Expanded Cognitive Factor Structure Revealed by High-Resolution Computerized Neuropsychological Assessment

David L. Woods[1], Kathleen Hall[1], Isabella Jaramillo[1], Mike Blank[1], Kristin Geraci[1], Peter Pebler[1], and David K. Johnson[2]

[1]*Neurobehavioral Systems, Inc., Berkeley, CA*

[2]*UC Davis Alzheimer's Disease Research Center, Walnut Creek, CA*

Supported by NIA R44AG097322.

Correspondence: drdlwoods@neurobs.com.

Short title: CCAB Cognitive Factor Structure

Keywords: computerized neuropsychological assessment; confirmatory factor analysis; bifactor model; cognitive aging; speech and language biomarkers

Key Findings

- We evaluated confirmatory factor analysis (CFA) models of measures of accuracy, response time, kinematics, and Speech and Language Biomarkers (SLBS) from 22 tests of the California Cognitive Assessment Battery (CCAB). CFA analysis of performance of 1,916 community-dwelling adults recovered a bifactor cognitive architecture: a general factor (g) with strong executive-function saturation, plus four orthogonal specific factors—Processing Speed (PS), Memory (MEM), Lexical/Story processing (LS), and Speech Fluency (SF).
- Model fit was acceptable on unresidualized indicators (robust CFI = 0.976, RMSEA = 0.076, SRMR = 0.027) and improved further after demographic residualization (robust CFI = 0.986, RMSEA = 0.043, SRMR = 0.022) revealing that demographic correction at the indicator level clarified latent cognitive structure (ΔAIC ≈ −3,151).
- Pairwise specific-factor covariances were modest ($|r| \leq 0.49$) supporting factor independence with metric, but not scalar independence, seen in comparisons of subsamples differing in age.
- Executive Function loaded near-exclusively on g (standardized $\lambda = 0.75–0.82$ after residualization) with no detectable EF-specific factor; attempts to model EF as a specific factor produced Heywood cases.
- Novel Lexical/Story processing and Speech Fluency factors, driven by SLBs, emerged as separable latent factors—dissociating lexical-content and speech-timing measures hypothesized by psycholinguistic theory.
- The bifactor architecture replicated under substantial indicator expansion in a subsample (n = 1,031) undergoing nine additional cognitive tests with 46 additional measures (robust CFI = 0.983, RMSEA = 0.050).

Abstract

Background. Traditional neuropsychological assessment batteries typically generate fewer than 15 performance scores, limiting the dimensionality of factor analytic models of latent cognitive structure. Computerized assessments can provide additional automated scores by adding process-level metrics such as response latencies and computationally derived measures (e.g., semantic clustering in verbal fluency tasks) and speech and language biomarkers (SLBs). We hypothesized that this expanded indicator space would reveal cognitive dimensions invisible with traditional scoring.

Methods. We analyzed 22 subtests of the California Cognitive Assessment Battery (CCAB) that included 107 measures of accuracy, response time, kinematics, and SLBs. Performance was assessed in a diverse sample of 1,916 native English-speaking participants (mean age 53.1 years) who completed a 2-hour, telemedically proctored at-home CCAB assessment. Indicators were grouped into five theory and correlation-driven cognitive domains—Executive Function (EF), Memory (MEM), Lexical/Story processing (LS), Processing Speed (PS), and Speech Fluency (SF). Each domain was divided into two clusters defined with correlation analysis to enable cross-validation of factor loadings. Bifactor confirmatory factor analysis (CFA) results were

compared using: (1) demographically unregressed z-scores and (2) demographically residualized z-scores, computed via a regression (C-model) accounting for age, vocabulary, sex, race/ethnicity, language, and other demographic variables. Model generalization was examined in two subsamples differing in mean age, and robustness analysis was conducted in one of the subsamples (n=1,031) who underwent nine additional cognitive tests with 46 additional measures.

