Social Science & Medicine 383 (2025) 118473

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Education shapes episodic memory measurement via test specifications: Evidence from the China Health and Retirement Longitudinal Study

Yizhou Chen *, James O'Donnell

School of Demography, Australian National University, Australia

ARTICLE INFO

Handling Editor: Susan J. Elliott

Keywords:
Education
Cognitive decline trajectory
Episodic memory measurement
Word recall test
Test specifications

ABSTRACT

Background: Many studies have examined how education influences cognitive decline trajectories, often reflected through episodic memory deficits measured by word recall tests. However, little is known about how education affects episodic memory measurement in longitudinal studies where word-list complexity and test form vary. Our study aims to explore whether education influences episodic memory measurement via test specifications in a Chinese context.

Methods: 23,951 respondents aged over 45 (78,364 person-years) from five waves (2011–2020) of the China Health and Retirement Longitudinal Study (CHARLS) were included. We fitted two random-intercept models—one for immediate and one for delayed test scores—to examine how education influences episodic memory measurement across varying test formats and word list complexities. Based on these results, we applied a hybrid frequency-estimation equating approach to facilitate longitudinal studies in CHARLS, accounting for education's impact when word recall tests use varying specifications.

Results: Respondents with higher education scored better on immediate and delayed word recall tests, but all education groups were negatively affected by increased word-list complexity, with lower-educated individuals more vulnerable. Higher-educated respondents also gained more improvement in word recall outcome from extra practice trials when complexity remained constant. After equating, the predicted trajectories reflected more accurate cognitive decline over time, enhancing the measurement's validity.

Conclusion: These findings suggest that education strongly influences episodic memory assessment, as test specifications—word-list complexity and test form—interact with participants' education and shape performance gaps between higher- and lower-educated groups. For equating techniques in longitudinal studies, frequency estimation suits waves with similar complexity, whereas equipercentile equating better addresses substantial complexity differences, thereby enhancing measurement validity.

1. Introduction

Cognitive aging poses a public health concern as populations worldwide experience rapid aging (WHO, 2021). Individuals commonly undergo memory-function declines through normal aging, including those caused by neurodegenerative diseases (Livingston et al., 2024). A growing body of longitudinal research employing multiple survey waves explores how various factors shape trajectories of cognitive decline from a life-course perspective, often using free word recall tests as a core measure of memory performance, in which participants recall a list immediately after presentation and again following a brief delay (Ford et al., 2021; Li et al., 2022).

A challenge though for longitudinal research is that recall tests can change across waves, both in terms of test forms and the complexity of word lists. This occurred in the most important longitudinal study of health and ageing in China, the China Health and Retirement Longitudinal Study (CHARLS), making it difficult to study and understand trajectories of cognitive health and decline in one of the largest and most rapidly aging countries in the world (UN, 2024). While prior approaches have been developed to address inconsistent estimates of cognitive ability under different test specifications (Guo et al., 2025), few longitudinal studies have systematically investigated how variations in test formats and word-list complexity interact with participants' education levels to influence recall scores and affect the reliability of estimated

This article is part of a special issue entitled: Cognitive Aging published in Social Science & Medicine.

* Corresponding author. Australian Capital Territory, 2601, Australia.
E-mail addresses: yizhou.chen@anu.edu.au (Y. Chen), james.odonnell@anu.edu.au (J. O'Donnell).

https://doi.org/10.1016/j.socscimed.2025.118473
Received 28 February 2025; Received in revised form 30 June 2025; Accepted 31 July 2025
Available online 5 August 2025

educational differences in memory trajectories over time. Education has been shown to correlate with higher cognitive ability and is even used as a "convenience proxy" for cognitive reserve (Stern et al., 2020), meaning that higher educated people potentially have greater capacity to respond to more complex word lists and tighter test protocols. Accordingly, this study examines the role of education in measuring free recall under diverse test specifications and proposes an equating method to align word recall outcomes across differing designs for longitudinal analysis.

2. Background

2.1. Episodic memory and its measurement

Cognitive decline, often reflected in inefficient functioning of memory, attention, and reasoning, is most prominently observed through a deficit in episodic memory (Dickerson and Eichenbaum, 2010). Such cognitive change is primarily attributed to normal aging among older adults, yet it is also caused by neurodegenerative diseases, including dementia, with Alzheimer's disease constituting the leading cause (Livingston et al., 2024). Episodic memory is a long-term declarative memory system characterized by retrieving previously stored information. Distinct from semantic memory, another component of long-term memory related to encyclopedic knowledge, episodic memory—often described as "mental time travel"—refers to the cognitive ability to encode, store, and retrieve information about personal experiences (Tulving, 2002).

Behavioral and neuroimaging approaches represent two primary methods for assessing how normal and pathological aging influence episodic memory deficits. The behavioral approach, the focus of this study, involves episodic memory tasks measuring subjects' performance in encoding, storage, and retrieval. Verbal memory tasks, also called word recall tests, are widely adopted in survey and clinical studies, in which subjects are given and asked to recall a list of words. Word recall typically declines gradually through normal aging and abruptly in early pathological aging and cognitive decline (Tromp et al., 2015). Two main types of word recall tests are used in specific scenarios, cued and free word recall. In cued recall tests, trained neuropsychologists provide visual cues to aid subject recall (Grober et al., 2010). In large-scale, less controlled studies, free word recall tests like the Consortium to Establish a Registry for Alzheimer's Disease (CERAD) are commonly used (Fillenbaum et al., 2008). These typically involve immediate and delayed recall tests, the former being where subjects are asked to recall a list of words immediately after hearing them and the latter being a follow-up test where subjects are asked to recall the same word list after a short period of time.

The consistency of word recall designs is a major issue in studying cognitive decline and dementia over time and through longitudinal analyses. One key issue is whether word lists change across waves in longitudinal designs or remain the same has implications for their semantic complexity. On the one hand, ordering issues in word-list tasks, where lists with lower entropy (more ordered) yield better recall, are well documented (Melin et al., 2022a,b), and word list complexity further depends on lexical frequency: participants tend to recall high-frequency words more accurately than low-frequency ones, therefore influencing overall task difficulty (Balota and Spieler, 1999). On the other hand, if the word list remains unchanged, its complexity may not remain stable, since respondents may obtain retest and learning benefits from repeating the same word list test over multiple waves - although such benefits usually appear in similarly structured environments rather than with identical content (Salthouse et al., 2004) and has also been partly explained by entropy in recent research (Melin et al., 2023).

A second key issue is that test protocols relating to practice trials are not always identical. The standard CERAD free recall test, for example, has three immediate recall trials before the delay test, providing substantial practice and learning opportunities for respondents. The adapted version of CERAD used in the US Health and Retirement Study (HRS) has only one immediate test before the delayed recall test. Delayed recall depends on how much information participants encode correctly, which can increase through repeated practice (the learning effect) (Karpicke and Roediger iii, 2007) and may be further shaped by word-list complexity.

2.2. Determinants of word recall test

Deficits in strategy, attention, and processing speed, occurring alongside and in conjunction with the aging process can contribute to performance gaps measured in word recall tasks (Anderson et al., 1998; Salthouse, 1996). The levels of processing framework suggests that establishing deep semantic links with words enhances recall performance (Craik and Tulving, 1975). Accordingly, successful word recall depends on encoding strategies that enrich information and facilitate retrieval of the semantic links through self-initiated processing (Tromp et al., 2015). Older adults often struggle with effective encoding strategies compared to younger individuals, leading to insufficient encoding and retrieval failures (Shing et al., 2010). Normal aging also negatively affects spontaneous cue generation, further impairing memory recall (Dunlosky et al., 2005). However, individuals aging normally can still employ effective encoding strategies and benefit from environmental cues. In contrast, patients with Alzheimer's Disease (AD) experience severe encoding deficits that hinder their ability to establish deep semantic connections (Monti et al., 1996). This impairment, coupled with retrieval-stage deficits, significantly reduces their ability to utilize semantic cueing in word recall tasks (Wagner et al., 2012).

