

How to assess cognitive decline when test administration changes across study waves? Harmonizing cognitive scores across waves in the China Health and Retirement Longitudinal Study

Journal of Alzheimer's
Disease Reports
Volume 8: 1661–1669
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25424823241302759
journals.sagepub.com/home/alr



Yingyan Wu¹ , Yuan S Zhang^{2,3}, Lindsay C Kobayashi⁴ , Elizabeth Rose Mayeda¹
and Alden L Gross⁵

Abstract

Background: Conducting longitudinal cognitive analyses is an essential part of understanding the underlying mechanism of Alzheimer's disease, especially for social and health behavior determinants. However, the cognitive test administration is highly likely to change across time and thus complicate the longitudinal analyses. The China Health and Retirement Longitudinal Study assessed memory through word recall tests across five study waves from 2011 to 2020. Since 2018, changes in the test stimuli and administration posed challenges for longitudinal cognitive analyses.

Objective: To address differences in administration and to preserve differences attributed to characteristics such as age and education and to derive equated scores for use in longitudinal analyses in CHARLS.

Methods: To ensure consistent underlying test ability across waves in the full sample (N = 19,364), we derived a calibration sample (N = 11,148) balancing age, gender, and education. Within this sample, we used weighted equipercentile equating to crosswalk percentile ranks between 2015 and 2018/2020 scores, then applied the algorithm to the full sample.

Results: Mean original delayed word recall was higher in 2018 (4.3 words) and 2020 (5.1 words) versus 2015 (3.2 words). Following equating, scores in 2018 and 2020 aligned better with previous waves (2015, 2018, 2020 immediate means: 4.1, 3.6, 4.0; delayed: 3.2, 2.4, 2.9 words).

Conclusions: Equipercentile equating enables the derivation of comparable scores, facilitating longitudinal analysis when cognitive test administration procedures change over time. We recommended the use of equated scores for longitudinal analyses using CHARLS cognitive data.

Keywords

Alzheimer's disease, calibration, forms and records control, longitudinal studies, neuropsychological tests, psychometrics

Received: 24 September 2024; accepted: 6 November 2024

¹Department of Epidemiology, Fielding School of Public Health, University of California Los Angeles, Los Angeles, CA, USA

²Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA

³Robert N. Butler Columbia Aging Center, Mailman School of Public Health, Columbia University, New York, NY, USA

⁴Center for Social Epidemiology and Population Health, Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA

⁵Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

Corresponding authors:

Alden L Gross, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 2024 E. Monument Street, Baltimore, MD, 21205, USA.
Email: agross14@jhu.edu

Yingyan Wu, Department of Epidemiology, UCLA Fielding School of Public Health, 650 Charles E Young Dr., Los Angeles, CA 90095, USA.
Email: yingyanwu@ucla.edu



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (<https://uk.sagepub.com/en-gb/eur/open-access-at-sage>).

Introduction

Cognitive decline and dementia are major public health concerns worldwide¹; however, most studies are conducted in high-income countries. It is essential to assess cognitive decline and dementia in low-income and middle-income countries to expand the current evidence base and further our understanding of the determinants of cognitive decline and dementia.² Since 2011, the China Health and Retirement Longitudinal Study (CHARLS) has collected data on individual and household characteristics and various aspects of aging-related health measures from a nationally representative sample of middle-aged and older adults in China every two or three years.³ As of early 2024, CHARLS has released five waves of data between 2011 and 2020 and has become a leading data resource for characterizing cognitive impairment and cognitive aging among middle-aged and older adults in China.

One critical advantage of CHARLS is its longitudinally repeated assessments of cognitive function, which is essential for examining changes in cognitive function with age. However, changes to the content and administration of cognitive tests over time are commonplace in research studies, due to changes in study design, priorities, or logistics. Alterations to cognitive test items over time can complicate analyses of cognitive change by mixing true changes in cognitive function with changes in test scores due to test administration differences. In 2018, CHARLS integrated the Harmonized Cognitive Assessment Protocol (HCAP) into its main household survey. The HCAP is an international research collaboration that aims to measure and understand the risk, prevalence, and consequences of dementia within ongoing longitudinal studies of aging worldwide using comparable data.⁴ The integration of the HCAP into the main CHARLS household survey in 2018 significantly expanded the battery of cognitive tests, but also coincided with modifications to existing cognitive tests that had been administered in prior waves.

Because of modifications to the test administration and word lists introduced in 2018, longitudinal differences in test performance do not necessarily reflect intra-individual changes in cognitive function but could be due instead to differences in perceived test difficulty or other test administration characteristics. We aimed to document these modifications made to the CHARLS word recall test and to apply a test-equating approach to adjust for the test item non-equivalence that stems from different test administrations to generate comparable scores that account for the differences in test content and administration between tests administered in 2011–2015 versus 2018–2020 to facilitate longitudinal analysis.

