Longitudinal measurement properties of the Montreal Cognitive Assessment

Hao Luo, Björn Andersson, Gloria H.Y. Wong & Terry Y.S. Lum

View supplementary material

Published online: 30 Nov 2022.

Submit your article to this journal

View related articles

View Crossmark data

Routledge
Taylor & Francis Group

Longitudinal measurement properties of the Montreal Cognitive Assessment

Hao Luo [a,b], Björn Andersson [c], Gloria H.Y. Wong [a,b] and Terry Y.S. Lum [a,b]

[a]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong SAR, China; [b]Sau Po Centre on Ageing, The University of Hong Kong, Hong Kong SAR, China; [c]Centre for Educational Measurement, University of Oslo, Oslo, Norway

ABSTRACT

Introduction: The Montreal Cognitive Assessment (MoCA) has started to be used in longitudinal investigations to measure cognition trends but its measurement properties over time are largely unknown. This study aimed to examine the longitudinal measurement invariance of individual MoCA items.

Method: We used four waves of data collected between 2014 and 2017 from a cohort study on health and well-being of older adults from twelve public housing estates in Hong Kong. We identified people aged 65 years or older at baseline who answered the MoCA items across all time points and had a valid indicator of educational level. A total of 1028 participants were included. We applied confirmatory factor analysis of ordinal variables to examine measurement invariance of the Chinese (Cantonese) MoCA (version 7.0) items across four time points, stratified by educational level, where invariant items were identified by sequential model comparisons.

Results: Four items exhibited a lack of measurement invariance across the four time points in both education groups (Clock Hand, abstraction, Delayed Recall, and Orientation). The items Cube and Sentence Repetition lacked longitudinal measurement invariance only in the "some education" group and the items Clock Shape and Clock Number only in the "no education" group. However, accounting for the lack of measurement invariance did not substantially affect classification properties for major neurocognitive disorder and mild cognitive impairment.

Conclusions: Our findings support using MoCA to assess changes in cognition over time in the study population while calling for future research in other populations.

Introduction

Dementia is a leading cause of disability and dependency among the older adults (Livingston et al., 2020). Fifty-five million people worldwide currently have dementia and its prevalence is expected to triple by 2050 (World Health Organization, 2021). Effective and accurate cognitive assessment is a critical component in early detection of dementia. Assessing cognition over time is essential for early identification of cognitive decline, monitoring disease progress, and examining the effectiveness of interventions.

The Montreal Cognitive Assessment (MoCA) is one of the most widely used screening instruments for detecting suspected dementia (Nasreddine et al., 2005). The validity evidence supporting the use of MoCA scores as a measure of cognitive performance has been extensively evaluated by previous studies in a wide range of populations employing various modeling frameworks. This includes validity evidence based on content, evidence based on the relationships between the MoCA scores and external variables and evidence concerning internal structure

(Julayanont & Nasreddine, 2017). Although the test has been shown to have overall satisfactory performance in detecting mild cognitive impairment and dementia, substantial variability in scale performance (in terms of sensitivity, specificity, and psychometric properties) has been observed across populations from different geographical locations and with different educational levels, resulting in population-specific cutoff values. This variability is partially attributable to item bias, the situation where the response to an item is dependent on item characteristics that are unrelated to the latent construct cognition (Balsis et al., 2018). For example, a recent Hong Kong study in a low-education older population found that the properties of the Cube, Clock Number, and Clock Hand items of the MoCA varied with educational level, suggesting that more sophisticated modeling is needed to account for the effect of education on individual items (Luo et al., 2020). Other putative characteristics that may be associated with item bias include age, culture, visual and hearing impairment, and previous exposure to an item.

In addition to the intended one-off usage of the MoCA for cognitive screening (Nasreddine et al., 2005), the scale has been adopted in a growing number of longitudinal investigations to track changes in cognition over time (Costa et al., 2014; Freitas et al., 2013; Krishnan et al., 2017). For example, a prospective study of non-demented older adults used the MoCA to measure change in global cognitive function over a period of two years and found that cognitive decline was associated with physical frailty (Chen et al., 2018). A two-year longitudinal study in Hong Kong used the MoCA and found that a mentally active lifestyle and a structured cognitive program were associated with better cognition (Tang et al., 2020). The MoCA was also used to assess cognition and cognitive decline in people with Parkinson's disease and Lewy body dementia in clinical practice (Biundo et al., 2016; Lessig et al., 2012). Changes in MoCA scores have also been used in randomized controlled trials as quick outcome measures to evaluate the effectiveness of an intervention (Apóstolo et al., 2014). However, for any of the subsequent inferences regarding changes in cognition to be valid, longitudinal measurement invariance of the MoCA needs to be established first.

For longitudinal measurement invariance to hold, the repeated measurement of the same items must be comparable across time in terms of discriminatory power and difficulty (Y. Liu et al., 2017). Drawing valid conclusions regarding changes in the latent construct over time is possible when longitudinal measurement invariance is fulfilled. The assumption may, however, not always hold since the same item can measure a different construct as people age or have its measurement properties change due to a possible practice or learning effect (Wong et al., 2018).

Measurement invariance of any cognitive assessment instrument, across populations and over time, needs to be evaluated using techniques such as confirmatory factor analysis or item response theory. A few studies have investigated measurement invariance of cognitive scales commonly used in older adults, including the Functional Assessment Questionnaire, the Neuropsychiatric Inventory Questionnaire, the NIH Toolbox Cognition Battery, the Brief Assessment of Impaired Cognition Questionnaire, the Mini-mental State Examination, and the MoCA (Li et al., 2022; Luo et al., 2020; Ma et al., 2021; Prieto et al., 2011; Sayegh & Knight, 2014). However, all previous investigations were limited to cross-sectional examinations of measurement invariance across demographic characteristics such as sex, age, educational level, and ethnicity. Despite the extensive adoption of cognitive scales in both general and clinical populations, no study has investigated longitudinal measurement invariance of the MoCA. In this study, we aimed to investigate the measurement invariance of individual MoCA items across four time points over a period of three years in a sample of older persons in Hong Kong.

Methods

Data source and participants

Longitudinal data collected at four time points between 2014 and 2017 were available from a prospective cohort study on the health and well-being of Cantonese-speaking older persons in Hong Kong. The study was conducted in 12 public housing estates. In each housing estate, participants were sampled using stratified random sampling by three age groups: 65 to 74 years, 75 to 84 years, and 85 years and older, with target sample sizes in each estate of 50, 60, and 70, respectively (T. Liu et al., 2018). The oldest age group was purposely over-sampled to cover the full spectrum of cognitive and/or functional abilities. Information on demographic characteristics, physical and mental health status, cognitive ability and life style were collected by trained interviewers. Consents were obtained from all participants and data were de-identified before submission to the data analyst.