2.3. Education and variability in measuring episodic memory

Neuroimaging and behavioral approaches reveal discrepancies in episodic memory measurement. Some people with AD pathology detected through neuroimaging nevertheless perform better than established cut-offs for impairment on cognitive tests, while others fall below (Katzman et al., 1988). Inter-individual variability is also observed within the same measurement approach, as differences in word recall performance may stem from factors beyond normal or pathological aging (Gross et al., 2015). Cognitive reserve (CR) provides an explanation for these inconsistencies, reflecting the brain's adaptability in mitigating the impacts of aging (Pettigrew and Soldan, 2019). As we explore in this study, such adaptability also potentially shapes the capacity of individuals to respond to differences and variety in how word recall tests are administered. Although CR is not directly measurable, education is thought to be related to CR in important ways (Pettigrew and Soldan, 2019; Stern et al., 2020). Higher education levels are associated with enhanced episodic memory performance, evidenced by neuroimaging and behavioral studies. Neuroimaging research links advanced education to greater cortical thickness in aging-sensitive brain regions, suggesting structural resilience against age-related atrophy (Kim et al., 2015). This protective effect extends to episodic memory performance, potentially due to cognitive strategies developed through formal education (Manly et al., 2004). However, education's protective role is complex: while it delays accelerated cognitive decline, its onset may progress more rapidly in highly educated individuals (Hall et al., 2007). Recent longitudinal studies across diverse cultural settings have thus prioritized disentangling the relationship between cognitive proxies and cognitive decline trajectories (Du et al., 2023; Li et al., 2020).

Methodological challenges persist, particularly in cross-wave comparisons. Variations in word-list complexity and test formats across longitudinal waves complicate efforts to measure education-related differences in cognitive trajectories and other cognitive research in longitudinal studies. The China Health and Retirement Longitudinal Study (CHARLS) serves as a prime example of how a longitudinal survey can feature variations in both word-list complexity and test formats across waves. Specifically, the HRS version of the CERAD test was

administered during Waves 1–3, whereas Waves 4–5 introduced the standard CERAD, which included additional practice trials and unfamiliar word lists. Research using CHARLS data to examine cognitive health in a longitudinal context often treats word recall as a core outcome measure, applying diverse analytical approaches. Most studies use raw scores without adjustments (Cadar et al., 2023; Du et al., 2023; Li et al., 2020), even though mean delayed recall in Waves 4–5 is higher than in earlier waves (Wu et al., 2024). Others standardize recall scores via z-scores, further stratifying by education to screen for possible dementia (Chen et al., 2024; Zeng et al., 2025). However, such z-scoring can complicate longitudinal cognitive trajectory analyses, due to confounded frames of reference (Moeller, 2015), if participants' performance is affected by shifting test specifications or if different individuals are sampled across waves. A recent study used data from Waves 1 to 4 of (CHARLS), adopting equated word recall results to address changes in test specifications (Guo et al., 2025). This weighted equipercentile approach employed a two-stage method (Gross et al., 2012): first, deriving an equating sample across waves with similar underlying ability, then using it to create the equipercentile algorithm, which was applied to the full dataset. The equated results better captured the cognitive trajectory, particularly between Waves 3 and 4 (Wu et al., 2024). However, the mean immediate and delayed recall in Wave 5 remains higher than in Wave 4, suggesting room for further refinement of the CHARLS equating process, and additional exploration is needed to clarify the interplay between education, word-list complexity, and test forms.

Equating techniques would be likely enhanced by explicitly taking education into account. Some research suggests that semantically organized word lists provide a disproportionate advantage to older adults with higher education in English-speaking contexts, highlighting a potential spillover effect of education (Frick et al., 2022). Yet, the impact of word complexity in non-English contexts, particularly where content adaptations are absent, remains unexplored. Additionally, variations in test formats, including differences in the number of recall trials, as seen in CHARLS (Zhao et al., 2020), raise questions about whether and how education level moderates episodic memory measures—an area requiring further investigation.

This study examines how education influences episodic memory measurement across varying test formats and word-list complexities. We hypothesize that individuals with higher education deploy advanced strategies during multiple practice trials and handle complex word lists more effectively through deeper semantic encoding. Consequently, alterations in test form and complexity may differentially affect measured cognitive health across education levels. After testing this hypothesis, we adapt Gross et al.'s (2012) equating approach to produce harmonized word-recall scores that remove changes in test form and complexity for longitudinal research in CHARLS. Our innovation harnesses findings from our hypothesis testing to enhance the validity and robustness of the equating technique by: (a) accounting for the hypothesized interacting effects of education with test form and complexity, and (b) exploiting the relative consistency of immediate-recall test form across waves while addressing shifts in word-list complexity.

3. Methodology

3.1. Dataset and study sample

The China Health and Retirement Longitudinal Study (CHARLS) is a nationally representative longitudinal survey targeting adults aged 45 and older, along with their spouses in China. The initial sample was selected through multistage probability sampling, with biennial follow-ups. The CHARLS survey instrument aligns with the Health and Retirement Study (HRS) model (Zhao et al., 2014). From Wave 4 (2018), CHARLS incorporated the Harmonized Cognitive Assessment Protocol (HCAP) to facilitate international and cross-study comparisons of cognitive health among older adults (Gross et al., 2023).

This study utilized CHARLS data from Waves 1 to 5, giving a sample size of 23951 respondents. Respondents younger than 45 or those who partially completed or abandoned the episodic memory test were excluded from the analysis. A flowchart detailing the sample selection process is provided in Fig. S1.

3.1.1. Key variables

Episodic memory was assessed using word recall tasks. In the immediate word recall task, respondents repeated 10 Chinese words read to them. After a short delay, respondents were asked to recall as many words as possible, referred to as delayed word recall. As explained above, the HRS version of the CERAD protocol was used in Waves 1 to 3, which included only one trial for immediate word recall, while the standard CERAD was adopted for Waves 4 and 5, which allows three trials before the delayed recall test. The shift in test format ensured alignment with international cognitive assessment protocols and facilitated cross-study comparisons of episodic memory performance (Zhao et al., 2020) but hampers longitudinal analysis.

The scores for 'immediate word recall' and 'delayed word recall' ranged from 0 to 10, with 1 point awarded for each correct recall and 0 for incorrect recalls. Only the first trial of 'immediate word recall,' completed without prior practice, was recorded in each wave. This approach provided some reliability and consistency in the 'immediate word recall' measure across waves. The complexity of the word list, however, varied across waves because Waves 1–3 used four randomly assigned lists of commonly used Mandarin words, whereas Wave 4 introduced unfamiliar items (e.g., 'butter' and 'queen') for older adults in China, repeating them in Wave 5 (See Table S1).

Education was categorized based on the highest educational attainment reported by respondents: (1) illiterate or incomplete primary school, (2) completed primary school but not high school, and (3) completed at least high school. This classification, rather than a binary division into 'literate' and 'illiterate,' aimed to observe the impact of increasing educational attainment on episodic memory performance and to divide the sample into approximately equal groups for analysis.

Wave serves as an explanatory variable that captures both word-list complexity and test-form differences across Waves 1–5 of CHARLS. Wave 5 serves as the reference category, creating four dummy variables (Waves 1, 2, 3, and 4). We posit that this variable is a proxy for any changes in the word recall test. An effect of these test changes on measured cognitive health is indicated if wave remains a significant predictor of recall scores after controlling for other explanatory and control variables.

Although all waves used the same test form for the immediate test, they employed different word lists. Consequently, significant associations between the wave dummies and immediate-recall scores are interpreted as changes in word-list complexity. For the delayed test, form differences arose specifically between Waves 3 and 4, whereas Waves 1–3 and Waves 4–5 shared identical formats (See Table S1). Hence, significant associations for Waves 1–3 (relative to Wave 5) are attributed to complexity or test-form effects, while Wave 4's significance is interpreted as shifting word-list complexity between Waves 4 and 5.

