-target scales negatively affect the precision of the item estimates, do not make for an efficient measurement and do not provide enough information along the desired population range.[24] The expected positive effect of adding mobility limitations to our metric in order to expand the construct may have been cancelled out by the inclusion of relatively healthy adults aged 50+. Previous authors have demonstrated that the dimensionality of ADL/IADL items could vary depending on if disabled or non-disabled people were included in the analysis.[9] Therefore, we carried out an additional analysis, selecting persons aged 65 years and over (n=14 339; results not shown but available on request), to observe the impact on reliability and targeting. We found a lowering in the floor effect (from 48.5% to 35%), but a similar reliability (PR=0.77, separation=1.86) and

Table 4 Fit statistics and hierarchy of the disability items

	Activity	Infit MnSq	Outfit MnSq	Location	SE
Stooping, kneeling or crouching	Mobility	1.04	1.00	−3.56	0.02
Climbing several flights of stairs without resting	Mobility	1.02	0.98	−3.36	0.02
Lifting or carrying weights over 10 pounds/5 kilos, like a heavy bag of groceries	Mobility	0.99	0.96	−2.57	0.02
Getting up from a chair after sitting for long periods	Mobility	1.08	1.10	−2.24	0.02
Pulling or pushing large objects like a living room chair	Mobility	0.97	0.91	−1.66	0.02
Climbing 1 flight of stairs without resting	Mobility	1.02	0.98	−1.22	0.02
Sitting for about 2 hours	Mobility	1.38	1.76	−1.07	0.03
Walking 100 m	Mobility	0.93	0.84	−1.05	0.03
Doing work around the house or garden	IADL	0.89	0.74	−1.04	0.03
Reaching or extending your arms above shoulder level	Mobility	1.21	1.44	−0.78	0.03
Dressing, including putting on shoes and socks	ADL	0.99	0.78	−0.21	0.03
Using a map to figure out how to get around in a strange place	IADL	1.14	1.05	0.04	0.03
Shopping for groceries	IADL	0.80	0.43	0.23	0.03
Bathing or showering	ADL	0.79	0.41	0.46	0.04
Getting into or out of bed	ADL	0.95	0.45	0.97	0.04
Managing money, such as paying bills and keeping track of expenses	IADL	1.00	0.50	1.23	0.04
Picking up a small coin from a table	Mobility	1.19	0.69	1.31	0.04
Preparing a hot meal	IADL	0.81	0.29	1.41	0.05
Using the toilet, including getting up or down	ADL	0.83	0.21	2.06	0.06
Walking across a room	ADL	0.82	0.17	2.21	0.06
Making telephone calls	IADL	0.93	0.20	2.83	0.07
Eating, such as cutting up your food	ADL	0.93	0.20	2.96	0.08
Taking medications	IADL	0.90	0.17	3.06	0.08

Item difficulty, items ordered according to the expected hierarchy of difficulty from the easiest to endorse (stopping, kneeling or crouching= −3.56) to the most difficult (taking medications=3.06 logits).
ADL, activities of daily living; IADL, instrumental activities of daily living; MnSq, mean square residual.

targeting (mean person score=−2.26), which represented a non-significant improvement. Attempts to expand the construct of disability in a single metric for a community-dwelling population should include mental health functions, more infrequent and demanding tasks, physical performance measures, sensory and communicating limitations, as well as pain, fatigue and tiredness[35] to better target the general population.

Another aspect to which some thought must be given is related to the hierarchical structure of disability. It has been widely accepted that ADL/IADL items can be ordered by the complexity of neuropsychological organisation involved with the decline in IADLs and the ambulation preceding ADLs.[15 16 37 38] In contrast, we provide additional evidence supporting a blurred hierarchical structure of functional decline when ADLs and IADLs are combined.[9 14 16 39–41] We found a disordered hierarchy among activities of moderate difficulty,[9] as well as among easy activities such as 'toileting (δ=2.06) and 'taking medications' (δ=3.06). As has been suggested,[4] the relative overlap of ADL and IADL items in aggregated scales may be reflecting different disability profiles resulting from the interaction of multiple factors, and therefore the purported strict hierarchy is only achieved in terms of general dimensions instead of specific activities or tasks. Studies with more homogeneous samples, for example, with specific chronic diseases or physical impairments, may reveal the existence of different formal hierarchies.