A total of 2081 participants were included in the study at baseline. Older people with a self-reported dementia diagnosis were not assessed with MoCA and hence excluded in this current analysis. The number of participants who were assessed with the MoCA were 2078 in 2014, 1597 in 2015, 1392 in 2016, and 1294 in 2017. In the analysis of longitudinal measurement invariance, we included the 1028 participants that answered the MoCA items across all four time points and had a valid indicator of educational level. The study was approved by the Review Board of the Human Research Ethics Committee for Non-Clinical Faculties at The University of Hong Kong with reference number EA050814.

Measures

We used the Chinese (Cantonese) MoCA (version 7.0) that covers eight cognitive domains with a total of 14 items (Nasreddine et al., 2005). These items were Trail Making, Cube, Clock Shape, Clock Number, Clock Hand, Naming, Digit Span, Attention, Serial Subtraction, Sentence Repetition, Verbal Fluency, Abstraction, Delayed Recall, and Orientation. This version of the MoCA has been validated in Hong Kong and is extensively used by

both researchers and service providers (Chu et al., 2015). Items were scored as integers starting from 0 with maximum scores between 1 and 5. Item scores were treated as ordinal in the statistical analysis. The MoCA is typically scored by summing the item scores, resulting in sum scores that range from 0 to 30. We collapsed the scores 0, 1, 2 into a single category for the item Orientation since our preliminary analysis showed that there were too few observations in the lowest two categories to estimate the parameters of the statistical models employed. This low frequency was expected as this is a non-clinical sample.

Educational level was assessed using the following categories: no formal education (has not attended school of any kind), primary school, junior school, high school, postsecondary (non-degree), and college and above. Since only a small proportion of participants had an educational level of junior school or above, we recoded educational level into a binary variable indicating whether a participant had a primary school education or higher. We hereafter refer to the two groups as "no education" and "some education" groups.

Participants were also asked whether they had a diagnosis of a range of chronic conditions. Relevant neuropsychiatric conditions included stroke, Parkinson's disease, depression, schizophrenia, anxiety disorder, and bipolar disorder.

Statistical analysis

We fitted a series of confirmatory factor analysis models to the ordinal observed variables (Olsson, 1979; Samejima, 1969) to evaluate longitudinal measurement invariance across four time points. Evaluations were carried out separately in each education group. This is because a previous study of the same population found that characteristics of the MoCA items differed between people with and without education at baseline (Luo et al., 2020). To avoid conflating cross-sectional and longitudinal measurement invariance, the analyses were therefore conducted by education group.

We used a multidimensional model (Figure 1) to examine longitudinal measurement invariance by testing the equality of the item parameters across measurement time points (Y. Liu et al., 2017). For each education group, cognitive performances at each time point were specified as separate factors that correlated with each other, where each factor was measured by 14 observed items. The repeated measurements with the same items had residual correlations. We utilized the theta parametrization with a standardized latent variable when specifying the models (Wu & Estabrook, 2016). The theta parametrization treats residual variances as parameters, which allows testing equality of residual variances across time points. Item parameters of interest included factor loadings, thresholds, intercepts, and residual variances.

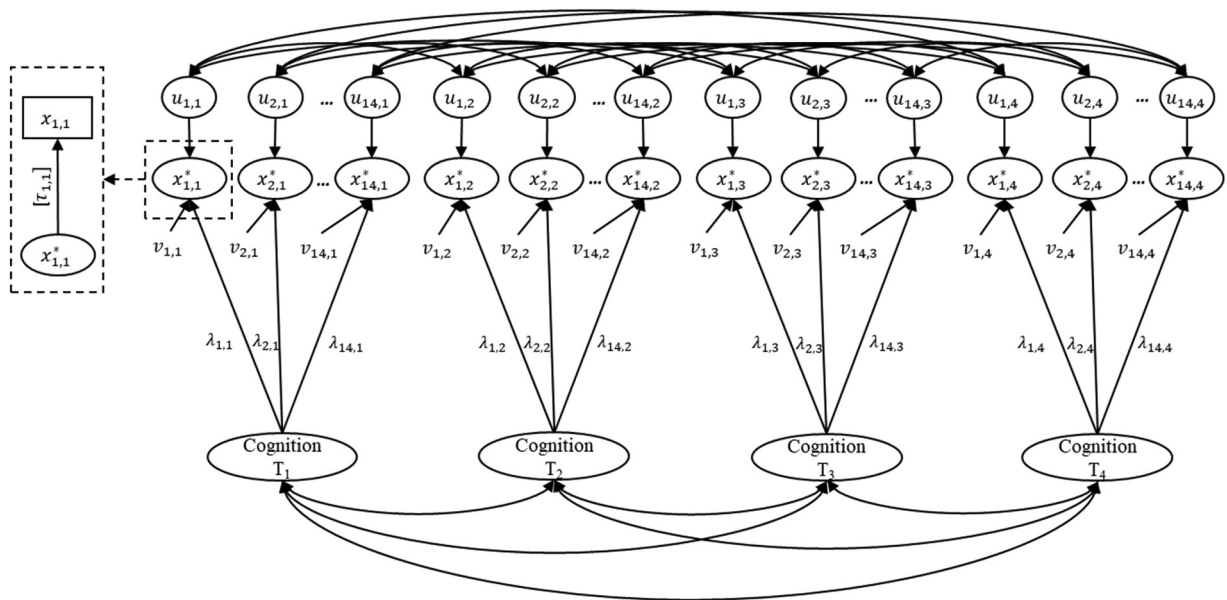Figure 1. The longitudinal model for the 14 items used across the four time points in each education group, where $x_{j,t}$ and $x^{*}_{j,t}$ denote the observed variable and its continuous latent response for item j at time point t in each group, respectively; $\tau_{j,k}$, $\lambda_{j,t}$, and $v_{j,t}$ denote the threshold, factor loading and intercept, respectively; and $u_{j,t}$ denotes the residual.

Analytical procedure

We adopted the procedures for evaluation of measurement invariance proposed in a previous work (Wu & Estabrook, 2016). This framework outlines models with different levels of restrictions corresponding to different levels of invariance, including (1) configural models, (2) models assuming threshold, loading and intercept invariance of all items, and (3) models assuming full invariance. We also considered both invariance across all items and invariance for only a subset of the items, so called partial invariance (Byrne et al., 1989). In situations where complete or partial threshold, loading and intercept invariance is established, any differences in both the means and variances of the factors at different time points can be interpreted as changes in the underlying cognitive performance over time in the study population. In the desirable situation where full invariance can be established, changes in sum scores of the MoCA over time can be interpreted as changes in the underlying construct (Y. Liu et al., 2017). The sequence of model comparisons is illustrated in Figure 2. A summary of the parameter restrictions for different models is provided in Table S1 of the supplementary material. Details of the analytical protocol are described below.