Methods

Data and participants

Data were from CHARLS, an ongoing nationally representative longitudinal survey of persons in China 45 years of age or older and their spouses regardless of age.³ The national baseline was conducted between June 2011 and March 2012 and included 17,708 respondents. CHARLS respondents are followed approximately every 2–3 years. Respondents have been followed in 2013, 2015, 2018, and 2020. The CHARLS-HCAP was introduced as part of the main household survey in 2018 and was also administered in 2020. Further details of the CHARLS-HCAP can be found elsewhere.⁵

Immediate and delayed recall tests: before 2018 and 2018 onward

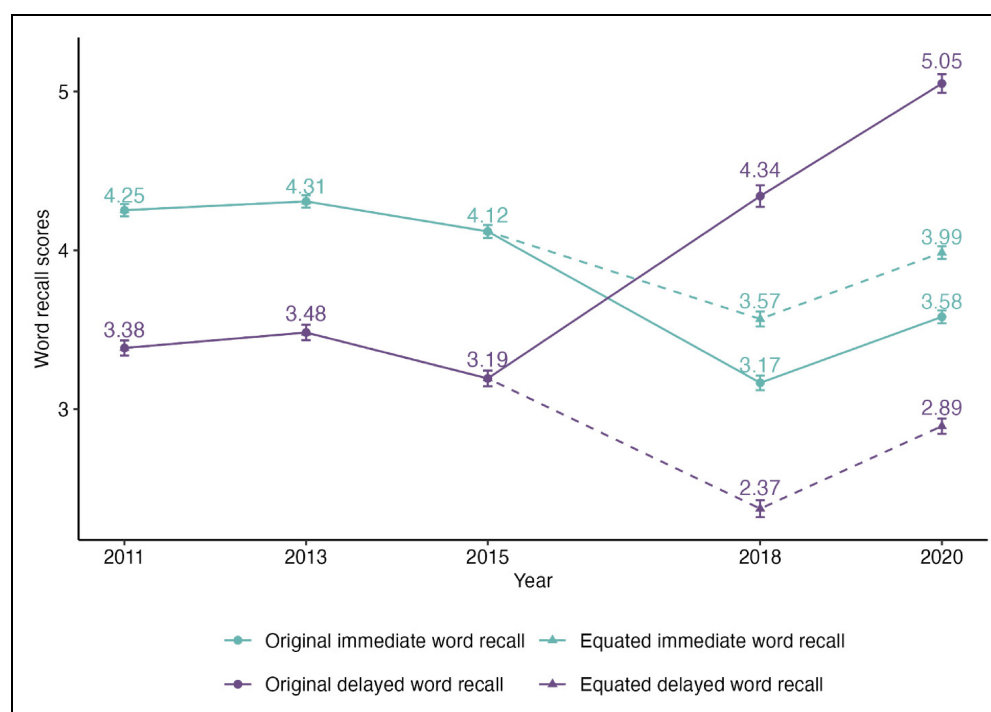
At each wave since its baseline in 2011, CHARLS has administered immediate and delayed word recall tests, a widely used measure of episodic memory. Respondents are asked to immediately recall ten nouns that are read aloud (immediate recall) and then recall the same words after a five-minute delay (delayed recall).⁶

With the adoption of the HCAP version of the recall test from 2018 onward were two major changes in the CHARLS word recall test. First, the specific nouns used in the test were changed. Earlier survey waves included commonly used words in Mandarin, as spoken by Chinese individuals. The ten nouns used in the CHARLS HCAP tests starting from 2018 were translated directly from a Consortium to Establish a Registry for Alzheimer's Disease (CERAD) word list⁷ used in the U.S. Health and Retirement Study (HRS) HCAP that was distinct from those used in the 2011–2015 waves (Table 1). Some words used in the 2018 CHARLS HCAP were less familiar to many older Chinese people, such as butter and shore. Neuropsychological research has suggested that low-frequency words often yield poorer recall performance compared to high-frequency words.^{8,9} A second change to the CHARLS word recall test starting in the 2018 wave was to administer three trials of immediate word recall tests instead of one trial. This change led to higher delayed recall scores in the 2018 and 2020 waves relative to prior survey waves (solid lines in Figure 1). Indeed, delayed recall scores greatly exceeded recall on the first trial for most respondents in 2018 and 2020 because participants were given three times as much exposure to the word list. Additionally, starting in the 2018 wave, the word lists were read and presented visually to the participants, whereas the words were only read to the participants in the 2011–2015 waves. Moreover, in the 2011–2015 waves, if a person did not remember any words in the immediate word recall test,

Table 1. Word list used in CHARLS word recall test.

Waves 1–3 (2011–2015)				Waves 4–5 (2018, 2020)
List a	List b	List c	List d	
rice 米饭	stool 凳子	mountain 山	water 水	butter 奶油
river 河流	foot 脚	stone 石头	hospital 医院	arm 胳膊
doctor 医生	sky 天空	blood 血液	tree 树木	shore 海滨
clothes 衣服	money 金钱	mother 妈妈	father 爸爸	letter 信件
egg 鸡蛋	pillow 枕头	shoes 鞋子	fire 火	queen 王后
cat 小猫	dog 小狗	eye 眼睛	tooth 牙齿	cabin 木屋
bowl 饭碗	house 房子	girl 女孩	moon 月亮	pole 扁担
child 小孩	wood 木头	house 房子	village 村子	ticket 票
hand 手	school 小学	road 马路	boy 男孩	grass 草
book 书	tea 茶	sun 太阳	table 桌子	engine 发动机

CHARLS: China Health and Retirement Longitudinal Study.