3.2. Statistical analysis

3.2.1. Regression analysis of word recall test changes

We employed a random intercept model to examine the differential impact of test specifications on measured episodic memory performance across education levels. This multi-level model separates residual variation into within-group (level 1) and between-group (level 2) components. In our analysis, measurement records (person-year) were treated as level 1 and respondents (denoted by 'ID') as level 2.

We conducted two random intercept models to analyze the outcomes

of immediate and delayed word recall. Model 1 for immediate recall is specified as follows:

$$\text{Immediate}_{ij} = \beta_0 + u_{0j} + \beta_1\left(\text{age}_{ij}\right) + \beta_2\left(\text{age}_{ij}^2\right) + \beta_3\left(\text{gender}_j\right) + \beta_4\left(\text{education}_j\right) + \beta_5\left(\text{Wave}_i\right) + \beta_6\left(\text{education}_j \times \text{Wave}_i\right) + \varepsilon_{ij}$$

Where *i* indexes repeated measurements (Level 1), *j* indexes individuals (Level 2), and represents the within-person residual. captures person-specific deviation from the overall intercept, assuming a normal distribution.

Model 2 for delayed recall was constructed similarly to Model 1, with the inclusion of the immediate recall score as a control variable:

$$\text{Delay}_{ij} = \beta_0 + u_{0j} + \beta_1\left(\text{Immediate}_{ij}\right) + \beta_2\left(\text{age}_{ij}\right) + \beta_3\left(\text{age}_{ij}^2\right) + \beta_4\left(\text{gender}_j\right) + \beta_5\left(\text{education}_j\right) + \beta_6\left(\text{Wave}_i\right) + \beta_7\left(\text{education}_j \times \text{Wave}_i\right) + \varepsilon_{ij}$$

As explained, 'Wave' is treated as a proxy for the administration and difficulty of the word recall test. In Model 1, Wave is assumed to reflect word list difficulty, while in Model 2, Wave may reflect word list difficulty and/or the number or trials (test form) before the immediate test. The interaction between education and Wave is assumed to reflect the differential effect of the test specifications across education groups.

3.2.2. *Equating word list recall scores across waves*

We adapted the two-stage approach proposed by Gross et al. (2012) to equate word recall scores in CHARLS. Our objective here was to derive a revised set of word recall test scores in Waves 1 to 4 that removed the effect of test administration and made them comparable to Wave 5 scores. Wave 5 was chosen as the reference wave as this wave is governed by the CERAD form and facilitates easier comparison of equated data with future waves, which are likely to continue using the standard CERAD form rather than the HRS-adapted version.

The Gross et al. (2012) approach is based on equipercentile equating. In our context, equipercentile equating involves calculating the percentile rank of each participant's word recall test score in each wave and then adjusting each individual score to match what their score would be in Wave 5 if they had the same percentile rank. For example, if a score gave an individual a rank of 25th percentile in Wave 1, that raw score would be revised to the score that corresponds to the 25th percentile in Wave 5. If the 25th percentile was a score of 5 in Wave 1 and 10 in Wave 5, the Wave 1 score would be revised upwards from five to 10. The problem with this approach in its basic form is that it assumes that underlying cognitive performance remains constant between waves, where in fact it is likely in a longitudinal context that factors like biological aging have contributed to cognitive decline.

We adopted the idea of a two-stage approach to control for aging. The first stage aimed to define sub-samples in each wave. Respondents were included in sampling frames if they remained in the same quartile of total word recall scores in Wave 5. These individuals were then stratified into ten-year age groups and randomly selected into final subsamples, with age distributions matching the waves being equated (Waves 1–4). During sampling, respondents were divided by their educational levels. This subsampling ensured that the equating algorithm accounted for the hypothesized interacting effect of test specifications and education. When identifying respondents with consistent episodic memory performance between targeted waves, comparisons were made within their respective educational levels, rather than across all levels. In the second stage, we derived equipercentile equating algorithms and applied them to the full sample. The overall word recall test (immediate + delayed recall) in Wave 5 was treated as the reference test. To find the equipercentile equivalent of Wave 1 scores in the Wave 5 scale, the percentile rank of scores in Wave 1's distribution was identified, and the corresponding score in the Wave 5 distribution was determined (Kolen and Brennan, 2014).

Despite the strengths of equipercentile equating, the approach rests on the strong assumption that respondents maintain consistent episodic memory function between waves (albeit after stratifying by age and education), and the order of waves does not affect their scores (Kolen and Brennan, 2014). To relax this assumption further and strengthen the robustness of the equating process, we propose to combine equipercentile equating with frequency estimation in what is known as frequency estimation in common-item non-equivalent groups. In our context, the idea is that we can equate scores across waves even among very different groups of respondents by utilizing a common set of questions or an anchor test that is consistently administered across survey waves. The strength of this approach is that rather than equating scores where we have to assume that our groups are the same across waves, we can equate on the basis of equivalent responses to the common items. In this study, we propose utilizing the results of the immediate recall test as the anchor test.

The analytical process of frequency estimation is similar to standard equipercentile equating. The main difference is that in frequency estimation, we perform equipercentile equating on synthetic cumulative distributions rather than the observed distributions in each wave. Adapting the steps set out by Kolen and Brennan (2014), we create two synthetic distributions for each pair of waves, 1 and 5, 2 and 5, 3 and 5, and 4 and 5. For the Wave 1 and 5 pair, the first distribution comprises total recall scores against the Wave 1 test form, combining the actual scores of Wave 1 respondents with predicted scores of Wave 5 respondents on the Wave 1 test form, given their observed scores on the immediate recall test:

$$f_s(t_1) = 0.5\ (f_1(t_1) + f_1(t_1\ |v)h_5(v))$$

is the cumulative distribution of total word recall scores for Wave 1 respondents on the Wave 1 test, ; is the cumulative distribution of scores that we predict Wave 5 respondents would have scored if they took the Wave 1 test. Predicted scores are calculated by taking Wave 5 respondents' observed scores on the immediate recall test, and multiplying them by the observed relationship between the immediate and total recall scores among Wave 1 respondents, . We create the synthetic distribution by taking the average of these distributions.

The second distribution comprises total recall scores against the Wave 5 test form. The calculation is the same as for the first distribution except in reverse, combining the actual scores of Wave 5 respondents with the predicted scores of Wave 1 respondents on the Wave 5 test form:

$$f_s(t_5) = 0.5\ (f_5(t_5) + f_5(t_5\ |v)h_1(v))$$

We then perform equipercentile equating on the synthetic distributions, calculating percentile ranks for individual scores in each distribution and then revising each Wave 1 respondents' scores by finding the percentile rank on the Wave 5 synthetic distribution - and its corresponding total recall score - that matches their percentile rank on the synthetic Wave 1 distribution.

Having introduced the two designs for equating - single item

equipercentile equating and frequency estimation in common-item nonequivalent groups - the decision on which design to adopt depends on the compatibility of the test forms with the assumptions of each design. As mentioned, the single-group design has stricter assumptions about the sample. While respondents in the sample may have the same rank of word recall scores across waves, their true episodic memory function may differ in unobserved ways. In contrast, the common-item nonequivalent groups design has less restrictive assumptions about the sample but introduces an additional assumption requiring that the immediate recall test is delivered consistently in administration and content, including in relation to word list difficulty.

The key limitation in using the immediate recall test as the common items is that word lists changed across waves and as already explained, likely resulted in changes in their complexity. A potential solution is to apply frequency estimation only to waves with minimal anchor-test complexity variation (Liu et al., 2011a, 2011b), which the regression model we constructed above identifies by determining waves with similar complexity. Consequently, the common-item nonequivalent groups design is adopted if there is minimal variability in word list difficulty between targeted waves. Otherwise, the single-group design would be chosen to ensure methodological appropriateness.