Results. Solutions showed acceptable fit on raw indicators (robust CFI=0.976, RMSEA=0.076, SRMR=0.027) and excellent fit after demographic residualization (robust CFI=0.986, RMSEA=0.043, SRMR=0.022; ΔAIC ≈ −3,151). EF parcels loaded near-exclusively on g in both models (residualized λ=0.75–0.82) ; attempts to model EF as a specific factor produced Heywood cases. Specific-factor loadings on PS, LS, MEM, and SF ranged 0.61–0.77 in residualized estimation with inter-factor covariances uniformly modest (|r| ≤ 0.49). The bifactor architecture showed metric invariance across subsamples, and replicated with expanded indicator coverage in the expanded-measure subsample.

Conclusions. The CCAB five-factor cognitive architecture revealed separable SLB-derived factors, Lexical/Story and Speech Fluency, as distinct dimensions that reflect a content-versus-timing dissociation in language production hypothesized by psycholinguistic theory. The novel cognitive structure was robust to indicator enrichment and demographic residualization, consistent with a stable latent cognitive architecture unconfounded by demographic variance.

Keywords: neuropsychological assessment; factor analysis; bifactor model; computerized assessment; speech and language biomarkers; processing speed; cognitive aging.

Introduction

Factor analytic models of cognition have long sought to characterize the latent dimensions of performance in neuropsychological assessments. Spearman's addition of a general factor (g) to the Cattell-Horn-Carroll (CHC) hierarchy of broad and narrow abilities (Schneider & McGrew, 2018) and Salthouse's structure of cognitive aging (Salthouse, 2010), have shown that adult cognition can be parsimoniously described by a small number of correlated factors organized hierarchically beneath general ability (Salthouse, 2010; Reise, 2012).

Despite this conceptual stability, the measures used to generate the data for factor analytic models have changed remarkably little. The dominant instruments in cognitive aging and Alzheimer's disease (AD) research include the National Alzheimer's Coordinating Center Uniform Data Set version 3 (NACC UDS-3) neuropsychological battery (Weintraub et al., 2018), the NIH Toolbox Cognition Battery (Mungas et al., 2014), the Wechsler Adult Intelligence Scale (Wechsler, 2008), and the Salthouse Virginia Cognitive Aging Project battery (Salthouse, 2010). These batteries analyze seven to 16 performance indicators based on summary metrics such as total accuracy, total time, and longest span. Factor analyses of these batteries necessarily recover whatever latent structure the limited set of summary scores support.

Recent factor analytic work on the UDS-3 illustrates both the achievements and the limits of this approach. Kiselica et al. (2020) reported a higher-order factor model with five lower-order factors—processing speed/executive functioning, visual processing, attention, language, and memory—in a sample of 2,520 cognitively normal older adults and reported good global fit (CFI=0.962, RMSEA=0.054, SRMR=0.036). Matusz et al. (2025) extended this analysis to a heterogeneous sample of 29,462 NACC participants spanning the cognitive continuum from cognitively normal to dementia. The original five-factor solution showed degraded absolute fit in the larger sample, the visuospatial factor (defined by only two indicators from a single test) became unstable, and the higher-order model produced a Heywood case in the speed/executive factor. Matusz et al. retained correlated-factors solution with four factors (memory, language, speed/executive, attention) but reported inter-factor correlations approaching unity (Speed/Executive vs. Attention r = 0.94; Speed/Executive vs. Language r = 0.89; Language vs. Memory r = 0.83) suggesting that several of the recovered factors were statistically indistinguishable from one another. Similar patterns of factor fusion and visuospatial instability appear in factor analyses of the NIH Toolbox in older adults (Hackett et al., 2024; Rose et al., 2025) and in CHC-based analyses of harmonized aging batteries (Gross et al., 2020).

We propose that these recurring challenges—high inter-factor correlations and factor solutions that vary substantially with sample composition—reflect underlying constraints of manual constraints on test quantification. When a battery samples five to seven cognitive tasks with one or two summary scores per task,

the resulting indicator covariance matrix can support only coarse factor resolution. Cognitive constructs that share substantial summary-score variance cannot be separated, even when underlying neuropsychological theory and clinical observation suggest that they are dissociable systems.