**Figure 1.** Empirical plot for original (solid lines) and equipercentile equated (dashed lines) immediate word recall scores and delayed word recall scores in a balanced sample – attended all 5 waves of immediate word recall tests ($n = 6830$) and delayed word recall tests ($n = 6324$). Mean $\pm 2 \times$ SE was shown for each wave.

the interviewer would reread the word list to the respondent, whereas this was not done in immediate word recall trials in 2018 and 2020. In the analysis, we used the first trial of immediate word recall tests from waves 2018 and 2020.

Statistical analysis

We calculated and compared the mean scores for immediate and delayed recall among participants in each study wave from 2011 to 2020. To address form differences introduced in 2018, we adapted a weighted version of equipercentile to

derive comparable versions of immediate and delayed word recall test scores for the 2018 and 2020 survey waves relative to the 2015 wave. To do this, we adopted a two-stage approach.^{10–12}

Given the consistent administration of word recall tests from 2011 to 2015, we selected wave 2015 as the reference wave for equating. We used the first trial of waves 2018 and 2020 as the focal scores for equating. In the first stage, we selected an equating sample to derive an equating algorithm based on characteristics of test score distributions. The goal of this stage was to define a calibration sample

of participants in CHARLS 2015, CHARLS 2018, and CHARLS 2020, such that any differences in memory scores between the waves are likely attributable to form differences in the tests and not attributable to demographic differences or real cognitive changes. This goal was achieved by restricting the calibration sample to a group with a common age range and similar education level across CHARLS waves in 2015, 2018, and 2020 and using weighting to further balance these characteristics across the waves. Specifically, the calibration sample was restricted to participants aged between 45 and 90 years who completed word recall tests in either 2015, 2018, or 2020 because this age range is present in each wave. Participants in the highest (post-graduate degree, e.g., Master, PhD) or lowest (less than elementary school) education category were excluded due to low numbers and to ensure a more comparable relationship between memory performance and education level. These restrictions helped to preserve differences between people in memory performance attributable to aging, and education level.

Next, we developed weights to ensure a consistent underlying test ability for completing each of the immediate and delayed word recall tasks in 2015 and in a combined group of 2018 and 2020 (hereafter, 2018/2020). Specifically, we fit a logistic regression model to obtain the probability (propensity score) of completing the test in 2018 or 2020 versus 2015. Time-varying measures include wave-specific age (years), rural residence (yes/no), marital status (yes/no), and working status (agricultural work/non-agriculture employed/unemployed/retired/never work). Time invariant measures include gender (women/men) and education level (elementary school or less/middle school level/college). Rural residence was defined using information recorded at the community level mapping to the rural or urban region definition by the National Bureau of Statistics of China.¹³ The rest of the covariate measures for modeling the propensity scores were self-reported. We then calculated stabilized inverse odds of weights (sIOW) for each respondent in the calibration sample who completed the test in CHARLS 2018 or 2020 by multiplying the inverse of the odds of completing the 2018 or 2020 recall test given the covariates, and the unconditional odds of completing the 2018 or 2020 testing. After developing sIOW based on the strongest determinants of cognitive test performance (age and education), we examined the covariate balance between the weighted CHARLS 2018 or 2020 participants and the CHARLS 2015 participants in the calibration sample, calculated as the standardized mean difference ($\text{mean}[\text{CHARLS 2018 or 2020}] - \text{mean}[\text{CHARLS 2015}] / \text{SD}[\text{CHARLS 2015}]$) for each variable. We iteratively adjusted covariates and interaction terms included in the propensity score models until the covariate balance was maximized based on available data. Larger standardized mean differences indicated stronger sources of underlying differences. Differences < 0.2 or < 0.25 are often deemed to indicate balance (i.e., small differences).¹⁴ We choose the most

parsimonious model for weight development based on covariate balance. The model included age, sex, education level, working status, rural residence, and interaction terms between rural residence with sex and working status.

In the second stage, we derived an equipercentile equating algorithm in the weighted calibration sample and applied it to the full CHARLS 2018 and 2020 samples to obtain the 2015-equivalent immediate and delayed word recall test scores. The equipercentile equating method accounts for the nonlinear relationship between the cumulative distribution function of the scores in 2018 and 2020, and those in 2015 because it leverages percentile ranks of scores. Given that the underlying abilities of the individuals in the calibration sample who completed the recall tests in both 2015 and 2018/2020 are assumed to be equal (due to restriction to common age and education ranges), the weighted percentile rankings of the recall scores obtained from the 2015 and 2018/2020 participants were set to be comparable. We then calculated the equipercentile equivalent of the recall scores for the 2018/2020 respondents on the 2015 survey scale. This calculation was achieved by finding the percentile rank of a particular score in the lumped 2018/2020 sample and finding the CHARLS 2015 score associated with that percentile rank in the CHARLS 2015 in the calibration sample. This crosswalk was then applied to the full study sample. We examined the equated scores and the original scores for three immediate word recall test trials and the delayed word recall test in waves 2018 and 2020 in the full study sample. Data construction and statistical analysis were conducted using R 4.3.1. The equipercentile equating analyses were run using the R package *equate*.¹⁵ The code is available on GitHub (https://github.com/Yingyan-Wu/charls_memory_equating).