Reliability

Although the reliability of scores of the single metric was adequate, we found that a very low reliability of ADL and IADL scores (as separate scales) yielded important effects on the measurement of disability. For example, low reliability attenuates effect sizes and increases the chance of type II errors. As a consequence, researchers may not find the expected differences across groups or some results could be misleading. The discrepancies observed in table 6 between Cronbach's α (0.78 for ADLs and IADLs) and the Rasch reliability (PR=0.26 and 0.36, respectively) reflect the negative impact of the different factors on the classical approach to reliability. All factors are present in the ADL and IADL scales: low number of items, skewed distributions, marked floor effect, low TIF and high SEs. If the requirement measurements are violated, coefficient α yields spuriously high estimates of reliability that do not reveal the poor metric quality of the scores.[42]

The alternative non-parametric approach

While we addressed the issue of cross-cultural validity within the framework of a highly restrictive parametric model, non-parametric IRT models (eg, Mokken scaling) have been successfully applied to evaluate the measurement invariance of disability scales.[3 14 43] Non-parametric models relax some of the strong assumptions of measurement that are required for

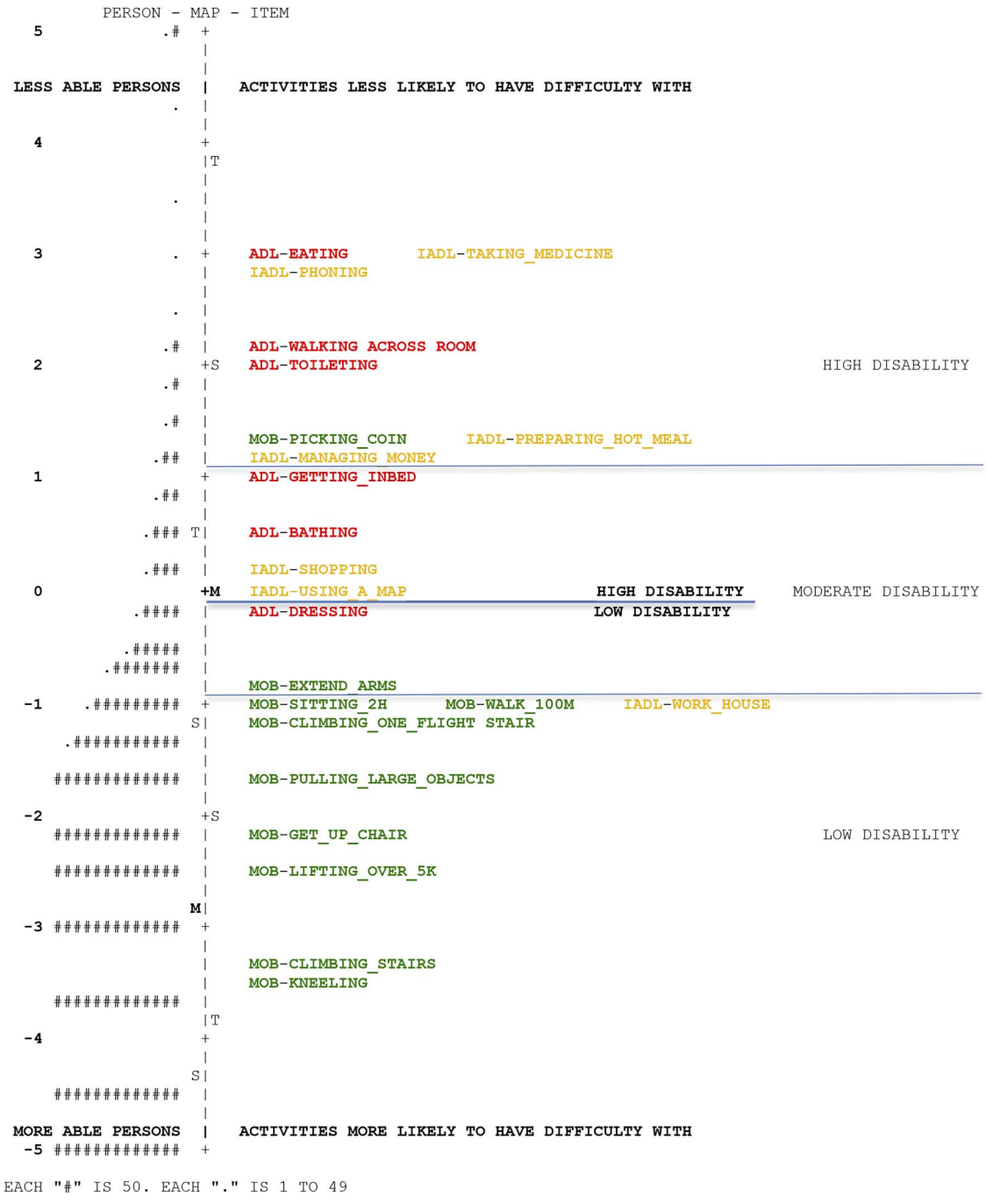

Figure 1 Hierarchical structure of the disability scale. The person–item map displays the joint locations of person disability measures (left side) and item difficulty calibrations (right side). In the left column, the more disabled participants are located near the top of the figure (positive values), and the less disabled at the bottom (negative values). In the right column, the items difficult to endorse (easiest tasks) are located near the top of the map. Continuous lines with labels represent limits for levels of disability according to the reliability indices and the test information function as are described in