We first estimate a configural model without restrictions in the item parameters across time and evaluate the model fit. This configural model tests if the same general factorial structure holds across time points. If the configural model fits well, we evaluate if all the thresholds, factor loadings, and intercepts are equal across time. Such invariance can be established if the model fit does not differ from the configural model according to chi-square difference tests (Y. Liu et al., 2017). If threshold, loading and intercept longitudinal measurement invariance is rejected, we evaluate partial longitudinal measurement invariance with an approach similar to that in Fischer et al. (2018). The following series of steps are taken:

Step 1. We define a restricted baseline model with threshold, loading and intercept invariance imposed for all items across the four time points.

Step 2. We remove the threshold and factor loading parameter restrictions across time for each item (one at a time) and compare the model fit to the restricted baseline model in step a) via a chi-square difference test using significance level $\alpha = 0.05$. This will be repeated for each of the 14 MoCA items.

Step 3. The parameters of the items that exhibited noninvariance in step 2 are freed and the partial threshold, loading and intercept invariance model is estimated. This model is then compared to the configural model with a chi-square difference test and evaluated for model fit.

If the threshold, loading and intercept invariance model holds, we further evaluate the full invariance model, in which all factor loadings, thresholds, intercepts, and residual variances are constrained to be equal over time. If full longitudinal measurement invariance is rejected, we evaluate partial invariance following similar steps as described in evaluating partial threshold, loading and intercept invariance.

Model estimation and evaluation

Estimation was done with the limited information estimator DWLS in combination with polychoric correlations (Browne, 1984) using the R package *lavaan* (Rosseel, 2012). Utilizing full-information maximum likelihood was not possible due to the high latent dimensionality of the specified longitudinal model. To
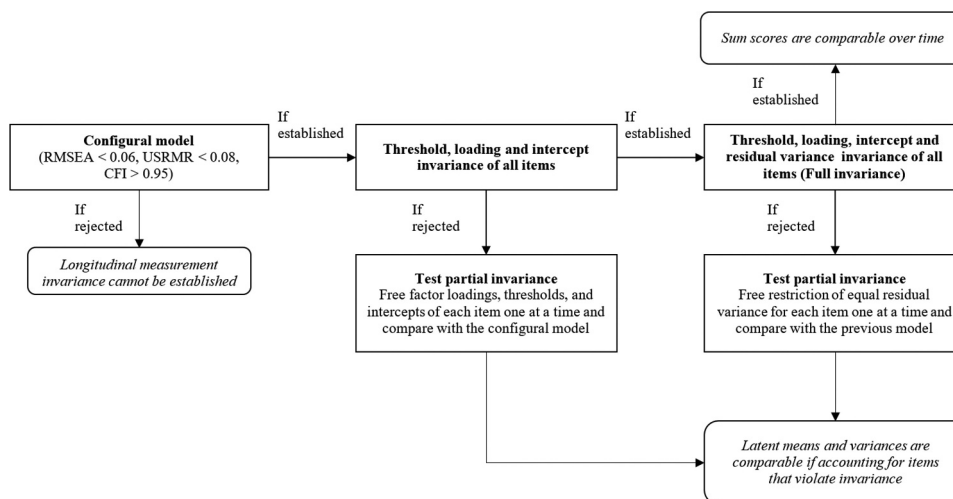

Figure 2. A flowchart of the procedure for the evaluation of measurement invariance.

specify the models corresponding to a) configural, b) threshold, loading and intercept invariance, and c) full invariance, we utilized the R package *semTools* (Jorgensen et al., 2021). Model fit was evaluated with the Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA) and the Unbiased Standardized Root Mean square Residual (USRMR; Shi et al., 2018). Observed values of the CFI higher than 0.95, RMSEA less than 0.06 and values of the USRMR less than 0.08 indicated acceptable model-data fit (Hu & Bentler, 1999). The R code used is available on the website https://github.com/bjoernhandersson/Longitudinal-MoCA.

Implications of potential violation of measurement invariance

The implications of the lack of measurement invariance were explored by comparing trajectories of cognitive performance and implied diagnostic categories resulting from sum scores (which assumes measurement invariance) and factor scores (which incorporates and adjusts for violations to measurement invariance). To assess the classification properties, we used the 2nd and 16th percentile for major and mild neurocognitive disorders based on the Diagnostic and Statistical Manual of Mental Disorders-Fifth edition (American Psychiatric Association, 2013), respectively.

We also used the 7th percentile for mild cognitive impairment based on Petersen's revised diagnostic criteria (Petersen et al., 2009). These cutoffs were also adopted by previous local studies (Luo et al., 2020; Wong et al., 2015). We also compared the estimated factor scores of each participant at each time point with different levels of invariance using scatterplots.

Results

Sample characteristics

Out of the 1028 participants, 46.50% had no formal education. The basic demographic characteristics, clinical characteristics, and the mean MoCA sum scores of participants with and without education are shown in Table 1. The mean age of the sample at baseline was 78.80 and 58.75% were female. The sum scores of MoCA in the "some education" group were considerably higher than in the "no education" group at all time points. The mean scores remained stable for both groups and no clear trend can be concluded.

Longitudinal measurement invariance of the MoCA over four time points

The configural model, without restrictions on the item parameters except those needed for identification, showed acceptable model fit in each of the two groups (Table 2). We thus proceeded with evaluating threshold, intercept, and loading longitudinal measurement invariance. The chi-square difference tests showed that the threshold, loading and intercept invariance model fit the data much worse than the configural model, suggesting that longitudinal measurement invariance did not hold for either of the "no education" ($\Delta\chi^2 = 174.71$, df = 78, p < 0.0001) and "some education" groups ($\Delta\chi^2 = 145.22$, df = 78, p < 0.0001). However, when evaluating the models in terms of the absolute fit, all models fitted the data well (Table 2).

Since the chi-square difference tests rejected the threshold, loading, and intercept invariance models, we proceeded to evaluate partial longitudinal

Table 1. Sample characteristics by educational level for the 1028 included participants.