Computerized administration of neuropsychological tests enhances the indicator space available for factor analysis beyond conventional accuracy and total scores. The manual administration of cognitive tests cannot capture millisecond-resolution response latencies, kinematic trajectories of pen movement, articulation rates during connected speech, or lexical-semantic properties of discursive speech. For example, the CCAB (Woods et al., 2024) generates four classes of additional process-level measures that are invisible to manual scoring:

- **Temporal measures.** While conventional batteries operationalize processing speed largely through Trails A completion time; the CCAB quantifies processing speed with dozens of timing measures, including reaction times, down times (the time of finger contact with the touch screen or mouse response), and time-per-response measures of verbal responses.
- **Kinematic measures.** Digital pen capture during figure drawing and trail making, yields drawing speed, acceleration, pause duration, and stroke segmentation features. These dissociate motor planning from execution and provide quantitative substrates for visuoconstruction performance that manual scoring collapses into a single accuracy score.
- **Lexical-statistical measures.** Automated SLB analysis of recorded speech yield lexical diversity metrics including Brunet's index, Honoré's statistic, type-token ratio, word count, and entropy. These measures quantify the lexical content and diversity of language output in parallel to the automatic analysis of response correctness.
- **Temporal speech markers.** SLB analysis yield speech rate, pause ratio, articulation rate, and inter-word intervals. These measures quantify how fluently speech is produced independently of the lexical content being articulated.

In the current dataset, the CCAB generated 107 measures (including conventional scores) across 22 subtests in each 2-hour assessment including conventional scores. This enhanced indicator space enabled the separation of performance components (e.g., , accuracy, processing speed, organization) on individual tests that are conflated in traditional batteries. For example, in logical memory tests SLBs permit the analysis of speech production and speech organization separately from speech content (e.g., target words recalled).

A second methodological gap in the existing factor analytic literature concerns the treatment of demographic variance. Demographic variables—age, education, sex, race/ethnicity, and vocabulary—account for substantial variance in neuropsychological test scores. Prior factor analyses have addressed this in two ways. Kiselica et al. (2020) extracted factor scores from the unregressed CFA solution and applied subsequent demographic adjustment as a regression-based scoring correction. Matusz et al. (2025) tested measurement invariance across demographic groups using multi-group CFA on unregressed indicators. However, neither approach addressed the extent to which factor structure itself is subject to demographic influences.

The present study addresses both gaps. We estimate a bifactor CFA on 107 CCAB metrics, organized into 10 paired measurement clusters in five cognitive domains and compare the resulting factor solution under two estimation conditions: (1) demographically unregressed z-scores and (2) demographically residualized z-scores computed via a C-model regression. Preliminary results suggested that (a) the bifactor architecture would resolve five separable cognitive constructs—a general factor with strong executive-function saturation, plus specific factors for processing speed, memory, lexical/story processing, and speech fluency—reflecting the expanded indicator coverage of the CCAB; (b) the structure would replicate across raw and residualized estimation, demonstrating that the latent architecture is robust to demographic correction. We additionally tested model generalization and model robustness in an older (mean age 62.2 years) temporally defined subsample (n=1,031) that completed an expanded battery incorporating nine additional cognitive tests and 46 additional metrics.

Methods

Participants

Participants were 1,916 native English-speaking adults aged 18–89 years (mean = 53.1, SD = 17.3) recruited from multiple cohorts contributing to the CCAB normative database. The sample was 56.6% female and averaged 15.4 years of education. Self-reported race was 37.2% White, 22.6% Black/African American, 17.9% Asian, and 22.3% other/multiracial; 15.6% of participants reported Hispanic/Latino ethnicity. Nearly 30% reported post-college education. Inclusion criteria required native English language proficiency, adequate visual and auditory acuity for telemedical assessment, sufficient computer use proficiency to operate the testing interface, and absence of self-reported neurological or psychiatric diagnosis. A subsample of 1,031 participants—the earliest temporally enrolled cohort— completed additional tests in an expanded battery on a sparate day. All participants provided written informed consent under a protocol approved by the WCG Institutional Review Board.