Results

Descriptive statistics

In the full sample, among those $N = 14,440$ aged 60+ years in 2018, $N = 7676$ had complete data on trial 1 immediate recall and delayed recall, of whom 52% had delayed recall scores greater than the first of three trials of immediate recall scores. Similarly, in 2020, the proportion of participants having higher delayed recall scores than the first of three trials of immediate recall scores was 65%. For the 2018 and 2020 waves where CHARLS implemented three immediate word recall trials, we examined scores of the three immediate word recall trials in the full sample; the median of the immediate word recall trials increased with each trial, and the delayed word recall test scores were always lower than the third trial of immediate word recall test scores (Figure 2).

In a balanced sample among which the participants completed all 5 waves of immediate word recall tests or delayed

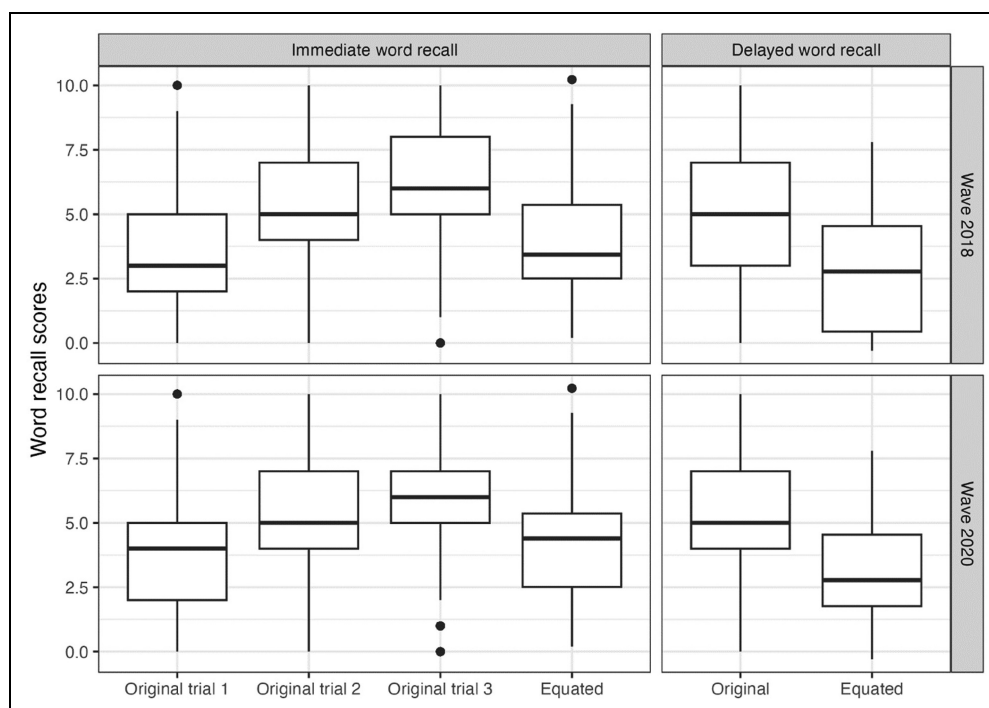


Figure 2. Distribution of original and equated immediate and delayed word recall scores in the full sample. In waves 2018 and 2020, immediate word recall tests were given 3 times to the participants before the delayed word recall test.

Table 2. Characteristics of restricted calibration sample in CHARLS (before applying sIOW).

Characteristic	2015, N = 9464 ¹	2018, N = 8202 ¹	2020, N = 8297 ¹
Age at survey			
Mean (SD)	58.2 (8.8)	60.2 (8.7)	61.6 (8.7)
Female (%)	3772 (39.9)	3403 (41.5)	3493 (42.1)
Early life education level (simplified) (%)			
Elementary school	4002 (42.3)	3406 (41.5)	3910 (47.1)
Vocational school/middle school/high school	5148 (54.4)	4541 (55.4)	4148 (50.0)
College	314 (3.3)	255 (3.1)	239 (2.9)
Rural residence (%)	5037 (53.2)	4435 (54.1)	4847 (58.4)
Work status (%)			
Agriculture employment	2848 (30.1)	2520 (30.7)	3822 (46.1)
Non-agriculture employment	3539 (37.4)	2874 (35.0)	1790 (21.6)
Unemployed	52 (0.5)	36 (0.4)	607 (7.3)
Retired	2919 (30.8)	2704 (33.0)	2034 (24.5)
Never work	106 (1.1)	68 (0.8)	44 (0.5)
Hukou status (%)			
Agricultural hukou	6527 (69.0)	5634 (68.7)	5514 (66.5)
Non-agricultural hukou	2721 (28.8)	2299 (28.0)	1726 (20.8)
Unified residence hukou	205 (2.2)	260 (3.2)	1054 (12.7)
Do not have hukou	11 (0.1)	9 (0.1)	3 (0.0)
Immediate word recall			
Mean (SD)	4.4 (1.7)	3.8 (1.8)	4.0 (1.6)
Delayed word recall			
Mean (SD)	3.4 (2.0)	5.3 (2.4)	5.6 (2.2)