	Some education (N = 550)	No education (N = 478)	Total sample (N = 1028)
Female, n (%)	269 (48.9)	335 (70.1)	604 (58.8)
Age in 2014, M(SD)	76.8 (7.8)	81.1 (7.0)	78.80 (7.7)
Educational level			
Primary school, n (%)	306 (55.6)	-	306 (29.8)
Junior school, n (%)	135 (24.5)	-	135 (13.1)
High school, n (%)	90 (16.4)	-	90 (8.8)
Postsecondary school, n (%)	8 (1.5)	-	8 (1.1)
College and above, n (%)	11 (2.0)	-	11 (4.2)
Neuropsychiatric conditions			
Stroke, n (%)	43 (7.8)	23 (4.8)	66 (6.4)
Parkinson's disease, n (%)	5 (0.9)	6 (1.3)	11 (1.1)
Depression, n (%)	20 (3.6)	4 (0.8)	24 (2.3)
Schizophrenia, n (%)	2 (0.4)	2 (0.4)	4 (0.4)
Anxiety disorder, n (%)	5 (0.9)	6 (1.3)	11 (1.1)
Bipolar disorder, n (%)	1 (0.2)	1 (0.2)	2 (0.2)
MoCA sum scores			
MoCA sum score 2014, M(SD)	22.2 (4.6)	17.5 (5.2)	20.38 (5.4)
MoCA sum score 2015, M(SD)	22.6 (4.5)	18.4 (5.4)	21.00 (5.3)
MoCA sum score 2016, M(SD)	22.4 (4.7)	17.5 (5.2)	20.46 (5.4)
MoCA sum score 2017, M(SD)	22.4 (4.8)	18.2 (5.7)	20.77 (5.5)

Table 2. Model fit statistics for longitudinal models in two education groups.

Group	Model	χ^2 (df)	CFI[a]	RMSEA[b]	SRMR[c]	USRMR[d]
Education	Configural	1857.293 (1394)	0.989	0.025	0.077	0.043
	Threshold, loading and intercept	2075.175 (1472)	0.986	0.027	0.077	0.043
	Partial threshold, loading and intercept	1893.180 (1427)	0.989	0.024	0.077	0.043
No education	Configural	1960.351 (1394)	0.986	0.029	0.080	0.046
	Threshold, loading and intercept	2177.300 (1472)	0.984	0.032	0.080	0.047
	Partial threshold, loading and intercept	2006.800 (1430)	0.987	0.029	0.080	0.046

[a] CFI = Comparative Fit Index; [b] RMSEA = Root Mean Square Error of Approximation; [c] SRMR = Standardized Root Mean square Residual; [d] USRMR = Unbiased Standardized Root Mean square Residual

measurement invariance. Invariant items were identified via chi-square difference tests between the restricted baseline model and models that freed the parameters of each item separately (in total 14 models). This procedure resulted in identifying eight items in the "some education" group (Trail Making, Clock Shape, Clock Number, Naming, Digit Span, Attention, Serial Subtraction, Verbal Fluency) and eight items in the "no education" group (Trail Making, Cube, Naming, Digit Span, Attention, Serial Subtraction, Sentence Repetition, Verbal Fluency) that exhibited measurement invariance across all time points. We summarize the model fit statistics for the configural models, threshold, loading and intercept invariance models, and the partial threshold, loading, and intercept invariance models with eight restricted items in Table 2. The two partial invariance models were not significantly different to the configural models, with $\Delta\chi^2 = 25.421$ (df = 33, p = 0.8243) in the "some education" group and $\Delta\chi^2 = 41.068$ (df = 36, p = 0.2581) in the "no education" group. Meanwhile, the threshold, loading, and intercept invariance model displayed noticeably worse absolute and relative fit compared to the less restrictive models, although the model fit, like all models considered, is still considered acceptable by the pre-specified criteria.

Four items exhibited a lack of measurement invariance across the four time points in both education groups (Clock Hand, Abstraction, Delayed Recall, and Orientation), while the items Cube and Sentence Repetition lacked longitudinal measurement invariance only in the "some education" group and the items Clock Shape and Clock Number lacked longitudinal measurement invariance only in the "no education" group. The factor loading and threshold estimates for the items without measurement invariance are displayed in Figure 3. There were generally small differences in the factor loading estimates without an obvious general trend over time. The items Clock Shape, Clock Number and Clock Hand displayed a slight decreasing trend in the thresholds in the "no education" group, effectively meaning that these items became easier (on average), conditional on the cognitive performance, at the later time points.

Trend estimation

The final longitudinal model identified eight items with measurement invariance across the time points in each group, which enabled comparisons of the mean and variance of latent cognition over time. Note that the mean and variance were fixed to 0 and 1 at the first time point. The mean cognition estimates and standard errors for the "some education" and "no education" groups, respectively, were 0.175 (0.076) and 0.105 (0.059) in 2015, −0.053 (0.068) and −0.023 (0.058) in 2016, and 0.020 (0.073) and 0.042 (0.058) in 2017. Hence, there did not exist an increasing or decreasing trend in cognitive performance over time for either of the two groups.