Demographic characteristics of the full sample (n=1,916) and the initial subsample (n=1,031) are summarized in Table 1. The full sample spanned a wide age range (18–89 years, mean = 53.1) and was demographically diverse, with substantial representation across White, Black, Asian, and other race/ethnicity groups, and highly educated with a mean 15.4 years of education. The initial subsample showed a similar racial/ethnic distribution and educational attainment but with older mean age (62.2 years versus 53.1 years in the full sample) reflecting the targeted enrollment of older participants in the earliest recruited cohorts.

Table 1. Demographic characteristics of the full sample and INITIAL subsample.

Characteristic	Full sample (n=1,916)	Initial subsample (n=1,031)
Age, M (SD)	53.1 (17.3)	62.2 (13.8)
Age range (years)	18–89	18–89
Education, years, M (SD)	15.4 (2.4)	15.8 (2.5)
Female, n (%)	1,084 (56.6%)	578 (56.1%)
Race/Ethnicity, n (%)		
White (non-Hispanic)	712 (37.2%)	379 (36.8%)
Black/African American	432 (22.6%)	286 (27.7%)
Asian	343 (17.9%)	189 (18.3%)
Hispanic/Latino	298 (15.6%)	133 (12.9%)
Other / Multiracial	427 (22.3%)	177 (17.2%)

Note. *Counts and percentages are computed directly from the analyzed datasets. Race categories (White, Black/African American, Asian, Other/Multiracial) are mutually exclusive; Hispanic/Latino ethnicity is reported separately and may overlap with any race category.*

California Cognitive Assessment Battery (CCAB)

The CCAB is a comprehensive computerized neuropsychological battery designed for at-home, telemedical assessment of older adults (Woods et al., 2024). The core 22-subtest battery requires 2-2.5 hours and is largely self-administered on a tablet computer with examiner proctoring via video link. Tests cover memory (Bay Area Verbal Learning Test, Logical Memory, Figure Drawing recall, Digit Span Forward and Reverse), processing speed (Symbol-Number, Trails A, Trails B, Stroop, Hidden Patterns, Identical Pictures, Symbol Search, Speeded Naming, Continuous Picture Naming), executive function (Stroop interference, Trails B, Verbal Fluency in six semantic categories with category-switching analysis, Semantic Stroop), language and discursive speech (Picture Description encoding and delayed recall, Logical Memory encoding and recall, Automatic Speech Recognition Q/A testing, and Story Reading), and visuoconstruction (Figure Drawing copy and Recall, Design Fluency). The expanded battery included nine additional tests: Finger Tapping, Mental Rotation, Simple Reaction Time, Choice Reaction Time, a Continuous Picture Naming Variant, Face-Name Binding, a Logical Memory Task with Read-Aloud Encoding, a Short-Form Bay Area Verbal Learning Test, and

a simplified Figure Drawing and Recall Task. All recorded responses were processed through automated pipelines that extracted conventional accuracy and timing scores along with kinematic features (for drawing tasks) and acoustic-phonetic and SLB features (for spoken responses). Detailed descriptions of subtest design, administration, automated scoring (including Consensus Automatic Speech Recognition), test–retest reliability, and prior comparisons to manual neuropsychological tests are provided in Woods et al. (2024) and in individual CCAB technical reports (www.ccabresearch.com).

The 107 measures used in the core analysis comprised approximately 34 accuracy scores, 10 response time and processing rate indicators (e.g., reaction times, completion times), 12 kinematic indicators (e.g., segment duration variability in trail making), 12 temporal SLBs (e.g., speaking rate, articulation rate, pause ratio), and 40 lexical and syntactic SLBs (e.g., Honoré's statistic, word entropy, tree depth, content keywords, etc.).