the next section about reliability of scores. The M and S on the vertical line between the two columns refer to mean and SD (S=1 SD, T=2 SD) statistics for persons and items measured in logit. According to the general formula, the probability of endorsing any item can be calculated by using the item difficulty (δ) and person ability estimates (θ). Thus, a respondent with the average ability of the sample (θ=−2.77, raw score=4) has a 69% probability of endorsing the item 'stooping, kneeling or crouching', whereas for the same persons the probability of endorsing the item 'preparing a hot meal' is 1%. When the ability-difficulty difference $|\theta-\delta|$ reaches 3 logits, the items are said to be 'rather off-target'.[24] ADL, activities of daily living; IADL, instrumental activities of daily living; MOB, mobility.

Table 5 Spearman's correlation coefficients of the single metric item calibrations across countries

	AT	DE	SE	NL	ES	IT	FR	DK	CH	BE	CZ	PL	HU	PT	SI	EE
AT	1.00															
DE	0.95	1.00														
SE	0.95	0.94	1.00													
NL	0.97	0.92	0.94	1.00												
ES	0.96	0.92	0.90	0.95	1.00											
IT	0.95	0.90	0.92	0.95	0.98	1.00										
FR	0.98	0.96	0.93	0.96	0.96	0.95	1.00									
DK	0.89	0.92	0.90	0.88	0.92	0.90	0.91	1.00								
CH	0.93	0.92	0.92	0.91	0.92	0.93	0.93	0.88	1.00							
BE	0.96	0.93	0.94	0.95	0.95	0.96	0.96	0.91	0.92	1.00						
CZ	0.97	0.95	0.95	0.95	0.96	0.96	0.96	0.92	0.94	0.98	1.00					
PL	0.92	0.89	0.90	0.93	0.93	0.94	0.93	0.90	0.88	0.96	0.94	1.00				
HU	0.94	0.88	0.93	0.92	0.94	0.96	0.95	0.89	0.88	0.96	0.95	0.97	1.00			
PT	0.91	0.88	0.90	0.94	0.92	0.94	0.92	0.88	0.90	0.96	0.93	0.97	0.93	1.00		
SI	0.96	0.92	0.92	0.95	0.96	0.97	0.95	0.89	0.91	0.96	0.96	0.96	0.97	0.97	1.00	
EE	0.97	0.95	0.95	0.95	0.96	0.96	0.98	0.90	0.94	0.99	0.98	0.96	0.97	0.97	0.98	1.00

p<0.001 in all cases.
AT, Austria; BE, Belgium; CH, Switzerland; CZ, Czechia; DE, Germany; DK, Denmark; EE, Estonia; ES, Spain; FR, France; HU, Hungary; IT, Italy; NL, the Netherlands; PL, Poland; PT, Portugal; SE, Sweden; SI, Slovenia.