Consequences of the lack of measurement invariance

We evaluated the effects of accounting for lack of measurement invariance versus not accounting for lack of measurement invariance by estimating individual factor scores and sum scores and subsequently applying a classification decision at each of the four time points. At each time point, we classified individuals into the 2nd, 7th, and 16th percentiles compared to the empirical distributions at baseline with either method. The resulting classification consistency between sum scores and factor scores are presented in Table 3, showing a high agreement between the methods and that this agreement did not substantially change over time. The lowest classification consistency for any categorization and time point was with respect to classification in the 16th percentile for the "some education" group at time point 4, where the classification consistency was 93.0% between sum scores and factor scores. This meant that making classification decisions based on the longitudinal model which accounted for measurement invariance did not substantially change the conclusions made compared to using the sum scores. Scatterplots of the estimated factor scores of each participant at each time point with the full invariance and the partial invariance models showed that the correlations between the factor scores from different models were extremely high (Supplementary Figures S1 and S2).
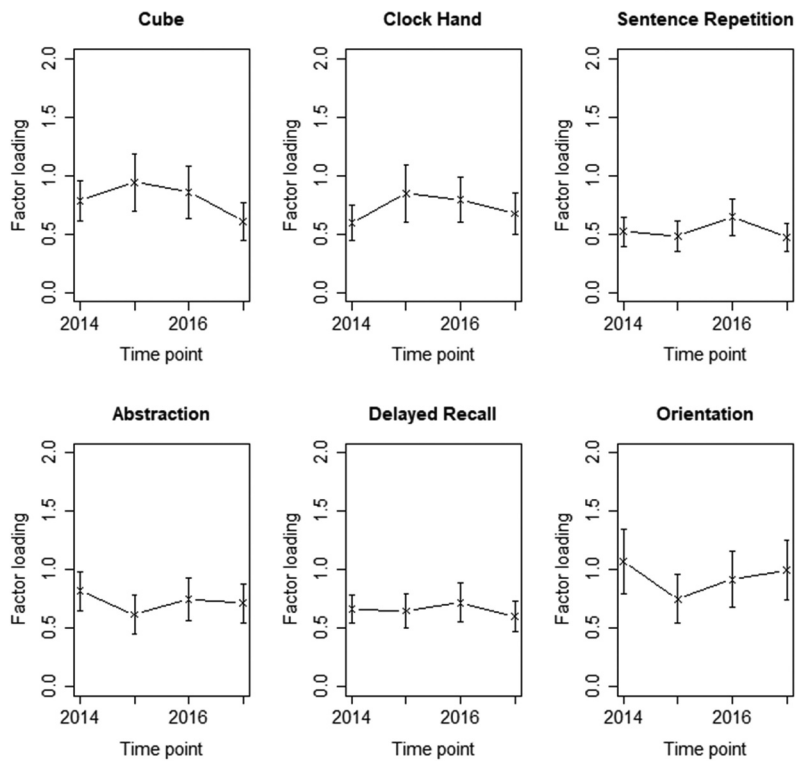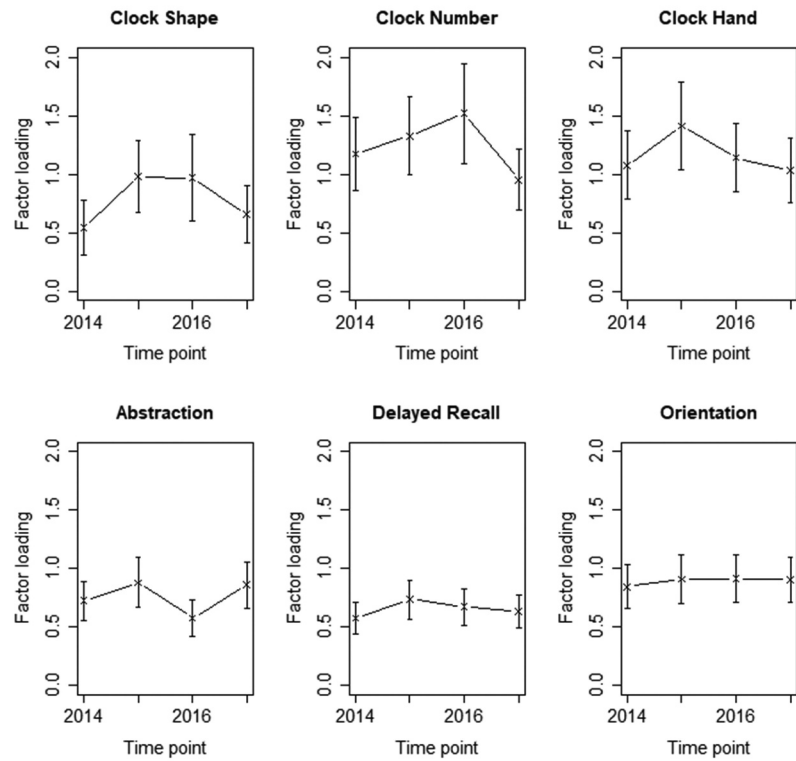
Figure 3. Factor loading and threshold estimates at each time point for each Montreal Cognitive Assessment item that lacked measurement invariance across time points.
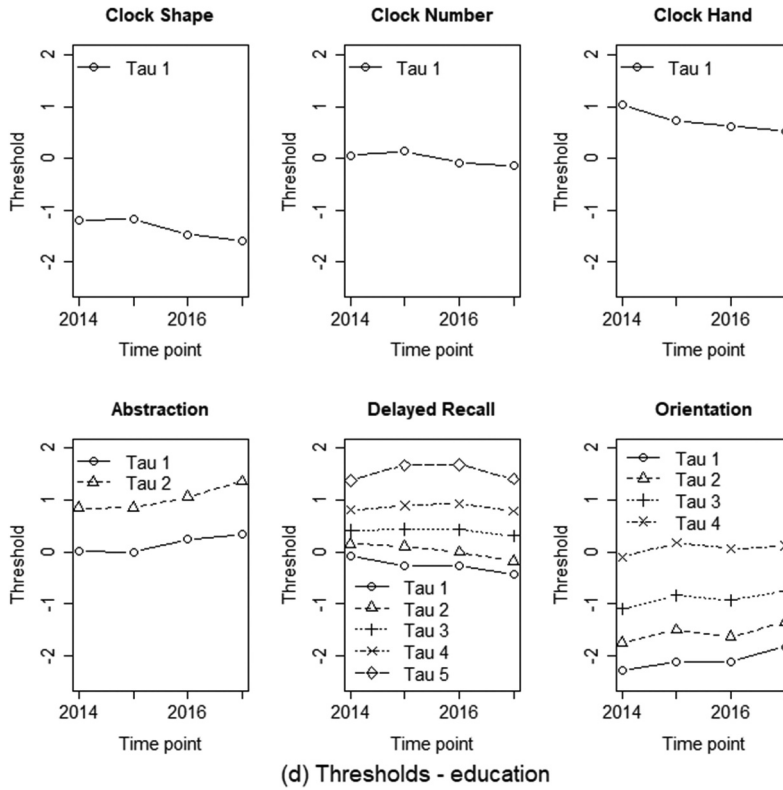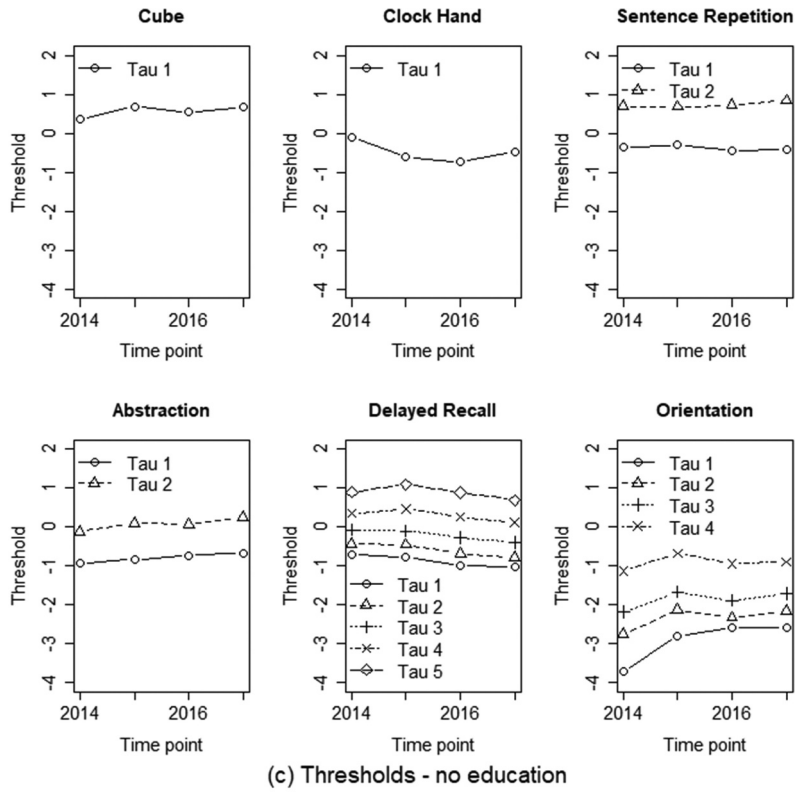
(c) Thresholds - no education

(d) Thresholds - education

Figure 3. (Continued).