¹Mean (SD) for continuous variables; n (%) for categorical variables

CHARLS: China Health and Retirement Longitudinal Study; sIOW: stabilized inverse odds weights; SD: standard deviation

word recall tests, mean scores with standard errors for both immediate ($N = 6830$) and delayed ($N = 6324$) recall tests across the four CHARLS waves from 2011 to 2020 are shown in Figure 1. From 2011–2015, mean immediate recall was consistently lower than that of delayed recall. However, the patterns changed starting in 2018 with the introduction of the HCAP: there was a marked drop in the mean score of the immediate recall test in 2018 compared to previous waves, coinciding with comparatively higher delayed recall scores (solid lines in Figure 1).

Calibration sample

Restricting the sample to exclude those with less than elementary school education or those who have a postgraduate degree (master's or PhD), resulted in analytical samples of $n = 9464$ in 2015, $n = 8202$ in 2018, and $n = 8297$ in 2020 (unique participants: $n = 11,148$). The characteristics of the calibration sample by study wave (2015, 2018, and 2020) are shown in Table 2. After applying weights, covariate balance was achieved with <0.1 SD units for all the

variables including the variables not included in the weight-generating models: standardized mean differences were close to 0 (Figure 3).

Crosswalk of equated scores and original scores

The crosswalk of equated scores and original scores for survey waves 2018 and 2020 are in Figure 4 and Table 3. We did not scale the scores back to a 0–10 scale to preserve the discrimination ability of the equated score distribution. As is shown in Figure 1, in the balanced sample where participants completed all 5 waves of immediate word recall tests or delayed word recall tests, mean equipercentile equated immediate word recall scores were higher than original scores in both 2018 and 2020 (2018 equated versus original score 3.57 versus 3.17; 2020 equated versus original score 3.99 versus 3.58). Mean equated delayed word recall scores were lower than the original scores in both 2018 and 2020 (2018 equated versus original score: 2.37 versus 4.34; 2020 equated versus original score 2.89 versus 5.05). Using equated scores, a general decrease in