Table 6 Reliability statistics for the single metric, and separate scales of self-care activities, instrumental activities and mobility limitations

	Single metric	ADL	IADL	Mobility
Extreme persons	**48.5%**	**91%**	**84.6%**	**49.4%**
Cronbach's α*	0.90	0.78	0.78	0.84
PR	0.74	0.26	0.36	0.61
Gp	1.70	0.59	0.74	1.24
Strata	2.66	1.12	1.32	1.98
TIF	3.52	0.93	0.99	1.89
SE	0.55	1.04	1.01	0.73

SE (measured where information is maximum).
*Cronbach's α coefficient is included only for comparative purposes that are detailed in the Discussion section.
ADL, activities of daily living, Gp, person separation; IADL, instrumental activities of daily living; PR, person (Rasch-based) reliability; TIF, test information function indicating the maximum information of the scale.

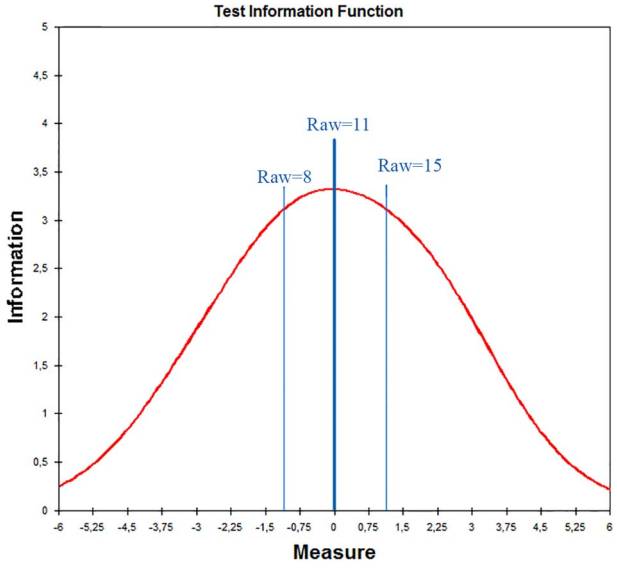

Figure 2 Test information function representing how well each (dis)ability level is being estimated with the scale. The amount of information is maximum at the person ability location of 0 logits (raw score 11), and about 3.15 for the locations of ±1.5 logits (raw score=8 and 15, respectively). (Dis)ability cannot be estimated with precision when outside of this range.

Rasch analysis. This can lead to more general conclusions and is more conservative; for example, when researchers are interested in retaining more items from a pool yielding higher reliability and better coverage of the latent trait.[44] For this reason, Mokken has been widely used for scale development and psychometric studies of scales with a small number of items. Moreover, Mokken yields ordinal-level measures that can be enough to order items, persons or both in most cases, especially when persons are performing at or near the midpoint of the range of the scale, or can also be used to determine whether change in an individual's health status has occurred. In contrast, interval-level measures allow estimates of how much more (or less) change has occurred, produce gains in precision over ordinal scores in discriminating between groups, and are ideally suited for studying longitudinal change.[18 30 36 45] The conjoint representation of persons and items on a common metric in Rasch modelling provides an easy evaluation of the reliability of scores (by means of item targeting), the construct validity (by means of the item-difficulty hierarchy) and the predictive validity (by means of the person-ability hierarchy).[18 24 30]

Table 7 Comparisons of the RP values of the two scoring methods for discriminating between groups differing in disability severity levels across medical conditions