Table 3. Classification decisions in % into the lowest 2nd, 7th, and 16th percentiles for sum scores (Sum) and maximum likelihood factor scores (Factor) (N = 775). Note. Factor(yes) and Sum(yes) indicates classification into the 2nd, 7th, or 16th percentile while Factor(no) and Sum(no) indicates not classifying into these percentiles.

Group	Percentile	Year	Sum(no) & Factor (no)	Sum(no) & Factor (yes)	Sum(yes) & Factor (no)	Sum(yes) & Factor (yes)	Classification Consistency
Education	2nd percentile	2014	97.9	0.4	0.0	1.7	99.6
		2015	96.8	1.3	0.0	1.9	98.7
		2016	97.9	0.0	0.4	1.7	98.7
		2017	96.8	0.8	0.0	2.3	99.1
	7th percentile	2014	92.8	1.9	0.0	5.3	98.1
		2015	93.2	2.3	0.0	4.4	97.6
		2016	94.9	1.1	0.4	3.6	98.5
		2017	93.7	1.7	0.4	4.2	97.9
	16th percentile	2014	83.5	3.4	0.4	12.7	96.1
		2015	83.3	6.5	0.0	10.1	93.4
		2016	83.8	5.3	0.4	10.5	96.2
		2017	81.4	6.5	0.4	11.6	93.0
No education	2nd percentile	2014	98.0	1.0	0.0	1.0	99.0
		2015	98.7	0.3	0.0	1.0	99.7
		2016	97.7	1.0	0.0	1.3	99.0
		2017	97.3	1.0	0.0	1.7	99.0
	7th percentile	2014	92.7	0.7	0.3	6.3	99.0
		2015	94.4	0.3	0.3	5.0	99.4
		2016	92.4	1.0	2.0	4.7	97.1
		2017	92.4	0.7	0.7	6.3	98.7
	16th percentile	2014	82.7	3.0	1.3	13.0	95.7
		2015	87.7	1.3	1.0	10.0	97.7
		2016	84.4	1.3	2.7	11.6	96
		2017	85.5	2.0	2.3	10.6	96.1

Discussion

This study employed four-wave longitudinal data of 1028 older adults living in public housing estates in Hong Kong to examine longitudinal measurement invariance of individual items of the MoCA. To the best of our knowledge, this is the first study that systematically investigated the longitudinal measurement properties of the MoCA. Using confirmatory factor analysis, we found evidence of violations to measurement invariance by time point in six items in each education group. However, the differences in parameter estimates of the confirmatory factor analysis model over time, although showing statistical significance, were marginal and had limited impact on the subsequent inference in terms of estimation of factor means and variances and for classification decisions based on the individual MoCA scores. This supports the usage of the MoCA for assessing changes in cognition over time in the study population. However, it is important to note that as information on reliable clinical diagnosis of dementia was not available, this research focused on the examination of validity evidence based on internal structure and did not investigate evidence based on relationships to external criterion variables. Future research is warranted to further examine the impact of not accounting for longitudinal measurement invariance on classification decisions by comparing the sum score and factor score implied diagnostic categories with clinical diagnosis.

Assessing longitudinal measurement invariance is of paramount importance when inferring general and individual trends across time in order for the interpretations of the observed trends in, for example, mean scale scores to be valid. Practice effects and learning as a result of the repeated use of the same scale over time can impact the resulting scores in unexpected ways that need to be accounted for. Although potential learning effects in serial MoCA assessments have been suspected, and versions for repeated measure of the MoCA have been made available to reduce such effects in previous studies (Jassal & Farragher, 2020; Siciliano et al., 2019), whether learning effects exist with the MoCA in the first place has not been rigorously investigated using sufficiently powered samples. Our study is one of the first that has conducted a detailed evaluation of the longitudinal measurement properties of a cognitive screening scale for mild and severe cognitive impairment using modern measurement techniques. The approach we employed addresses the potential lack of measurement invariance at the level of the items, which enables the identification of problematic items for consideration of revision in the application of the scale in a longitudinal setting. We found evidence on violations of longitudinal measurement invariance in some but not

all items, supporting the generation of hypotheses on different learning effects across cognitive domains. This warrants future research to examine domain-specific measurement invariance.

We found that four items exhibited noninvariance in both of the studied groups: Clock Hand, Abstraction, Delayed Recall, and Orientation. The differences in item properties across time for these items mainly concerned the threshold parameters, which implies a differing difficulty level across time for these items. Two trends were common to both groups: a decrease in the threshold parameters for the item Delayed Recall and an increase in the threshold parameters for the items Abstraction and Orientation. While this does influence the properties of the scale scores, the magnitudes of the differences across time were small and the differences furthermore went in opposite directions for different items. Hence, the relative difficulty of the MoCA over time did not change substantially for either of the two groups.

While our study identified violations to longitudinal measurement invariance for six out of 14 items on the MoCA scale in each of the education groups, the subsequent analysis of the implications of the lack of invariance revealed that the items which exhibited a lack of invariance did not substantially affect the measurement properties of the scale as a whole across the time points. Furthermore, applying decision rules based on diagnostic criteria with either the sum scores (which assume measurement invariance) or the model-based factor scores (which account for measurement invariance) yielded highly similar classifications. This implies that, for the MoCA scale, the statistically significant differences in the item properties over time did not result in practically important differences for the scale as a whole. This supports the validity of MoCA scores as a measure of cognitive performance, not only for one-off cognitive screening, but also for measuring changes in cognition longitudinally at the individual level. It is also important to note that our findings are limited to a non-clinical sample of participants who did not have a formal diagnosis of dementia. For people with dementia diagnosis, changes in item functioning associated with different cognitive domains, if any, would arguably provide important clinical information for advancing the understanding of dementia.

From a methodological perspective, this study can form the basis for future analyses of longitudinal measurement invariance. However, a potential roadblock to assessing longitudinal measurement invariance in practice is the complex model specification needed and the numerous sequential model comparisons that are required. A more streamlined approach to assessing longitudinal measurement invariance in terms of the method used and the software implementation thereof is desirable from a practical perspective.