Indicator Pre-Processing Pipeline

Polarities of all indicators were adjusted such that higher values consistently reflected better cognitive performance. Temporal measures (e.g., reaction times, inter-word intervals, completion times etc.) were log-transformed and inverted so that higher values consistently reflected faster, more efficient performance. Indicators were then winsorized at ±3 standard deviations to reduce the influence of extreme values. Indicators were z-scored against the full normative sample, with z-scores subsequently re-winsorized to ±4 SD to control for residual outliers. Participants missing more than 25% of scores (<1%) were excluded from analysis. All missing individual scores (< 2%) were then imputed using an R implementation of the random-forest-based missForest algorithm (Stekhoven & Bühlmann, 2012).

Domain and Parcel Construction

Clustering was necessary because the indicator count substantially exceeded the feasible parameter space for stable bifactor estimation. The 107 indicators were assigned to five cognitive domains based on theoretical content and empirical correlation structure: Executive Function (EF), Memory (MEM), Lexical/Story processing (LS), Processing Speed (PS), and Speech Fluency (SF). Domain assignment was guided by measurement type and process. EF indicators included verbal fluency core scores and computational measures (semantic clusters, switch counts), Stroop and Semantic Stroop interference and accuracy, Trails B-Trails A, and Figure Drawing efficiency scores. MEM indicators included BAVLT trial-by-trial hits and recall, BAVLT-F variants, Figure Drawing recall accuracy, Face-Name encoding and retrieval, and mean digit span in Forward and Reverse testing. LS indicators included lexical-statistical features (Brunet's index, Honoré's statistic, word entropy, word counts) and content-keyword retrieval (total match counts) from connected speech tasks (Logical Memory, Picture Description). PS indicators included response time, decision time, and total processing rate measures from speeded tasks (Hidden Patterns, Identical Pictures, Symbol Search, Simple Reaction Time, Choice Reaction time,) as well as motor speed measures (Finger Tapping, button depression time, and screen-touch duration), and kinematic features from trail making. SF indicators comprised temporal markers of speech production: speaking rate, pause ratio, articulation rate, inter-word interval, and response times from naming tasks.

Each domain was further partitioned into two subclusters (e.g., EF_1, EF_2) by balanced random assignment of tasks within domains, yielding ten parcels that maximized task segregation across subclusters. Subcluster scores were computed as the mean of constituent indicators. Subcluster-pairing enabled cross-validation of factor loadings within a domain: if a domain reflects a coherent latent construct, subclusters should load comparably on the corresponding specific factor.

Demographic Residualization of all scores.

For the demographically residualized analysis, a parallel set of indicators was derived from each unregressed score using a regression model (the Comprehensive C-model; Woods et al., 2024) with candidate predictors of age, education, gender, residualized age², vocabulary (assessed via the CCAB adaptive Vocabulary subtest), race (Black, Asian, other), Latino ethnicity, computer use, and daily medications. The C-model used LASSO to identify only significant predictors for inclusion in scoring models (median number of included predictors = 4) and accounted for 35–50% of domain variance across domain indicators. The same z-scoring and

winsorization procedure was applied to residualized indicators. Residuals from the C-model regression were used as inputs to create the residualized CFA.

Confirmatory Factor Analysis

Bifactor confirmatory factor analyses were estimated in the lavaan package (Rosseel, 2012) for R version 4.4.0. The model specified a general cognitive ability factor (g) loading on all ten parcel indicators, plus four orthogonal specific factors (PS, LS, MEM, SF) loading on their corresponding parcel pairs. Executive Function indicators loaded almost exclusively on g, with attempts to include a separate Executive Function resulting in Heywood cases. The EF clusters showed broad cross-domain correlations indicative of high g-saturation.

Specific factors were modeled as orthogonal to g (a defining feature of the bifactor parameterization) and were allowed to covary with one another. Within each specific factor, loadings of the two cluster indicators were constrained to equality to identify the model and reduce parameter count, given the symmetric construction of randomly partitioned parcels.