3.2.3. Constructing trajectories of cognitive test scores

Line graphs illustrating predicted mean scores over time were generated to compare the trajectory of both equated and raw outcomes on the word recall test. Utilizing previous results, we identify our proposed equating method and compare the resulting trajectories against those produced through other equating approaches, including equipercentile equating in single-group designs and frequency estimation in non-equivalent groups with and without stratifying by education. Predicted scores were estimated via a random-intercept model that included wave, education, and age group (younger than 65 vs. 65 or older), plus their interactions, with random intercepts for participants to account for individual-level variability. We also fit latent growth curve models to raw and equated scores, allowing us to calculate the model fit statistics, the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). As our approach involves fitting the same model to different sets of underlying data rather than the conventional application of testing different models on the one dataset, our model fit statistics do not tell us which is the best equating approach. Rather, they provide insights as to which approach provides the basis for well-fitting models that potentially pave the way for further research and greater understanding of cognitive health trajectories, their determinants and consequences.

4. Results

4.1. Sample characteristics

In total, 75,404 person-years were included in the study, with general characteristics across waves presented in Table 1. Respondents with low educational attainment comprised a larger portion of the sample. As educational attainment levels increased, the sex ratio reversed, with males more likely to achieve higher education.

Immediate recall scores showed a steady decline from Wave 1 to Wave 3, followed by a significant drop in Wave 4 and a rebound in Wave 5. Delayed recall scores similarly declined from Wave 1 to Wave 3 but were followed by two consecutive increases. This trend was consistent across educational levels, except for respondents with high educational attainment. Notably, in each wave, respondents with higher educational attainment consistently achieved higher scores in both immediate and delayed word recall tests, highlighting the positive association between education and memory performance.

4.2. Predicting the impact of changes in word list complexity and test form

The stage one regression results are shown in Table 2. As indicated by the Intraclass Correlation Coefficients (ICC), 30 % and 13 % of the unexplained variation in Immediate and Delayed recall test scores, respectively, are due to unexplained differences within individuals across waves. In Model 1 for Immediate recall, the difficulty of the word list varies across waves after controlling for fixed effect covariates and unexplained variation in random effects. Specifically, the coefficient for Wave 4 (−0.619) indicates that the word list administered in Wave 4

Table 1
Basic characteristics of participants by education levels, CHARLS, Wave 1 to Wave 5.

Characteristics	2011	2013	2015	2018	2020
a. Low Education					
Age at survey					
Mean (SD)	60.8 (9.79)	61.5 (9.57)	61.3 (9.95)	62.9 (9.5)	64.8 (9.24)
Female (%)	4055 (66.5)	4407 (66.7)	4857 (66.5)	3159 (67.5)	4308 (69.2)
Immediate word recall					
Mean (SD)	3.53 (1.58)	3.37 (1.72)	3.24 (1.72)	2.38 (1.83)	2.84 (1.68)
Delayed word recall					
Mean (SD)	2.77 (1.59)	2.53 (1.90)	2.20 (1.91)	3.02 (2.61)	3.91 (2.51)
Total Counts (%)	6094 (44.9)	6612 (44.4)	7299 (41.2)	4677 (34.7)	6223 (39.5)
b. Medium Education					
Age at survey					
Mean (SD)	56.7 (8.85)	57.5 (9.04)	57.1 (9.57)	59.5 (9.27)	61.1 (9.08)
Female (%)	2378 (41.4)	2712 (42,1)	3446 (42.3)	2914 (43.4)	3226 (43.8)
Immediate word recall					
Mean (SD)	4.39 (1.58)	4.42 (1.63)	4.33 (1.65)	3.79 (1.78)	3.92 (1.56)
Delayed word recall					
Mean (SD)	3.56 (1.76)	3.56 (1.90)	3.34 (1.94)	5.32 (2.41)	5.54 (2.19)
Total Counts (%)	5743(42.2)	6443 (43.3)	8152 (46.0)	6727 (49.9)	7368 (46.8)
b. High Education					
Age at survey					
Mean (SD)	55.6 (8.74)	56.6 (8.40)	56.7 (9.05)	59.0 (8.76)	60.7 (8.27)
Female (%)	665 (38.0)	665 (36.5)	816 (35.9)	785 (38.0)	806 (37.3)
Immediate word recall					
Mean (SD)	5.18 (1.60)	5.26 (1.62)	5.21 (1.61)	4.47 (1.66)	4.62 (1.52)
Delayed word recall					
Mean (SD)	4.31 (1.82)	4.44 (1.93)	4.23 (1.87)	6.27 (2.17)	6.35 (2.06)
Total Counts (%)	1749(12.9)	1824 (12.3)	2270 (12.8)	2065 (15.3)	2158 (13.7)

Note: CHARLS: China Health and Retirement Longitudinal Study; SD: standard deviation.

Table 2
Random intercept model results for Immediate and delay word recall scores.

Variables	Model 1 (Immediate)	Model 2 (Delay)
	Coefficient (95 %CI)	
Fixed effect		
Intercept	3.028 (2.574, 3.481) ***	3.142 (2.718, 3.567) ***
Immediate word recall		0.713 (0.706, 0.729) ***
Age	0.040 (0.026, 0.055) ***	−0.012 (−0.025, 0.002).
Age^2	−0.001 (−0.001, 0.000) ***	−0.0001 (−0.0002, 0.000) *
Gender (Ref. Male)	0.127 (0.095, 0.159) ***	0.066 (0.038, 0.093) ***
Education (Ref. Low)		
Medium	0.955 (0.900, 1.009) ***	0.772 (0.717, 0.826) ***
High	1.659 (1.581, 1.738) ***	1.077 (0.998, 1.156) ***
Wave (Ref. Wave 5)		
Wave 1	0.518 0.467, 0.569) ***	−1.730 (−1.784, −1.677) ***
Wave 2	0.393 (0.344, 0.442) ***	−1.841 (−1.893, −1.789) ***
Wave 3	0.249 (0.202, 0.296) ***	−2.087 (−2.138, 2.036) ***
Wave 4	−0.619 (−0.672, −0.566) ***	−0.629 (−0.686, −0.572) ***
Interaction of Wave x Education (Ref. Wave 5 x Low)		
Wave 1 x Medium	−0.216 (−0.285, −0.146) ***	−0.698 (−0.772, −0.623) ***
Wave 1 x High	−0.146 (−0.248, −0.043) **	−0.839 (−0.948, −0.729) ***
Wave 2 x Medium	−0.022 (−0.089, 0.045)	−0.583 (−0.656, −0.511) ***
Wave 2 x High	0.094 (−0.005, 0.194).	−0.619 (−0.726, −0.512) ***
Wave 3 x Medium	0.021 (−0.043, 0.085)	−0.514 (−0.583, −0.445) ***
Wave 3 x High	0.186 (0.092, 0.280) ***	−0.557 (−0.658, −0.456) ***
Wave 4 x Medium	0.389 (0.319, 0.458) ***	0.452 (0.376, 0.527) ***
Wave 4 x High	0.386 (0.288, 0.485) ***	0.595 (0.489, 0.702) ***
Random effect		
ID	0.7815	0.3291
ICC	0.3000	0.1320
AIC	278873.3000	281608.4000

Note: 95 % confidential intervals are in parentheses. ***p < 0.001, **p < 0.01, *p < 0.05, .p < 0.1.

was the most challenging for low-educated people, while Wave 5 was slightly harder but comparable to other waves (Wave 5: baseline; Wave 1: 0.518; Wave 2: 0.393; Wave 3: 0.249). Education was significantly associated with Immediate recall scores, with individuals with higher education outperforming their lower-educated counterparts in Wave 5 (Education Low as baseline vs. Education High: 1.659). Significant interactions between the high-education level and wave suggest that changes in word-list difficulty were moderated by education, whereas this effect may not apply to individuals with a medium-education level in Waves 2 and 3. Higher-educated respondents performed relatively better with the more challenging Wave 4 word list, widening the performance gap. For instance, if tested in Wave 1 (the easiest word list), the gap between the highest and lowest education levels was 1.513. Conversely, in Wave 4 (the most difficult word list), the gap widened to 2.045.

In Model 2 for the Delayed recall test, the coefficient of Wave steadily declined in the first three waves (−1.730, −1.841, −2.087) and rose in the last two waves (−0.629, 0). Thus, in test forms with more practice opportunities (Waves 4 and 5), the coefficient of Wave was significantly higher compared to those with only one practice opportunity (Waves 1–3), irrespective of word list difficulty.