In this study we employed confirmatory factor analysis (CFA) with ordinal data in a multidimensional measurement model for the MoCA item scores across time. Some prior studies have instead used the framework of item response theory (IRT) to evaluate the measurement properties of the MoCA (Luo et al., 2020; Tsai et al., 2012). We utilized CFA instead of IRT because the high dimensionality of the longitudinal model made the typical estimation procedure used in IRT, full-information maximum likelihood, practically impossible to use in this setting. Such observations of the difficulty to apply complex longitudinal IRT models have previously been done in the literature (Paek et al., 2016). We note also that CFA with ordinal indicators is equivalent to a particular IRT graded response model (Takane & de Leeuw, 1987) and, as such, our analysis can be viewed as a type of IRT analysis. However, since our estimation approach only allowed for a specific type of model, utilizing a more flexible modeling framework is still desirable to potentially improve the model-data fit. The application of longitudinal IRT in complex settings remains a difficult challenge that requires further methodological development to solve.

While the MoCA has so far been used primarily as a screening tool for specific cognitive disorders at a single point in time, we believe that our results point to the potential in utilizing MoCA as a tool for assessing individual trends and, thus, identification of potential individual decline over time. By simultaneously analyzing MoCA item scores across multiple time points it is possible to improve individual measurement precision and utilize the added information of the repeated measurements to predict decline more effectively. We encourage future studies to evaluate the diagnostic accuracy of such usage of the MoCA scores, which can guide clinical practice beyond what was possible in the present study.

Several limitations of this study are worth noting. First, our sample, although large, were randomly selected from a relatively homogeneous population residing in public housing estates in Hong Kong. Since public housing estates in Hong Kong are only available for low-income households, findings from this study are limited to older city-dwellers in Hong Kong with a lower educational level and socioeconomic status compared to the general older population (Lum et al., 2016). In the absence of an external

sample, one potential method to assess the robustness of our findings is to further split the current sample. However, the no education group already had a sample size of 478. Further reducing the sample size may compromise the power of the analysis, especially when the model to be estimated is high-dimensional. Future research is needed to examine the robustness of our findings. Second, very few participants in our sample had an education at middle school or above. Our measure of educational level was hence collapsed into a dichotomous variable indicating any formal education at all or some formal education. The potential effect of high educational attainment on measurement invariance of the MoCA could not be investigated. Third, given that the study sample comprised old-age participants who suffered from various levels of functional impairments, only half of the initial sample completed the MoCA at all four time points. The results are hence skewed toward more healthy and robust participants. Consequently, findings from this study can only be generalized to individuals who answered the MoCA items across all four time points and had a valid indicator of educational level. This may also partially explain why no clear trend in cognitive changes can be concluded. In addition, the analysis for comparing classification decisions with and without accounting for measurement invariance was based on an even smaller sample of 775 participants. This is because although it is possible to handle missing values when estimating the factor scores from the confirmatory factor analysis model, conventional sum scores need to be calculated using participants that completed all items. Fourth, we only evaluated measurement invariance at the level of the items and not at the level of individual item parameters. Hence, we did not establish which type of violation to measurement invariance the identified items exhibited. Finally, as mentioned, we did not investigate criterion validity against diagnostic accuracy as indicated by self-reported dementia diagnosis as participants who had a diagnosis of dementia were exempted from further cognitive assessment. Our baseline data showed that only 0.5% (N = 10) of the participants reported having a diagnosis of dementia, suggesting that most people living with dementia in public housing estates did not receive any formal diagnosis. This suggests that our study sample can be viewed as representative of the general older population living in public housing estates in Hong Kong, covering the whole cognitive spectrum.

Notwithstanding these limitations, our study provides important first evidence on longitudinal measurement invariance of the MoCA. Our findings support the longitudinal usage of the tool for assessing cognitive changes in a low-education older population in Hong Kong. Future studies are needed to establish longitudinal measurement invariance of the MoCA in other populations. It is also of interest to explore whether the same items that violated measurement invariance, although with very marginal effect, can be observed in other samples.

ORCID

Hao Luo http://orcid.org/0000-0003-4261-3414
Björn Andersson http://orcid.org/0000-0002-9007-2440
Gloria H.Y. Wong http://orcid.org/0000-0002-1331-942X
Terry Y.S. Lum http://orcid.org/0000-0003-1196-5345

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Apóstolo, J. L. A., Cardoso, D. F. B., Rosa, A. I., & Paúl, C. (2014). The effect of cognitive stimulation on nursing home elders: A randomized controlled trial. *Journal of Nursing Scholarship*, *46*(3), 157–166. https://doi.org/10.1111/jnu.12072

Balsis, S., Choudhury, T. K., Geraci, L., Benge, J. F., & Patrick, C. J. (2018). Alzheimer's disease assessment: A review and illustrations focusing on item response theory techniques. *Assessment*, *25*(3), 360–373. https://doi.org/10.1177/1073191117745125

Biundo, R., Weis, L., Bostantjopoulou, S., Stefanova, E., Falup-Pecurariu, C., Kramberger, M. G., Geurtsen, G. J., Antonini, A., Weintraub, D., & Aarsland, D. (2016). MMSE and MoCA in Parkinson's disease and dementia with Lewy bodies: A multicenter 1-year follow-up study. *Journal of Neural Transmission*, *123*(4), 431–438. https://doi.org/10.1007/s00702-016-1517-6

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83. https://doi.org/10.1111/j.2044-8317.1984.tb00789.x

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chen, S., Honda, T., Narazaki, K., Chen, T., Kishimoto, H., Haeuchi, Y., & Kumagai, S. (2018). Physical frailty is

associated with longitudinal decline in global cognitive function in non-demented older adults: A prospective study. *The Journal of Nutrition, Health & Aging*, 22(1), 82–88. https://doi.org/10.1007/s12603-017-0924-1

Chu, L.-W., Ng, K. H., Law, A. C., Lee, A. M., & Kwan, F. (2015). Validity of the Cantonese Chinese Montreal cognitive assessment in Southern Chinese. *Geriatrics & Gerontology International*, 15(1), 96–103. https://doi.org/10.1111/ggi.12237

Costa, A. S., Reich, A., Fimm, B., Ketteler, S. T., Schulz, J. B., & Reetz, K. (2014). Evidence of the sensitivity of the MoCA alternate forms in monitoring cognitive change in early Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 37(1–2), 95–103. https://doi.org/10.1159/000351864

Fischer, F., Gibbons, C., Coste, J., Valderas, J. M., Rose, M., & Leplège, A. (2018). Measurement invariance and general population reference values of the PROMIS profile 29 in the UK, France, and Germany. *Quality of Life Research*, 27(4), 999–1014. https://doi.org/10.1007/s11136-018-1785-8