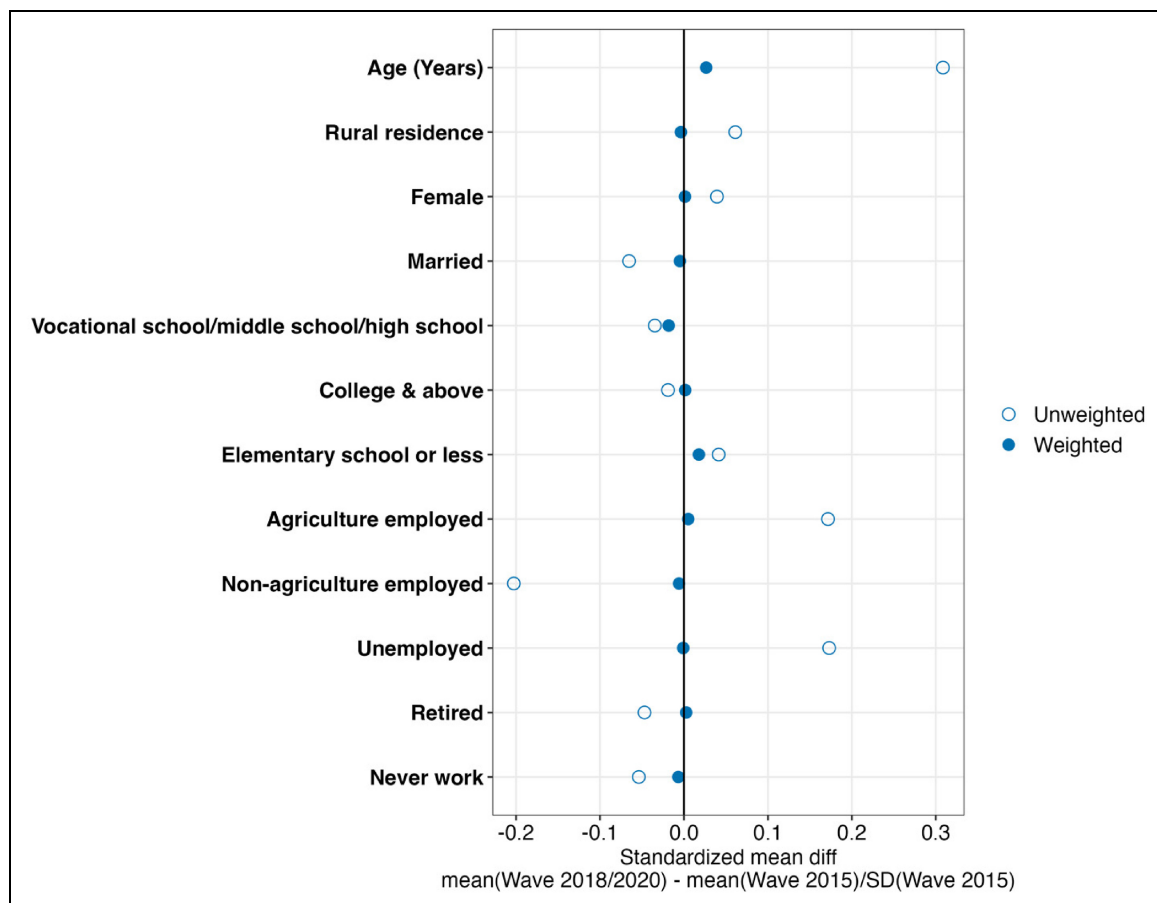


Figure 3. Covariate balance plot showing the standardized mean difference of important characteristics of testing ability between wave 2015 and wave 2018, and 2015 and 2020 before and after applying IOW to the calibration sample. Circles represent unweighted standardized mean difference and dots represent weighted standardized mean difference.

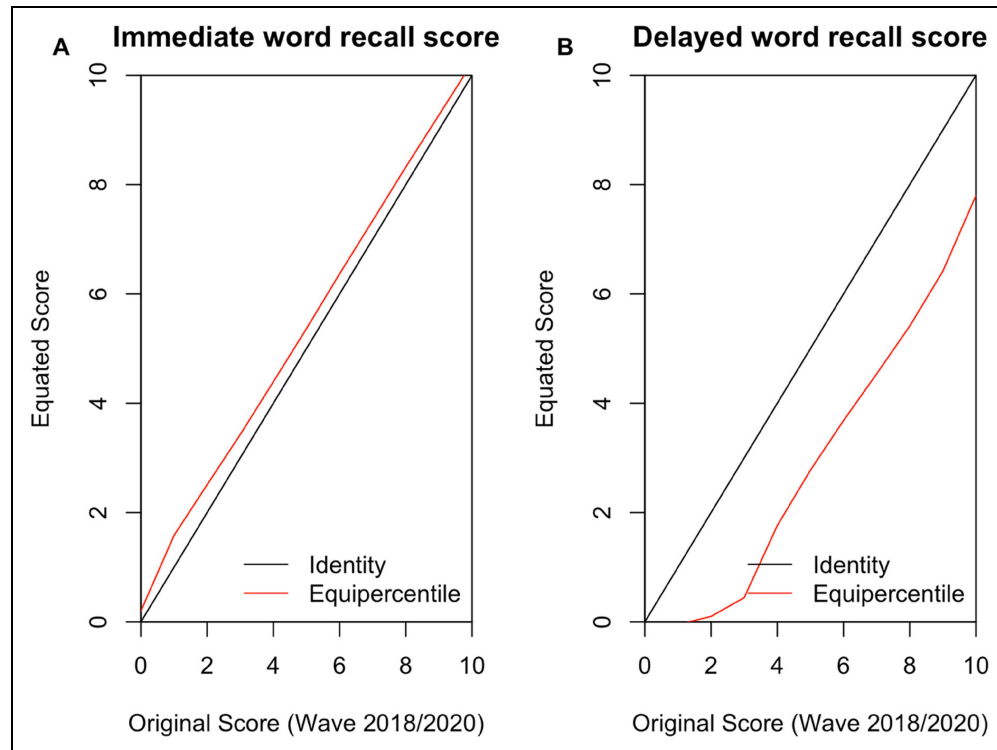


Figure 4. The relationship between equated and original immediate and delayed word recall scores from CHARLS waves 2018 and 2020 in the weighted calibration sample.

Table 3. The crosswalk between original immediate and delayed word recall scores and equated scores in waves 2018 and 2020.

Score	Original score	Wave 2018/2020 Equated score
Immediate	0	0.2
	1	1.58
	2	2.51
	3	3.43
	4	4.39
	5	5.36
	6	6.37
	7	7.34
	8	8.32
	9	9.27
	10	10.22
Delayed	0	−0.3
	1	−0.05
	2	0.1
	3	0.44
	4	1.76
	5	2.78
	6	3.68
	7	4.54
	8	5.41
	9	6.42
	10	7.8

mean recall was observed for both equated immediate and delayed word recall scores over time. In the full sample, equated immediate word recall scores were generally

higher than the equated delayed word recall scores in waves 2018 and 2020 (Figure 2).

Discussion

In this study, we described form differences in test items and test administration that led to spurious longitudinal differences in performance on immediate and delayed word recall tests across waves of CHARLS. We then applied equating methods to remove these form differences. Prior to applying equating methods, immediate word recall scores substantially decreased while delayed word recall scores drastically increased in 2018, compared with previous waves. Alterations to the content and administration of cognitive test items over time in a study are common but can obfuscate longitudinal analysis of change. This study demonstrates the application of equating methods when non-equivalent tests are administered in longitudinal surveys. By applying a test-equating method, we created CHARLS 2015-equivalent word recall scores for the CHARLS 2018 and 2020 wave respondents that could be used to leverage longitudinal CHARLS data to examine changes in cognitive function, particularly memory.