Education is also associated with delayed recall performance. High and medium educated individuals performed significantly better on the delay test in Wave 5, indicated by the main effects of education in Model 2 of Table 2. Notably, the interaction between education level and Wave was also significantly associated with delay test scores. Specifically, in Wave 4, with a more difficult word list but more practice opportunities, individuals with higher educational attainment performed better, widening the performance gap. Similarly, in Wave 5, with the same practice opportunities but a comparable word list in terms of difficulty Waves 1–3, higher educated people also performed better.

Fig. 1 illustrates the predicted Delayed recall scores for males in each wave, controlling for age (set to 70) and immediate recall score (set to the mean). The lines show predicted delay scores by wave and education level. The columns show the predicted difference or gap in delayed recall scores between highly educated and low-educated respondents. The gap is predicted to widen from 0.5 points in Wave 3 to 1.7 points in Wave 4. Thus, higher educated people are predicted to benefit more in test forms with more practice opportunities (Wave 4 and Wave 5). The significant decline in the education gap to Wave 5 (from 1.7 to 1.1 points) suggests that within the same test form, respondents with low educational attainment were more negatively impacted by increased word list difficulty than their higher-educated counterparts. The measured score gap widens in test forms with more practice opportunities and increases further with more challenging word lists, reflecting differences in the measurement of performance rather than true ability.

4.3. Equating and predicting total word recall scores

The regression analysis suggests that changes in the word recall test measurably affected raw scores, and this was significantly moderated by education. Incorporating education into the equating process is thus warranted to account for this differential effect. The results also suggest that the apparent consistency of the immediate word recall test was interrupted by changes in the Wave 4 content. Thus, while treating the immediate test as a set of common items between some wave pairs, it is not appropriate for equating Wave 4.

We therefore implement a hybrid equating approach. Specifically, we use frequency estimation with the immediate test as a common item to equate scores from Waves 1–3 to Wave 5, then apply standard equipercentile equating for Wave 4 to Wave 5. The sample sizes used for equating are shown in Table S5. Fig. 2B shows the results based on equating each education group separately, while Fig. 2D shows the results based on equating over all education groups. We also contrast these results with raw, un-equated outcomes (2A) and with standard equipercentile equating lacking education-based stratification (2C).

The hybrid frequency estimation approaches are likely to effectively capture declining word scores over time, a potential indicator of cognitive decline (though noting that there is not necessarily a linear relationship between raw scores and cognitive decline). In un-equated data (Fig. 2A), predicted word recall scores were higher after a nine-year gap across all age and education groups, echoing patterns in Fig. 1 and raising concerns about measurement reliability. Equipercentile equating without education subgroups (Fig. 2C) exhibited similar issues, as predicted trajectories rarely declined, contradicting known aging effects. Both hybrid approaches (Fig. 2B and D) yielded generally downward-trending scores, although patterns differed by education level. Specifically, for high-education groups, stratified equating produced an initial sharp decline followed by a plateau, while the method without education sub-groups produced a steady decline. This pattern was reversed for low-education groups. While we do not know which is closer to the true underlying pattern of cognitive performance, significant interactions between test form and education (Table 2) underscore the need to incorporate education into equating approaches.

The hybrid frequency estimation approaches also allow for better-fitting latent growth curve models. The un-equated data exhibited non-linear changes, leading to relatively poor fitting models (RMSEA = 0.103; CFI = 0.915). All equating methods allowed for better model fit. The hybrid approach with education sub-groups enabled a model with the best fit (RMSEA = 0.040; CFI = 0.986), followed by equipercentile equating with education sub-groups (RMSEA = 0.044; CFI = 0.983) and hybrid frequency estimation without education (RMSEA = 0.049; CFI =

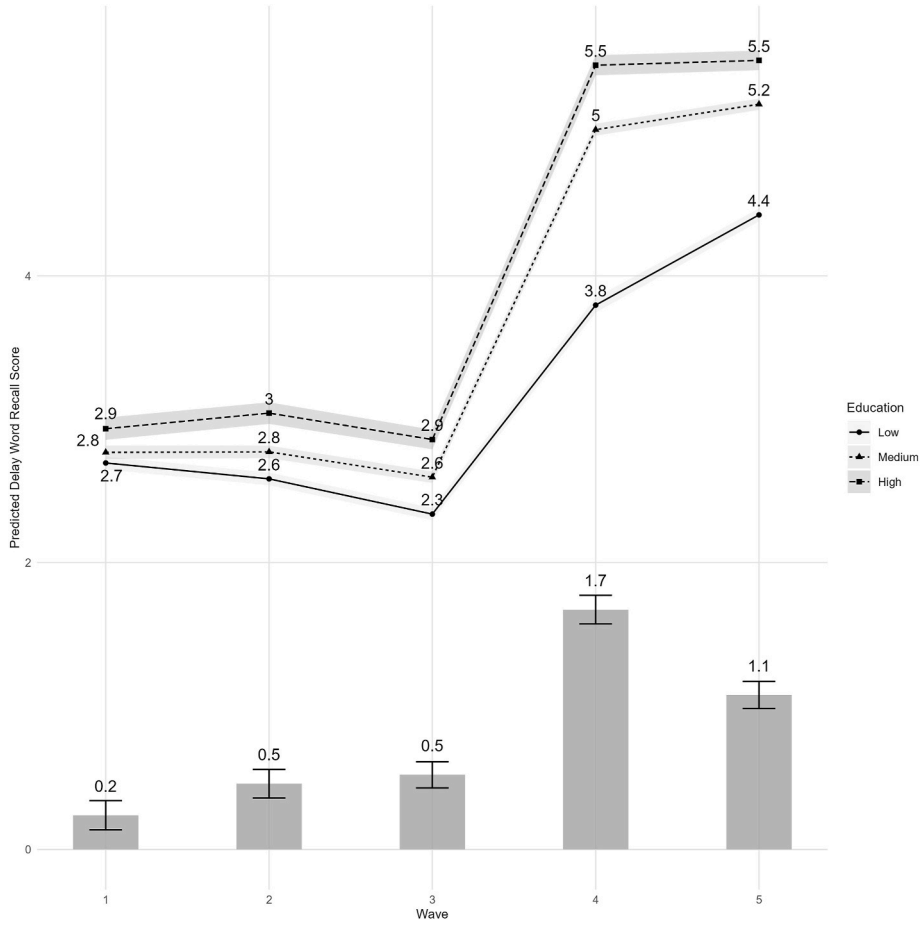

Fig. 1. Predicted Delayed Recall Scores by Wave and Education Level, Estimated by Random Intercept Model
Note: Gender is set as male, age is fixed at 70, and the immediate recall score is held at the mean. The shaded area around the line and the error bars indicates the 95 % confidence interval. The bar chart also shows the predicted gap in delayed recall scores between respondents with high and low education levels.

0.981). The full set of latent growth curve model results are provided in the online Supplementary Material (Table S3).

5. Discussion

Our study underscores the influence of education on word recall measurement within the quasi-experimental design observed in the CHARLS. Our findings indicate that education exerts a strong effect on episodic memory assessment, as test specifications—including word list complexity and test form—interact with participants' education level and thus shape the performance gap between those with higher and lower educational attainment. Specifically, when complexity remains consistent, participants across all educational levels tend to score higher under test forms offering more practice opportunities. However, these gains are uneven: those with higher education benefit more significantly, widening the performance gap. Conversely, if the test form remains unchanged, participants with lower educational attainment prove more vulnerable to variations in word list complexity, resulting in notably lower scores on more challenging lists. Meanwhile, those who have completed at least primary school appear less affected by such complexity.