Models were estimated using maximum likelihood with robust (Huber-White) standard errors and Satorra-Bentler scaled test statistics (MLR estimator), which provides consistent fit-index and parameter estimation under non-normality without requiring distributional transformations of indicators (Satorra & Bentler, 2001). Full information maximum likelihood (FIML) was used for missing data. Global model fit was evaluated using the comparative fit index (CFI; ≥0.90 acceptable, ≥0.95 good), root mean square error of approximation (RMSEA; ≤0.08 acceptable, ≤0.05 good) with associated 90% confidence intervals, and standardized root mean square residual (SRMR; ≤0.08 acceptable, ≤0.05 good). Robust (scaled) versions of these indices were primary, supplemented by standard versions and information criteria (AIC, BIC) for model comparison.

Two CFA estimations were performed on the full cohort: (1) on raw demographically unregressed z-score parcels and (2) on demographically residualized parcels. Bifactor model specification was identical across the two estimations to enable direct comparison of factor structure under demographic correction. Additional robustness analysis was conducted in the initial subsample (n=1,031) using the same parcel structure with 46 indicators from the nine additional tests folded into existing subclusters.

Construct Validity Analyses

To assess the cognitive interpretability of the recovered factors, we extracted factor scores from the unregressed bifactor solution and computed Pearson correlations with two demographic variables of theoretical interest in cognitive aging: age and vocabulary (assessed by the CCAB adaptive Vocabulary subtest; Woods et al., 2024) with vocabulary serving as an proxy of crystallized verbal knowledge as in the CHC framework. Differential patterns of age and vocabulary correlation across factors provided construct-level evidence for the fluid–crystallized distinction and the separability of recovered factors.

Metric invariance

To assess whether the bifactor measurement model was equivalent across enrollment cohorts, we conducted a multi-group invariance analysis comparing the earlier-enrolled cohort (n = 1,031) with the younger later enrolled group that did not complete the expanded battery (n = 885). Configural, metric (equal loadings), and scalar (equal loadings and intercepts) models were estimated and compared by scaled likelihood-ratio tests (Satorra & Bentler, 2001). Invariance was evaluated in both unregressed and demographically residualized indicators. Because cohort membership was confounded with both age and test wave, scalar non-invariance is interpreted as a level difference of indeterminate origin (cohort, age, or exposure) rather than as measurement bias per se.

Results

Subcluster Correlation Structure

Figure 1 shows the subcluster correlation matrices for demographically uncorrected (left) and residualized (right) indicators. Within-domain correlations on the diagonal blocks (EF_1×EF_2, EM_1×EM_2, LS_1×LS_2, PS_1×PS_2, SF_1×SF_2) were uniformly the strongest cells in their respective rows and columns in both

matrices, supporting the subcluster-pairing strategy. Within-domain correlations ranged from 0.62–0.88 in uncorrected indicators and 0.48–0.80 in residualized indicators, with LS showing the tightest within-domain coupling (0.88 uncorrected; 0.80 residualized) and EM showing the weakest under residualization (0.48).

The correlation matrices show three notable features. First, EF parcels exhibited the highest and most uniform off-diagonal correlations across both matrices (uncorrected: 0.24–0.57 with other domains; residualized: 0.15–0.32), consistent with executive control as a broadly distributed cognitive process. This empirical pattern is consistent with the bifactor specification in which EF fused with g. Second, demographic residualization compressed off-diagonal correlations by approximately 30–50% across most domain pairs while preserving the relative ordering of correlation magnitudes. Thus, the residualized matrix is visibly paler than the uncorrected matrix in cells representing cross-domain relations, with within-domain blocks remaining the most saturated. Third, several cross-domain correlations seen in unregressed scores were substantially reduced in residualized indicators—notably PS_1×LS_1 fell to 0.21 and PS_2×LS_2 to 0.15—suggesting that part of the shared variance in unregressed indicators was demographic in origin.