An advantage of using the 2015 wave as the reference wave against which to conduct equipercntile equating is that it retains practice effects observed across CHARLS study waves 2011, 2013, 2015, 2018, and 2020. Retest

effects have been shown to be present for up to 17 years between measurement occasions.¹⁶ The retest effects were probably present in CHARLS due to contextual contributions to practice,¹⁷ despite the use of alternate word lists in CHARLS across the 2011–2015 waves.¹⁰ However, the most pronounced evidence of retest typically occurs between the first and second occasions of measurement.^{17–19} Thus, we believe that changes from 2011 to 2013 may have been the most likely period during which to observe a retest effect. Ensuring the retention of practice effects is crucial for reliable longitudinal analysis for cognitive decline.

It is important to study cognitive decline and dementia using a nationally representative cohort in low-income and middle-income countries, e.g., CHARLS data in China. Most research to date on these topics has been conducted in high-income countries, a smaller portion of the world population, making the transportability of the current evidence uncertain to low-income and middle-income countries, a larger portion of the world population. It is also important to assess research questions in a low-income and middle-income context to further our understanding of determinants of cognitive decline and dementia.

Alterations in tests administered across time are common, particularly in longitudinal studies where changes in test format, difficulty, or administration procedures may occur due to various reasons such as updates in assessment protocols. These changes can pose challenges when attempting to compare repeated measures over time and may complicate longitudinal analyses aimed at tracking changes in cognitive function within individuals. Our method, which involves utilizing the weighted equipercentile equating method offers a solution to this issue for CHARLS and other studies. We presented a valuable tool for researchers seeking to conduct comprehensive longitudinal analyses for cognitive change with comparable cognitive scores over time.

Measurement adjustment techniques have been employed to account for shifts in cognitive assessments over time, such as transitions from in-person to telephone administration across study waves²⁰ and to evaluate the reliability of cognitive tests that measure the same psychological traits across time²¹ using methods including equipercentile equating.

Equipercentile-equated scores are unnecessary in cross-sectional research unless different forms were administered to different subgroups of participants in a non-random fashion at the same time. However, cautions merit attention for making comparisons to other studies when the tests were administered differently (e.g., varying scales¹² or item sets¹¹). Most of the major questions in modern cognitive aging research are longitudinal in nature and require consideration of change in cognitive performance because cognitive change is the fundamental feature of cognitive disorders in aging.²² Highlighting the importance of longitudinal data, most impactful trials and cohorts in cognitive aging research take repeated measures. Particularly for

investigators using data they did not personally collect, it is crucial to read study documentation in detail to understand the measures used to assess cognitive performance, including how tests were administered, scored, and coded. Form differences in word lists are a subtle, pernicious feature of many longitudinal studies that arise when not enough attention has been paid to making word lists equivalent.²³ Equating methods have a rich history of being used to resolve problems that arise in longitudinal analyses from form differences.^{10,11}

Several caveats merit attention. The equipercentile equating method relies on the assumption that two test-taking populations have the same underlying ability to be validly equated. In our case, we did not have the same study sample or two randomly assigned groups who were administered two versions of tests simultaneously. To address this challenge, prior to applying the equating algorithm, we restricted the sample and applied IOW in a calibration sample in which the underlying abilities of the respondents who took the two tests in 2015 and 2018, and 2015 and 2020, could be assumed to be identical. Our ability to identify a group of respondents who completed the two different tests and had the same underlying memory function depended on the availability of covariates that are predictive of memory and were limited to age and education. Another limitation, specific to the equipercentile equating method, is that the outcome should be continuously distributed and have enough range to reliably distinguish different percentiles. Given that the word recall scores only have 11 categories, ranging from 0 to 10, the equipercentile-equated scores we used are not integers.

In conclusion, our work documented our effort to address the challenges posed by non-equivalent cognitive tests with different administrations over time. We recommend using the equated word recall scores for longitudinal analysis of memory change utilizing word recall tests in CHARLS. Specifically, for the immediate word recall test, we advise using the first trial scores updated with the equated word recall scores from waves 2018 and 2020 in longitudinal analyses of immediate word recall performance. We have made the crosswalk of original and equated word recall scores available, and we have also distributed code to reproduce the equipercentile crosswalk. Harmonization methods such as equipercentile equating methods can be useful to generate comparable word recall test scores to facilitate longitudinal analyses for cognitive change.

Acknowledgments

We would like to express our gratitude to the China Health and Retirement Longitudinal Study (CHARLS) participants for providing their consent to the data. We also appreciate the efforts of the CHARLS team and Gateway to Global Aging Data team for making the original and harmonized nationally representative data available.