Scoring	Diagnosed	Non-diagnosed	Mean difference	F	RP
Heart attack					
n	7888	48 595			
Summative	4.50 (0.16)	2.37 (0.05)	2.13 (0.17)	15 762	1
Interval	23.76 (0.59)	13.99 (0.19)	9.77 (0.63)	233.83	1.48
Hypertension					
n	22 532	33 951			
Summative	3.20 (0.07)	2.29 (0.05)	0.90 (0.09)	90.73	1
Interval	18.31 (0.31)	13.30 (0.23)	5.00 (0.39)	157.24	1.73
Cholesterol					
n	13 210	43 273			
Summative	3.12 (0.09)	2.47 (0.04)	0.64 (0.11)	35.59	1
Interval	17.88 (0.40)	14.31 (0.21)	3.57 (0.45)	61.07	1.71
Stroke					
n	2490	53 993			
Summative	7.53 (0.36)	2.46 (0.04)	5.06 (0.37)	183.99	1
Interval	35.01 (1.24)	14.50 (0.18)	20.51 (1.26)	261.88	1.43
Diabetes					
n	7169	49 314			
Summative	4.23 (0.14)	2.38 (0.04)	1.85 (0.15)	148.60	1
Interval	22.61 (0.55)	14.01 (0.20)	8.59 (0.59)	207.01	1.39
Lung disease					
n	3751	52 732			
Summative	4.47 (0.18)	2.50 (0.04)	1.97 (0.19)	102.31	1
Interval	24.45 (0.72)	14.51 (0.19)	9.93 (0.75)	170.66	1.67
Asthma					
n	403	56 080			
Summative	4.01 (0.25)	2.63 (0.04)	1.38 (0.25)	29.14	1
Interval	23.13 (1.09)	15.17 (0.18)	7.86 (1.10)	50.23	1.72
Arthritis					
n	13 946	42 537			
Summative	4.16 (0.09)	2.02 (0.05)	2.14 (0.11)	385.23	1
Interval	23.37 (0.36)	11.90 (0.21)	11.47 (0.44)	676.29	1.75
Osteoporosis					
n	743	55 740			
Summative	3.24 (0.18)	2.63 (0.04)	0.60 (0.19)	9.87	1
Interval	18.85 (0.76)	15.17 (0.18)	3.68 (0.79)	21.32	2.16
Cancer					
n	3036	53 447			
Summative	3.78 (0.22)	2.57 (0.04)	1.20 (0.22)	28.48	1
Interval	21.05 (0.89)	14.89 (0.19)	6.15 (0.92)	44.73	1.57
Stomach ulcer					
n	3272	53 211			
Summative	4.58 (0.24)	2.54 (0.04)	2.04 (0.25)	66.77	1
Interval	24.71 (0.95)	14.73 (0.18)	9.98 (0.97)	105.33	1.58
Parkinson					
n	409	56 074			
Summative	9.82 (0.82)	2.58 (0.04)	7.23 (0.82)	77.60	1
Interval	42.97 (2.81)	14.98) (0.18)	27.99 (2.82)	98.31	1.27
Cataracts					
n	4876	51 607			
Summative	3.76 (0.20)	2.52 (0.04)	1.23 (0.21)	33.54	1
Interval	20.31 (0.79)	14.67 (0.19)	5.64 (083)	45.24	1.35
Hip fracture					
n	1364	55 119			
Summative	6.37 (0.38)	2.54 (0.04)	3.82 (0.38)	97.91	1
Interval	30.09 (1.33)	14.83 (0.18)	15.26 (1.36)	125.87	1.28

Continued

Table 7 Continued

Scoring	Diagnosed	Non-diagnosed	Mean difference	F	RP
Alzheimer					
n	765	55 718			
Summative	10.89 (0.53)	2.49 (0.04)	8.40 (0.53)	245.65	1
Interval	45.82 (1.90)	14.66 (0.18)	31.15 (1.92)	273.47	1.11
Benign tumour					
n	221	56 282			
Summative	2.90 (0.25)	2.63 (0.04)	0.23 (0.25)	1.03	1
Interval	17.04 (1.24)	15.19 (0.18)	1.84 (1.25)	2.15	2.08

p<0.001 in all cases, except for osteoporosis (p=0.002) and benign tumour (p=NS).
Values expressed in means and SE. Wald F test was used to compare groups. Age and gender were entered as covariates.
NS, not significant; RP, relative precision.

Final recommendations

Our findings raise an important question regarding the choice of scales. That is to say, is it better to use a single metric of disability instead of separate ADL, IADL and mobility scales? If a researcher aims to estimate the prevalence of disability using the traditional cut-off score ADL 1+, IADL 1+ or mobility 1+, and wants to report findings based on descriptive and non-parametric statistics, then separate scales can be adequate. In this case, each scale could even be replaced by a single question (with a binary response format), including all the activities that the respondent might have difficulty with. Alternatively, difficulties in daily activities and functional limitations can be summed, and the aforementioned statistics can also be performed. However, researchers have to face several related issues: (1) the well-known problem of construct under-representation even with aggregated ADL/IADL scales, (2) the presence of large floor effects (around 80–90%) that seriously threaten construct validity and (3) the inability of the scales to separate statistically persons with different levels of disability, which implies that additional cut-offs are not supported empirically. Moreover, some ADLs (eg, dressing) are more challenging than some IADLs (eg, preparing a hot meal), so the inferences regarding the hierarchy of functional disability of respondents can be misleading.