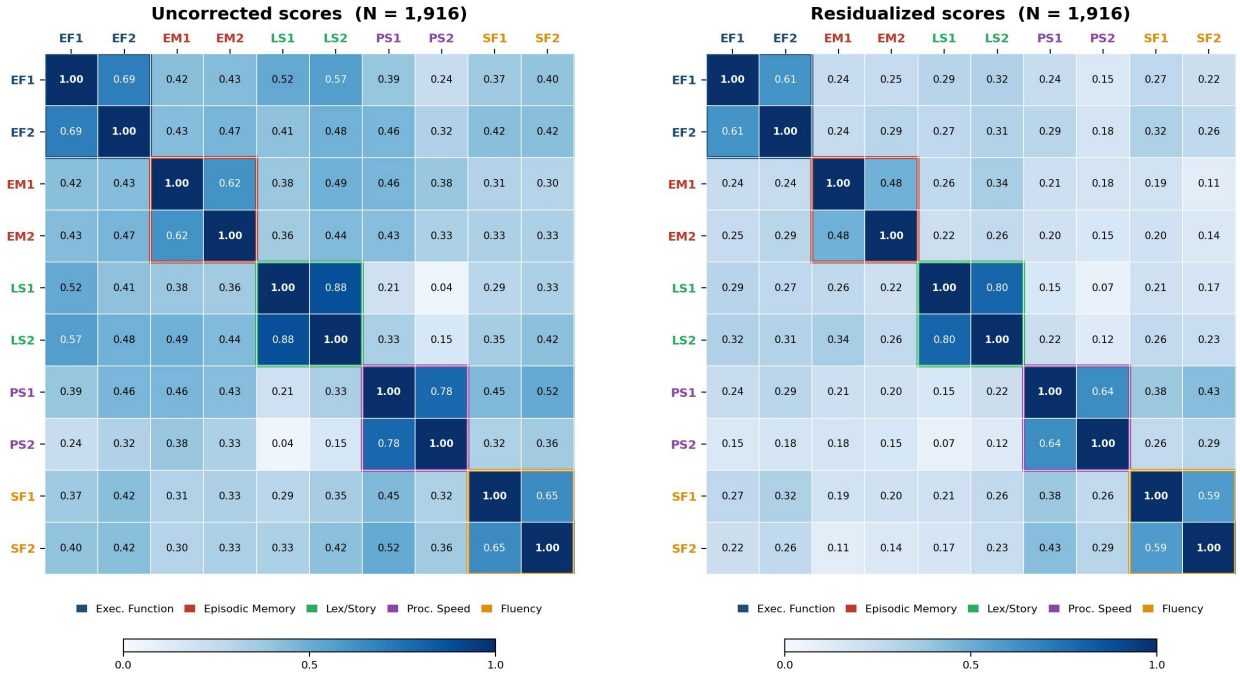

Figure 1. *Inter-subcluster correlation matrices for demographically uncorrected (left) and residualized (right) CCAB scores at baseline (n=1,916 with complete data across all 107 indicators). Each cognitive domain is represented by two subclusters constructed by random partition of indicators within domain. Within-domain correlations (colored boxes on the diagonal: EF=blue, MEM=red, LS=green, PS=purple, SF=orange) are uniformly the strongest cells in their rows and columns, supporting the parcel structure. EF parcels show the highest and most uniform off-diagonal correlations across both matrices, consistent with executive control reflecting a domain-general process. Demographic residualization compresses off-diagonal correlations by 30–50% while preserving the relative pattern of within- versus between-domain correlations. (Note: domain abbreviation 'EM' in the figure corresponds to 'MEM' in the manuscript text.)*

Bifactor Confirmatory Factor Analysis: Full Sample

Figure 2 displays CFA path diagrams for the unregressed (panel A) and residualized (panel B) estimations in the full sample (n=1,916). The unregressed model converged after 27 iterations and the residualized model after 29 iterations, without estimation problems in either fit. Standardized loadings, factor variances, and inter-factor covariances are reported in the path diagrams.