Freitas, S., Simões, M. R., Alves, L., & Santana, I. (2013). Montreal cognitive assessment: Validation study for mild cognitive impairment and Alzheimer disease. *Alzheimer Disease and Associated Disorders*, 27(1), 37–43. https://doi.org/10.1097/WAD.0b013e3182420bfe

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Jassal, S. V., & Farragher, J. F. (2020). MoCA: Turn your mind to it. *Journal of the American Society of Nephrology*, 31(4), 672–673. https://doi.org/10.1681/ASN.2020020173

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *SemTools: Useful tools for structural equation modeling* R package version 0.5-5. https://CRAN.R-project.org/package=semTools

Julayanont, P., & Nasreddine, Z. S. (2017). Montreal Cognitive Assessment (MoCA): Concept and clinical review. In L. A.j (Ed.), *Cognitive screening instruments* (pp. 139–195). Springer. https://doi.org/10.1007/978-3-319-44775-9_7

Krishnan, K., Rossetti, H., Hynan, L. S., Carter, K., Falkowski, J., Lacritz, L., Cullum, C. M., & Weiner, M. (2017). Changes in Montreal cognitive assessment scores over time. *Assessment*, 24(6), 772–777. https://doi.org/10.1177/1073191116654217

Lessig, S., Nie, D., Xu, R., & Corey-Bloom, J. (2012). Changes on brief cognitive instruments over time in Parkinson's disease. *Movement Disorders*, 27(9), 1125–1128. https://doi.org/10.1002/mds.25070

Li, S., Cui, G., Jørgensen, K., Cheng, Z., Li, Z., & Xu, H. (2022). Psychometric properties and measurement invariance of the Chinese version of the brief assessment of impaired cognition questionnaire in community-dwelling older adults. *Frontiers in Public Health*, 10, 908827. https://doi.org/10.3389/fpubh.2022.908827

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, 22(3), 486–506. https://doi.org/10.1037/met0000075

Liu, T., Wong, G. H., Luo, H., Tang, J. Y., Xu, J., Choy, J. C., & Lum, T. Y. (2018). Everyday cognitive functioning and global cognitive performance are differentially associated with physical frailty and chronological age in older Chinese men and women. *Aging & Mental Health*, 22(8), 936–941. https://doi.org/10.1080/13607863.2017.1320700

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., . . . Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet (London, England)*, 396(10248), 413–446. https://doi.org/10.1016/S0140-6736(20)30367-6

Lum, T. Y., Lou, V. W., Chen, Y., Wong, G. H., Luo, H., & Tong, T. L. (2016). Neighborhood support and aging-in-place preference among low-income elderly Chinese city-dwellers. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 71(1), 98–105. https://doi.org/10.1093/geronb/gbu154

Luo, H., Andersson, B., Tang, J. Y., & Wong, G. H. (2020). Applying item response theory analysis to the Montreal cognitive assessment in a low-education older population. *Assessment*, 27(7), 1416–1428. https://doi.org/10.1177/1073191118821733

Ma, Y., Carlsson, C. M., Wahoske, M. L., Blazel, H. M., Chappell, R. J., Johnson, S. C., Asthana, S., & Gleason, C. E. (2021). Latent factor structure and measurement invariance of the NIH toolbox cognition battery in an Alzheimer's Disease research sample. *Journal of the International Neuropsychological Society*, 27(5), 412–425. https://doi.org/10.1017/S1355617720000922

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4), 443–460. https://doi.org/10.1007/BF02296207

Paek, I., Li, Z., & Park, H.-J. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research*, 51(4), 569–580. https://doi.org/10.1080/00273171.2016.1178567

Petersen, R. C., Roberts, R. O., Knopman, D. S., Boeve, B. F., Geda, Y. E., Ivnik, R. J., Smith, G. E., & Jack, C. R., Jr. (2009). Mild cognitive impairment: Ten years later. *Archives of Neurology*, 66(12), 1447–1455. https://doi.org/10.1001/archneurol.2009.266

Prieto, G., Delgado, A. R., Perea, M. V., & Ladera, V. (2011). Differential functioning of mini-mental test items according to disease. *Neurología (English Edition)*, 26(8), 474–480. https://doi.org/10.1016/j.nrleng.2011.01.007

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17* https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Sayegh, P., & Knight, B. G. (2014). Functional assessment and neuropsychiatric inventory questionnaires: Measurement invariance across hispanics and non-Hispanic whites. *The Gerontologist*, *54*(3), 375–386. https://doi.org/10.1093/geront/gnt026

Shi, D., Maydeu-Olivares, A., & DiStefano, C. (2018). The relationship between the standardized root mean square residual and model misspecification in factor analysis models. *Multivariate Behavioral Research*, *53*(5), 676–694. https://doi.org/10.1080/00273171.2018.1476221

Siciliano, M., Chiorri, C., Passaniti, C., Sant'Elia, V., Trojano, L., & Santangelo, G. (2019). Comparison of alternate and original forms of the Montreal Cognitive Assessment (MoCA): An Italian normative study. *Neurological Sciences*, *40*(4), 691–702. https://doi.org/10.1007/s10072-019-3700-7

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Tang, J. Y. M., Wong, G. H. Y., Luo, H., Liu, T., & Lum, T. Y. S. (2020). Cognitive changes associated with mentally active lifestyle and structured cognitive programs: A 2-year longitudinal study. *Aging & Mental Health*, *24*(11), 1781–1788. https://doi.org/10.1080/13607863.2019.1636204

Tsai, C.-F., Lee, W.-J., Wang, S.-J., Shia, B.-C., Nasreddine, Z., & Fuh, J.-L. (2012). Psychometrics of the Montreal Cognitive Assessment (MoCA) and its subscales: Validation of the Taiwanese version of the MoCA and an item response theory analysis. *International Psychogeriatrics*, *24*(4), 651–658. https://doi.org/10.1017/S1041610211002298

Wong, A., Law, L. S., Liu, W., Wang, Z., Lo, E. S., Lau, A., Wong, L. K., & Mok, V. C. (2015). Montreal cognitive assessment: One cutoff never fits all. *Stroke*, *46*(12), 3547–3550. https://doi.org/10.1161/STROKEAHA.115.011226

Wong, A., Yiu, S., Nasreddine, Z., Leung, K. T., Lau, A., Soo, Y., Wong, L. K., & Mok, V. (2018). Validity and reliability of two alternate versions of the Montreal cognitive assessment (Hong Kong version) for screening of mild neurocognitive disorder. *Plos one*, *13*(5). https://doi.org/10.1371/journal.pone.0196344

World Health Organization. (2021). *World health organization fact sheet - dementia*. https://www.who.int/news-room/fact-sheets/detail/dementia

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. https://doi.org/10.1007/s11336-016-9506-0