Finally, we recommend that in situations where researchers are interested in (1) comparing disability severity using summative scores for parametric statistics, especially with markedly skewed distributions or expected minimal differences between groups, (2) estimating change scores in longitudinal studies, interval-level measures from the single metric should be used. In this way, researchers can be reasonably confident that any of the differences in disability detected between countries, age groups, gender, medical conditions, symptomatology and self-rated health are likely to be true differences. Furthermore, the availability of interval measures to conduct parametric statistical analysis without violating fundamental measurement requirement represents a promising field to explore the relationship between disability and a wide range of linear measures in health sciences (eg, blood pressure, cholesterol, C reactive protein, grip strength, etc).

Contributors JB conceived the study and created the data set from SHARE wave 4. JB and MC-R performed analyses and wrote the paper.

Funding The SHARE data collection has been primarily funded by the European Commission, through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812) and FP7 (SHARE-PREP: N°211909, SHARE-LEAP: N° 227822, SHARE M4: N°261982). Additional funding from the German Ministry of Education and Research, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064) and from various national funding sources is gratefully acknowledged. JB and MC-R are independent from the SHARE funding organisations.

Competing interests None declared.

Ethics approval SHARE has been approved by the Ethics Committee of the University of Mannheim and the Ethics Council of the Max-Planck-Society for the Advancement of Science.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Altman BM. Definitions, concepts, and measures of disability. *Ann Epidemiol* 2014;24:2–7.
2. Chan KS, Kasper JD, Brandt J, *et al.* Measurement equivalence in ADL and IADL difficulty across international surveys of aging: findings from the HRS, SHARE, and ELSA. *J Gerontol B Psychol Sci Soc Sci* 2012;67:121–32.
3. Fieo R, Manly JJ, Schupf N, *et al.* Functional status in the young-old: establishing a working prototype of an extended-instrumental activities of daily living scale. *J Gerontol A Biol Sci Med Sci* 2014;69:766–72.
4. Gross AL, Jones RN, Inouye SK. Development of an expanded measure of physical functioning for older persons in epidemiologic research. *Res Aging* 2014;37:671–94.
5. Haley SM, Jette AM, Coster WJ, *et al.* Late life function and disability instrument: development and evaluation of the function component. *J Gerontol A Biol Sci Med Sci* 2002;57:M217–22.
6. Laan W, Bleijenberg N, Drubbel I, *et al.* Factors associated with increasing functional decline in multimorbid independently living older people. *Maturitas* 2013;75:276–81.