We suggest that the presence of a semantic link, along with more effective encoding strategies, helps explain why education exerts an additional influence on word recall measurements. First, the complex version of the word list—lacking local adaptation—may hinder participants with lower education from forming semantic links crucial for establishing retrievable memory traces (Craik and Tulving, 1975). This obstacle stems from cultural gaps, particularly for older adults in China who have minimal foreign-language exposure and encounter Western items like "Butter" or "Queen" in CHARLS Wave 4. In contrast, those with greater schooling often receive English-language education, making them more familiar with such items. Language learning background, as indicated by higher education, thus plays a key role in bridging cultural references (Hossain, 2024). Moreover, individuals with more schooling can implement sophisticated encoding strategies, especially in repeated retrieval settings proven to support long-term retention (Karpicke and Roediger iii, 2007). Considering that they are accustomed to such environments, their performance gains from practice are amplified by employing more effective encoding strategies, further magnifying the measurement gap.

Our study proposes an equating method to address differences in test specifications across waves, taking into account education's impact on word recall measurement. We examine different sets of raw and equated scores, including those from alternative equating approaches. While we cannot say which method most closely approximates the unmeasured truth, our analysis of cognitive trajectories helps to determine their plausibility and utility for longitudinal research. Wu et al.'s (2024) recent study also applied equipercentile equating through a two-stage process and were likewise successful in producing more plausible trajectories. Nevertheless, they produce a significant and counter-intuitive uptick in recall scores between Waves 4 and 5. By incorporating education into our equating approach and utilizing frequency estimation, we are able to address this problem and build on existing literature. Compared with applying equipercentile methods across all waves, frequency estimation in a non-equivalent design—mainly employed in our Wave 1–3 equating—offers a more feasible theoretical assumption and

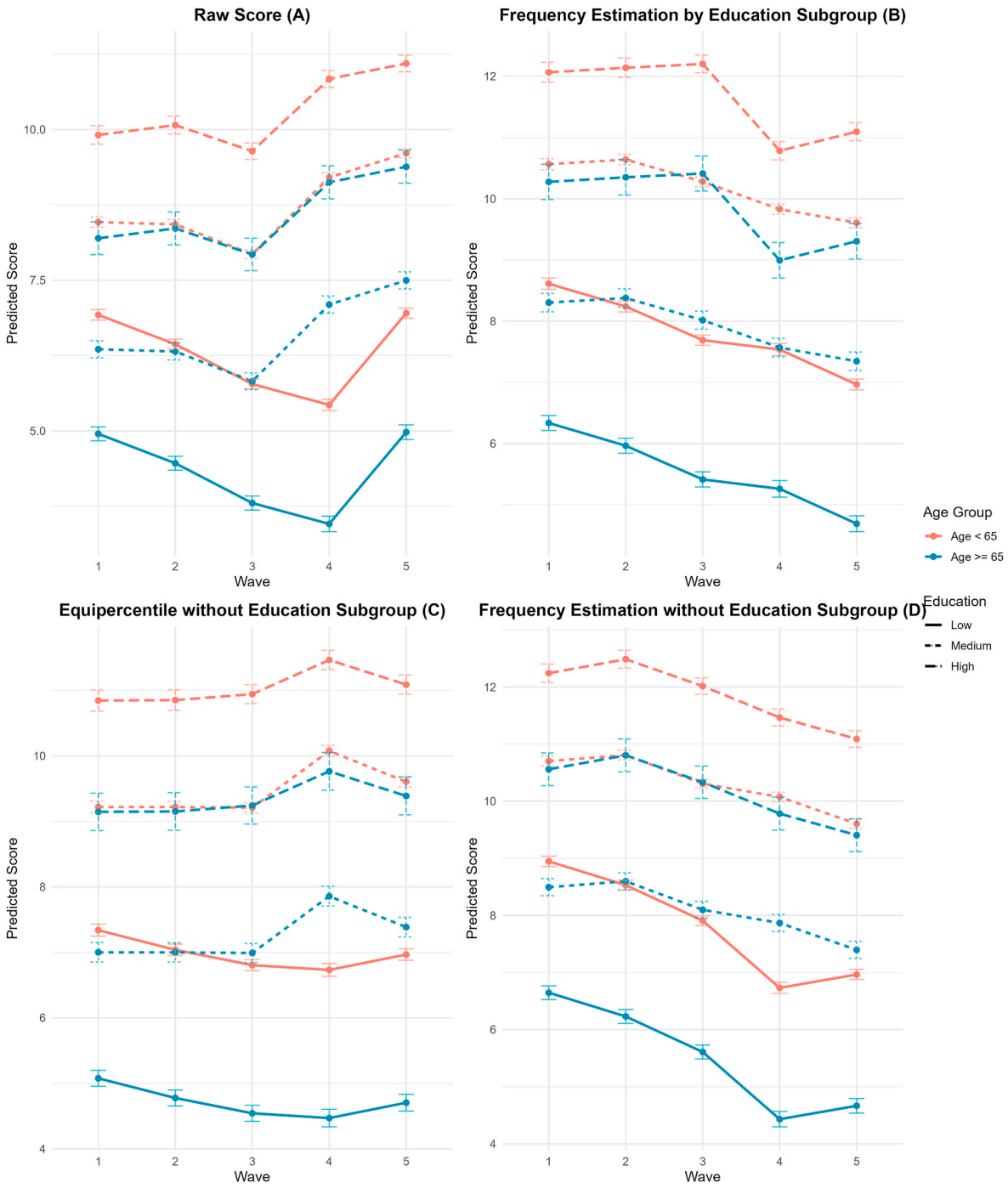

Fig. 2. Trajectories of Predicted Equated and Raw Word Recall Scores, Estimated by a Random Intercept Model
Note: A 95 % confidence interval is provided for each predicted score point.

better equated outcomes. Specifically, frequency estimation relies on similar 'content specifications', referring to comparable word list complexity, rather than presuming an identical underlying ability across tested populations (Kolen and Brennan, 2014). Thus, if future CHARLS waves maintain analogous word list complexity to Waves 1–3 and 5 and preserve the standard CERAD test form, their raw scores could be analyzed in tandem with equated results without further equating. Finally, applying equipercentile equating to Wave 4 with Wave 5 as the reference accounts for potential differences in word list complexity stemming from retest effects, thereby facilitating cross-wave comparisons. Survey attrition may potentially limit equipercentile equating, particularly by undermining sample representativeness when participants drop out. However, frequency estimation remains robust: the relationship between delay and immediate recall scores still holds after accounting for attrition (Table S6b), notably mitigating this concern.

Nevertheless, this study has certain other limitations. In analyzing the interacting effects of education and test administration on immediate and delayed recall scores, we used the survey wave as a proxy for test administration. Although we control for age and potential curvilinear impacts of aging, the survey wave nevertheless potentially conflates the effects of test administration with more complex effects of aging and other time-varying factors. Thus, we cannot determine with any certainty that the main and interacting associations between wave and

cognitive scores reflect administration effects. As discussed in the paper, the administration effects are themselves comprised of at least two issues, changes in the content and difficulty of the word list and changes in the number of practice tests administered before the delay test. While we can exploit aspects of test administration to an extent to infer the potential contributing effects of each, we lack a rigorous experimental design. Specifically, we lack a comparison wave in which a word list is administered with high complexity alongside only one practice trial, potentially questioning the logical basis for inferring complexity effects in limited practice conditions. Moreover, since CHARLS provided complete comparisons exclusively for multiple-practice forms, extrapolating these results to single-trial test form remains tentative, limiting the external validity of our findings. Finally, our analysis also did not examine word placement and its effect on recall, known as serial position effects (SPEs), which serve as a marker of neurodegeneration (Melin et al., 2022; Weitzner & Calania, 2020). Future studies should investigate potential interactions between SPEs and the test specifications assessed here within the present study.