ORCID iDs

Yingyan Wu  <https://orcid.org/0000-0003-1875-7493>

Lindsay C Kobayashi  <https://orcid.org/0000-0003-2725-3107>

Elizabeth Rose Mayeda  <https://orcid.org/0000-0002-6326-0163>

Alden L Gross  <https://orcid.org/0000-0002-4678-0562>

Statements of declaration

Author contributions

Yingyan Wu (Conceptualization; Data curation; Formal analysis; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing); Yuan S Zhang (Conceptualization; Funding acquisition; Methodology; Writing – review & editing); Lindsay C Kobayashi (Funding acquisition; Methodology; Writing – review & editing); Elizabeth Rose Mayeda (Conceptualization; Funding acquisition; Methodology; Writing – review & editing); Alden L Gross (Conceptualization; Funding acquisition; Methodology; Supervision; Writing – review & editing).

Funding

This work was supported by the National Institute on Aging [grant numbers R01AG070953 to L.C.K and A.L.G, R01AG063969 and R01AG052132 to E.R.M., and R00AG070274 to Y.S.Z.].

Declaration of conflicting interests

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

Data availability

The China Health and Retirement Longitudinal Study (CHARLS) data are publicly available at: <https://charls.pku.edu.cn/en/>

References

1. Plassman BL, Langa KM, Fisher GG, et al. Prevalence of cognitive impairment without dementia in the United States. *Ann Intern Med* 2008; 148: 427–434.
2. Mayeda ER and Wu Y. Identifying modifiable determinants of cognitive decline and dementia risk. *Neurology* 2024; 102: e209293.
3. Zhao Y, Hu Y, Smith JP, et al. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *Int J Epidemiol* 2014; 43: 61–68.
4. Langa KM, Ryan LH, McCammon RJ, et al. The Health and Retirement Study Harmonized Cognitive Assessment Protocol Project: study design and methods. *Neuroepidemiology* 2020; 54: 64–74.
5. Meng Q, Wang H, Strauss J, et al. Validation of neuropsychological tests for the China Health and Retirement Longitudinal Study Harmonized Cognitive Assessment Protocol. *Int Psychogeriatr* 2019; 31: 1709–1719.
6. Zhao Y, Strauss J, Chen X, et al. *China health and retirement longitudinal study wave 4 user's guide*. Peking: National School of Development, Peking University, 2020.
7. Fillenbaum GG, van Belle G, Morris JC, et al. Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers Dement* 2008; 4: 96–109.
8. Ryder RJ and Slater WH. The relationship between word frequency and word knowledge. *J Educ Res* 1988; 81: 312–317.
9. Balota DA and Spieler DH. Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. *J Exp Psychol Gen* 1999; 128: 32–55.
10. Gross AL, Inouye SK, Rebok GW, et al. Parallel but not equivalent: challenges and solutions for repeated assessment of cognition over time. *J Clin Exp Neuropsychol* 2012; 34: 758–772.
11. Gross AL, Kueider-Paisley AM, Sullivan C, et al. Comparison of approaches for equating different versions of the Mini-Mental State Examination administered in 22 studies. *Am J Epidemiol* 2019; 188: 2202–2212.
12. Briceño EM, Gross AL, Giordani BJ, et al. Pre-statistical considerations for harmonization of cognitive instruments: harmonization of ARIC, CARDIA, CHS, FHS, MESA, and NOMAS. *J Alzheimers Dis* 2021; 83: 1803–1813.
13. Phillips D, Green H, Petrosyan S, et al. *Harmonized CHARLS documentation version D (2011–2018)*. Los Angeles, CA, USA: USC Center for Economic and Social Research, Program on Global Aging, Health, and Policy, June 2021.
14. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci Rev J Inst Math Stat* 2010; 25: 1–21.
15. Albano AD. Equate: an R package for observed-score linking and equating. *J Stat Softw* 2016; 74: 1–36.
16. Salthouse TA, Schroeder DH and Ferrer E. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Dev Psychol* 2004; 40: 813–822.
17. Gross AL, Chu N, Anderson L, et al. Do people with Alzheimer's disease improve with repeated testing? Unpacking the role of content and context in retest effects. *Age Ageing* 2018; 47: 866–871.
18. Vivot A, Power MC, Glymour MM, et al. Jump, hop, or skip: modeling practice effects in studies of determinants of cognitive change in older adults. *Am J Epidemiol* 2016; 183: 302–314.
19. Gross AL, Benitez A, Shih R, et al. Predictors of retest effects in a longitudinal study of cognitive aging in a diverse community-based sample. *J Int Neuropsychol Soc* 2015; 21: 06.
20. Smith JR, Gibbons LE, Crane PK, et al. Shifting of cognitive assessments between face-to-face and telephone administration: measurement considerations. *J Gerontol Ser B* 2023; 78: 191–200.
21. Bertola L, Benseñor IM, Gross AL, et al. Longitudinal measurement invariance of neuropsychological tests in a diverse sample from the ELSA-Brasil study. *Braz J Psychiatry* 2020; 43: 254–261.
22. Salthouse T. *Major issues in cognitive aging*. New York: Oxford University Press, 2010.
23. Brandt J. The Hopkins Verbal Learning Test: development of a new memory test with six equivalent forms. *Clin Neuropsychol* 1991; 5: 125–142.