7. Ramsay SE, Whincup PH, Morris RW, *et al*. Extent of social inequalities in disability in the elderly: results from a population-based study of British Men. *Ann Epidemiol* 2008;18:896–903.
8. Schoufour JD, Mitnitski A, Rockwood K, *et al*. Predicting disabilities in daily functioning in older people with intellectual disabilities using a frailty index. *Res Dev Disabil* 2014;35:2267–77.
9. Cabrero-García J, López-Pina JA. Aggregated measures of functional disability in a nationally representative sample of disabled people: analysis of dimensionality according to gender and severity of disability. *Qual Life Res* 2008;17:425–36.
10. Fieo RA, Austin EJ, Starr JM, *et al*. Calibrating ADL-IADL scales to improve measurement accuracy and to extend the disability construct into the preclinical range: a systematic review. *BMC Geriatr* 2011;11:42–56.
11. Finlayson M, Mallinson T, Barbosa VM. Activities of daily living (ADL) and instrumental activities of daily living (IADL) items were stable over time in a longitudinal study on aging. *J Clin Epidemiol* 2005;58:338–49.
12. Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. *J Gerontol B Psychol Sci Soc Sci* 2002;57:S275–84.
13. Fortinsky RH, Garcia RI, Joseph Sheehan T, *et al*. Measuring disability in Medicare home care patients: application of Rasch modelling to outcome and assessment information set. *Med Care* 2003;41:601–15.
14. Kingston A, Collerton J, Davies K, *et al*. Losing the ability in activities of daily living in the oldest old: a hierarchic disability scale from the Newcastle 85+ study. *PLoS ONE* 2012;7:e31665.
15. LaPlante MP. The classic measure of disability in activities of daily living is biased by age but an expanded IADL/ADL measure is not. *J Gerontol B Psychol Sci Soc Sci* 2010;65:720–32.
16. Spector WD, Fleishman JA. Combining activities of daily living with instrumental activities of daily living to measure functional disability. *J Gerontol B Psychol Sci Soc Sci* 1998;53:46–57.
17. Khan A, Chien CW, Bagraith KS. Parametric analyses of summative scores may lead to conflicting inferences when comparing groups: a simulation study. *J Rehabil Med* 2015;47:300–4.
18. Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Arch Phys Med Rehabil* 1989;70:857–60.
19. Seidel D, Brayne C, Jagger C. Limitations in physical functioning among older people as a predictor of subsequent disability in instrumental activities of daily living. *Age Ageing* 2011;40:463–9.
20. Börsch-Supan A, Brandt M, Hunkler C, *et al*. Data resource profile: the survey of health, ageing and retirement in Europe (SHARE). *Int J Epidemiol* 2013;42:992–1001.
21. Malter F, Börsch-Supan A. *SHARE wave 4: innovations & methodology*. Munich, Germany: Munich Center for the Economics of Aging (MEA), Max-Planck-Institute for Social Law and Social Policy, 2013.
22. Stout WF. A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika* 1990;55:293–325.
23. Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct Equ Model A Multidiscip J* 2002;9:233–55.
24. Linacre JM. *A user's guide to WINSTEPS & MINISTEPS: Rasch model computer programs*. Chicago, IL: Winsteps.com, 2011.
25. Reckase MD. Unifactor latent trait models applied to multifactor tests: results and implications. *J Educ Stat* 1979;4:207–30.
26. Zwick R, Thayer DT, Lewis C. An empirical Bayes approach to Mantel-Haenszel analysis. *J Educ Meas* 1999;36:1–28.
27. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum, 1988.
28. Hambleton RK. Good practices for identifying differential item functioning. *Med Care* 2006;44(Suppl 3):182–8.
29. Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Med Care* 2006;44(Suppl 3):152–70.
30. Bond T, Fox C. *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.
31. Montanari GE, Ranalli MG, Eusebi P. Latent variable modeling of disability in people aged 65 or more. *Stat Methods Appl* 2011;20:49–63.
32. Cieza A, Oberhauser C, Bickenbach J, *et al*. The English are healthier than the Americans: really? *Int J Epidemiol* 2014;44:229–39.
33. Chaterjii S, Byles J, Cutler D, *et al*. Health, functioning, and disability in older adults-present status and future implications. *Lancet* 2015;385:563–75.
34. Tennant A, Penta M, Tesio L, *et al*. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model. The PRO-ESOR project. *Med Care* 2004;42:37–48.
35. Altman BM, Gulley SP. Convergence and divergence: differences in disability estimates in the United States and Canada based on four health survey instruments. *Soc Sci Med* 2009;69:543–52.
36. Las Hayas C, Bilbao A, Quintana JM, *et al*. A comparison of standard scoring versus Rasch scoring of the Visual Function Index-14 in patients with cataracts. *Invest Ophthalmol Vis Sci* 2011;52:4800–7.
37. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. *Gerontologist* 1969;9:179–86.
38. Lazaridis EN, Rudberg MA, Furner SE, *et al*. Do activities of daily living have a hierarchical structure? An analysis using the longitudinal study of aging. *J Gerontol* 1994;49:47–51.
39. Verbrugge LM, Yang LS, Juarez L. Severity, timing, and structure of disability. *Soz Präventivmed* 2004;49:110–21.
40. Thomas VS, Rockwood K, McDowell I. Multidimensionality in instrumental and basic activities of daily living. *J Clin Epidemiol* 1998;51:315–21.
41. Coster WJ, Haley SM, Andres PL, *et al*. Refining the conceptual basis for rehabilitation outcome measurement. Personal care and instrumental activities domain. *Med Care* 2004;42:62–72.
42. Green SB, Yang Y. Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 2009;74:121–35.
43. Kempen GIJM, Myers AM, Powell LE. Hierarchical structure in ADL and IADL analytical assumptions and applications for clinicians and researchers. *J Clin Epidemiol* 1995;48:1299–305.
44. Sijtsma K. Methodology review: nonparametric IRT approaches to the analysis of dichotomous item scores. *Appl Psych Meas* 1998;22:3–31.
45. Norquist JM, Fitzpatrick R, Dawson J, *et al*. Comparing alternative Rasch-based methods vs raw scores in measuring change in health. *Med Care* 2004;42:25–36.