Our proposed equating method has several limitations. First, the approach uses frequency estimation for Waves 1–3 and equipercentile for Wave 4. As equipercentile equating presumes identical underlying ability in test-taking populations, we addressed this theoretical concern through the two-step process, selecting participants for the equating algorithm who maintained the same score rank across both waves in the process of sample construction and maintaining the same age distribution across waves. However, critics might argue that other time-varying factors could shift cognitive performance and bring into question the comparability of the sub-samples used for equating. The use of frequency estimation to equate Waves 1–3, meanwhile, rests on the assumed consistency in the administration of the immediate recall test, where we know in fact that the word list content was different in Wave 5 (albeit with muted measured effects on performance). Second, our method focuses exclusively on the influence of education, one convenience proxy for cognitive reserve, in measuring word recall. Other proxies, such as social engagement and healthy lifestyles, may also shape test specifications by allowing older adults to devote more attention during encoding and retrieval (Stern et al., 2020).

Finally, the non-linearity observed in the prediction analysis of equated scores may reflect cognitive reserve yet might also stem from limitations of Classic Test Theory (CTT), which assumes performance scores function as interval-level data. As scores approach the upper or lower scale bound, this non-linearity challenges the validity of episodic memory trajectories. To mitigate this counted-fraction phenomenon, future studies should integrate equating methods with modern measurement theories such as the Rasch model, which has addressed ordering issues in word recall tasks (Melin et al., 2022a,b; Pendrill, 2018). By combining these frameworks, researchers can improve overall measurement precision of episodic memory and more rigorously test the validity of cognitive reserve across diverse test formats.

6. Conclusion

Population aging across the world will bring degenerative diseases, most especially dementia and cognitive impairment, to the forefront of public health policy and research in coming decades. While the body of research in the United States and Europe continues to grow, China stands out as a country at the beginning of a period of rapid population aging and a growing burden of dementia (Liu et al., 2024; United Nations Population Division, 2024). This underlines the importance of high-quality longitudinal data such as CHARLS, in quantifying this burden, its correlates and causes, and informing policy to prevent and manage its personal and societal impacts. To fully harness its capabilities and draw meaningful and reliable insights, however, we need strategies to ensure that the information fed into longitudinal analyses is consistent across time.

While work is well underway to ensure consistency in CHARLS, our study points to ways in which this can be further strengthened. Our study illustrates how education shapes the measurement of episodic memory, specifically through the word recall test. The measured outcome varies with two main test-specification factors influenced by respondents' education. Specifically, when the word list is more demanding, or the test form provides multiple trials, the performance gap between higher and lower educated participants widens. Drawing on our empirical findings, we propose an equating method to facilitate longitudinal studies where word recall tests use different specifications. Our approach accounts for education's impact while highlighting a role for frequency estimation in strengthening the equating solution. We suggest that frequency estimation should be employed between waves sharing similar word-list complexity, whereas equipercentile equating is more suitable when complexity differs substantially. Future research should explore whether additional cognitive reserve proxies, such as social integration or healthy lifestyles, also contribute to variability in measuring episodic memory across various test specifications.

CRediT authorship contribution statement

Yizhou Chen: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **James O'Donnell:** Writing – review & editing, Supervision, Methodology, Investigation.

Ethical approval

Ethical approval for China Health and Retirement Longitudinal Study (CHARLS) was obtained from the Institutional Review Board at Peking University. IRB approval number IRB00001052-11015.

Acknowledgements

Yizhou Chen was supported by the China Scholarship Council (File No. 202508190005) and Australian National University (No. 675/2014). James O'Donnell received funding from the Australian Research Council Discovery Early Career Researcher Award (Project No. DE240100232) funded by the Australian Government.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2025.118473.

Data availability

The data is available in the website of the China Health and Retirement Longitudinal Study (CHARLS) at https://charls.pku.edu.cn/en/.

References

Anderson, N.D., Craik, F.I., Naveh-Benjamin, M., 1998. The attentional demands of encoding and retrieval in younger and older adults: I. Evidence from divided attention costs. Psychol. Aging 13, 405.

Balota, D.A., Spieler, D.H., 1999. Word frequency, repetition, and lexicality effects in word recognition tasks: beyond measures of central tendency. J. Exp. Psychol. Gen. 128, 32.

Cadar, D., Brocklebank, L., Yan, L., Zhao, Y., Steptoe, A., 2023. Socioeconomic and contextual differentials in memory decline: a cross-country investigation between England and China. J. Gerontol.: Ser. Bibliogr. 78, 544–555.

Chen, S., Chen, X., Hou, X., Fang, H., Liu, G.G., Yan, L.L., 2024. Temporal trends and disparities of population attributable fractions of modifiable risk factors for dementia in China: a time-series study of the China health and retirement longitudinal study (2011–2018). The Lancet Regional Health–Western Pacific 47.

Craik, F.I., Tulving, E., 1975. Depth of processing and the retention of words in episodic memory. J. Exp. Psychol. Gen. 104, 268.

Dickerson, B.C., Eichenbaum, H., 2010. The episodic memory system: Neurocircuitry and disorders. Neuropsychopharmacology 35, 86–104. https://doi.org/10.1038/npp.2009.126.

Du, Y., Hu, N., Yu, Z., Liu, X., Ma, Y., Li, J., 2023. Characteristics of the cognitive function transition and influencing factors among Chinese older people: An 8-year longitudinal study. J. Affect. Disord. 324, 433–439. https://doi.org/10.1016/j.jad.2022.12.116.

Dunlosky, J., Hertzog, C., Powell-Moman, A., 2005. The contribution of mediator-based deficiencies to age differences in associative learning. Dev. Psychol. 41, 389.

Fillenbaum, G.G., van Belle, G., Morris, J.C., Mohs, R.C., Mirra, S.S., Davis, P.C., Tariot, P.N., Silverman, J.M., Clark, C.M., Welsh-Bohmer, K.A., others, 2008. Consortium to establish a Registry for Alzheimer's disease (CERAD): the first twenty years. Alzheimer's Dementia 4, 96–109.

Ford, K.J., Batty, G.D., Leist, A.K., 2021. Examining gender differentials in the association of low control work with cognitive performance in older workers. Eur. J. Publ. Health 31, 174–180.

Frick, A., Wright, H.R., Fay, S., Vanneste, S., Angel, L., Bouazzaoui, B., Taconnat, L., 2022. The protective effect of educational level varies as a function of the difficulty of the memory task in ageing. Eur. J. Ageing 19, 1407–1415. https://doi.org/10.1007/s10433-022-00724-z.

Grober, E., Sanders, A.E., Hall, C., Lipton, R.B., 2010. Free and cued selective reminding identifies very mild dementia in primary care. Alzheimer Dis. Assoc. Disord. 24, 284–290.

Gross, A.L., Inouye, S.K., Rebok, G.W., Brandt, J., Crane, P.K., Parisi, J.M., Tommet, D., Bandeen-Roche, K., Carlson, M.C., Jones, R.N., 2012. Parallel but not equivalent: challenges and solutions for repeated assessment of cognition over time. J. Clin. Exp. Neuropsychol. 34, 758–772. https://doi.org/10.1080/13803395.2012.681628.

Gross, A.L., Li, C., Briceño, E.M., Rentería, M.A., Jones, R.N., Langa, K.M., Manly, J.J., Nichols, E., Weir, D., Wong, R., others, 2023. Harmonisation of later-life cognitive function across national contexts: results from the Harmonized Cognitive Assessment Protocols. The Lancet Healthy Longevity 4, e573–e583.

Gross, A.L., Mungas, D.M., Crane, P.K., Gibbons, L.E., MacKay-Brandt, A., Manly, J.J., Mukherjee, S., Romero, H., Sachs, B., Thomas, M., others, 2015. Effects of education and race on cognitive decline: An integrative study of generalizability versus study-specific results. Psychol. Aging 30, 863.

Guo, M., Wu, Y., Gross, A.L., Karvonen-Gutierrez, C., Kobayashi, L.C., 2025. Age at menopause and cognitive function and decline among middle-aged and older women in the China Health and Retirement Longitudinal Study, 2011–2018. Alzheimer's Dementia 21, e14580.

Hall, C., Derby, C., LeValley, A., Katz, M., Verghese, J., Lipton, R., 2007. Education delays accelerated decline on a memory test in persons who develop dementia. Neurology 69, 1657–1664.

Hossain, K.I., 2024. Reviewing the role of culture in English language learning: Challenges and opportunities for educators. Soc. Sci. Humanit. Open 9, 100781.

Karpicke, J., Roediger‖, H., 2007. Repeated retrieval during learning is the key to long-term retention☆. J. Mem. Lang. 57, 151–162. https://doi.org/10.1016/j.jml.2006.09.004.

Katzman, R., Terry, R., DeTeresa, R., Brown, T., Davies, P., Fuld, P., Renbing, X., Peck, A., 1988. Clinical, pathological, and neurochemical changes in dementia: a subgroup with preserved mental status and numerous neocortical plaques. Ann. Neurol.: Official Journal of the American Neurological Association and the Child Neurology Society 23, 138–144.

Kim, J.P., Seo, S.W., Shin, H.Y., Ye, B.S., Yang, J.-J., Kim, C., Kang, M., Jeon, S., Kim, H. J., Cho, H., Kim, J.-H., Lee, J.-M., Kim, S.T., Na, D.L., Guallar, E., 2015. Effects of education on aging-related cortical thinning among cognitively normal individuals. Neurology 85, 806–812. https://doi.org/10.1212/WNL.0000000000001884.

Kolen, M.J., Brennan, R.L., 2014. Test Equating, Scaling, and Linking: Methods and Practices. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4939-0317-7.

Li, C., Zhu, Y., Ma, Y., Hua, R., Zhong, B., Xie, W., 2022. Association of cumulative blood pressure with cognitive decline, dementia, and mortality. J. Am. Coll. Cardiol. 79, 1321–1335. https://doi.org/10.1016/j.jacc.2022.01.045.

Li, H., Li, C., Wang, A., Qi, Y., Feng, W., Hou, C., Tao, L., Liu, X., Li, X., Wang, W., others, 2020. Associations between social and intellectual activities with cognitive trajectories in Chinese middle-aged and older adults: a nationally representative cohort study. Alzheimers Res. Ther. 12, 1–12.

Liu, J., Sinharay, S., Holland, P., Feigenbaum, M., Curley, E., 2011a. Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. Educ. Psychol. Meas. 71, 346–361. https://doi.org/10.1177/0013164410375571.

Liu, J., Sinharay, S., Holland, P.W., Curley, E., Feigenbaum, M., 2011b. Test score equating using a Mini-Version anchor and a midi anchor: a case study using SAT® data. J. Educ. Meas. 48, 361–379.

Liu, Yuyang, Wu, Y., Chen, Y., Lobanov-Rostovsky, S., Liu, Yixuan, Zeng, M., Bandosz, P., Xu, D.R., Wang, X., Liu, Yuanli, others, 2024. Projection for dementia burden in China to 2050: a macro-simulation study by scenarios of dementia incidence trends. The Lancet Regional Health–Western Pacific 50.

Livingston, G., Huntley, J., Liu, K.Y., Costafreda, S.G., Selbæk, G., Alladi, S., Ames, D., Banerjee, S., Burns, A., Brayne, C., others, 2024. Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission. Lancet 404, 572–628.

Manly, J.J., Byrd, D., Touradji, P., Sanchez, D., Stern, Y., 2004. Literacy and cognitive change among ethnically diverse elders. Int. J. Psychol. 39, 47–60.

Melin, J., Cano, S., Göeschel, L., Fillmer, A., Lehmann, S., Hirtz, C., Flöel, A., Pendrill, L., 2022a. Metrological references for person ability in memory tests. Measurement: Sensors 18. https://doi.org/10.1016/j.measen.2021.100289.

Melin, J., Cano, S., Flöel, A., Göeschel, L., Pendrill, L., 2022b. The role of entropy in construct specification equations (CSE) to improve the validity of memory test: extension to word lists. Entropy 24 (7), 934. https://doi.org/10.3390/e24070934.

Melin, J., Kettunen, P., Wallin, A., Pendrill, L., 2023. Entropy-based explanations of serial position and learning effects in ordinal responses to word list tests. Acta IMEKO 12 (4), 1–5.

Moeller, J., 2015. A word on standardization in longitudinal studies: don't. Front. Psychol. 6, 1389.

Monti, L.A., Gabrieli, J.D., Reminger, S.L., Rinaldi, J.A., Wilson, R.S., Fleischman, D.A., 1996. Differential effects of aging and Alzheimer's disease on conceptual implicit and explicit memory. Neuropsychology 10, 101.

Pendrill, L.R., 2018. Assuring measurement quality in person-centred healthcare. Meas. Sci. Technol. 29 (3). https://doi.org/10.1088/1361-6501/aa9cd2.

Pettigrew, C., Soldan, A., 2019. Defining cognitive reserve and implications for cognitive aging. Curr. Neurol. Neurosci. Rep. 19 (1), 1.

Salthouse, T.A., 1996. The processing-speed theory of adult age differences in cognition. Psychol. Rev. 103, 403.

Salthouse, T.A., Schroeder, D.H., Ferrer, E., 2004. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. Dev. Psychol. 40, 813.

Shing, Y.L., Werkle-Bergner, M., Brehmer, Y., Müller, V., Li, S.-C., Lindenberger, U., 2010. Episodic memory across the lifespan: the contributions of associative and strategic components. Neurosci. Biobehav. Rev. 34, 1080–1091.

Stern, Y., Arenaza-Urquijo, E.M., Bartrés-Faz, D., Belleville, S., Cantilon, M., Chetelat, G., Ewers, M., Franzmeier, N., Kempermann, G., Kremen, W.S., others, 2020. Whitepaper: Defining and investigating cognitive reserve, brain reserve, and brain maintenance. Alzheimer's Dementia 16, 1305–1311.

Tromp, D., Dufour, A., Lithfous, S., Pebayle, T., Després, O., 2015. Episodic memory in normal aging and Alzheimer disease: Insights from imaging and behavioral studies. Ageing Res. Rev. 24, 232–262. https://doi.org/10.1016/j.arr.2015.08.006.

Tulving, E., 2002. Episodic memory: from Mind to brain. Annu. Rev. Psychol. 53, 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114.

United Nations Population Division, 2024. World Population Prospects [Data set]. https://population.un.org/wpp/. (Accessed 25 February 2025).

Wagner, M., Wolf, S., Reischies, F.M., Daerr, M., Wolfsgruber, S., Jessen, F., Popp, J., Maier, W., Hüll, M., Frölich, L., Hampel, H., Perneczky, R., Peters, O., Jahn, H., Luckhaus, C., Gertz, H.-J., Schröder, J., Pantel, J., Lewczuk, P., Kornhuber, J., Wiltfang, J., 2012. Biomarker validation of a cued recall memory deficit in prodromal Alzheimer disease. Neurology 78, 379–386. https://doi.org/10.1212/WNL.0b013e318245f447.

World Health Organization, 2021. Global Status Report on the Public Health Response to Dementia, first ed. World Health Organization, Geneva.

Weitzner, D.S., Calamia, S., 2020. Serial position effects on list learning tasks in mild cognitive impairment and Alzheimer's disease. Neuropsychology 34 (4), 467–478. https://psycnet.apa.org/doi/10.1037/neu0000620.

Wu, Y., Zhang, Y.S., Kobayashi, L.C., Mayeda, E.R., Gross, A.L., 2024. 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 8, 1661–1669.

Zeng, M., Chen, Y., Lobanov-Rostovsky, S., Liu, Y., Steptoe, A., Brunner, E.J., Liao, J., 2025. Adiposity and dementia among Chinese adults: longitudinal study in the China Health and Retirement Longitudinal Study (CHARLS). Int. J. Obes. 49 (4), 706–714.

Zhao, Y., Hu, Y., Smith, J.P., Strauss, J., Yang, G., 2014. Cohort profile: the China health and retirement longitudinal study (CHARLS). Int. J. Epidemiol. 43, 61–68. https://doi.org/10.1093/ije/dys203.

Zhao, Y., Strauss, J., Chen, X., Wang, Y., Gong, J., Meng, Q., Wang, G., Wang, H., 2020. China Health and Retirement Longitudinal Study Wave 4 User's Guide.