The unregressed bifactor model showed acceptable absolute fit and good incremental fit (robust CFI=0.976, robust TLI=0.956, robust RMSEA=0.076 [90% CI: 0.068, 0.084], SRMR=0.027). The hypothesis test of robust RMSEA ≤ 0.050 was rejected ($p < .001$), and the hypothesis test of robust RMSEA ≥ 0.080 was not rejected ($p = 0.218$), indicating that absolute fit lay near the conventional acceptable boundary. AIC was 43,403 and BIC 43,626. EF subclusters loaded strongly on g (standardized $\lambda = 0.83$ for EF_1 and 0.82 for EF_2), with no specific EF factor estimated. Specific factor loadings were strong and consistent across paired parcels, with

equality constraints yielding PS = 0.77, LS = 0.70, MEM = 0.56, and SF = 0.63. The general factor loaded substantially on non-EF domains, with values 0.33–0.65 for PS, LS, MEM, and SF subclusters.

The residualized bifactor model showed substantially improved fit (robust CFI=0.986, robust TLI=0.976, robust RMSEA=0.043 [90% CI: 0.035, 0.052], SRMR=0.022). The hypothesis test of robust RMSEA ≤ 0.050 was not rejected (p = 0.885), placing absolute fit at the good-fit boundary. AIC was 40,252—a decrease of approximately 3,151 from the unregressed model—and BIC was 40,474. EF subclusters remained strongly loaded on g (λ = 0.75 for EF_1 and 0.82 for EF_2). Specific factor loadings remained stable or strengthened relative to the unregressed model: PS = 0.77, LS = 0.75, MEM = 0.61, and SF = 0.66. The g factor loadings on non-EF subclusters declined substantially after residualization, falling to 0.20–0.40 for PS, LS, and MEM parcels and to 0.32–0.38 for SF.

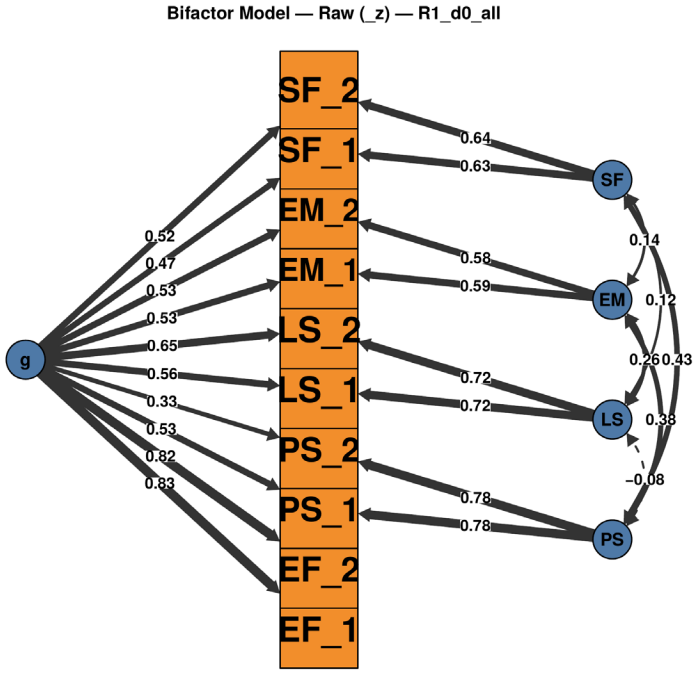

Figure 2A. *Bifactor confirmatory factor analysis path diagram for the full sample (n=1,916) using demographically unregressed indicators. Path coefficients are standardized factor loadings and standardized covariances (curved arrows). The general factor (g) loads on all ten parcel indicators including EF, while four orthogonal specific factors (PS, LS, MEM, SF) load on their respective parcel pairs. Within each specific factor, loadings of the two parcels were constrained to equality. EF parcels show strong g loadings (0.83, 0.82) with no residual specific-factor variance. (Note: 'EM' in the figure corresponds to 'MEM' in the manuscript.)*

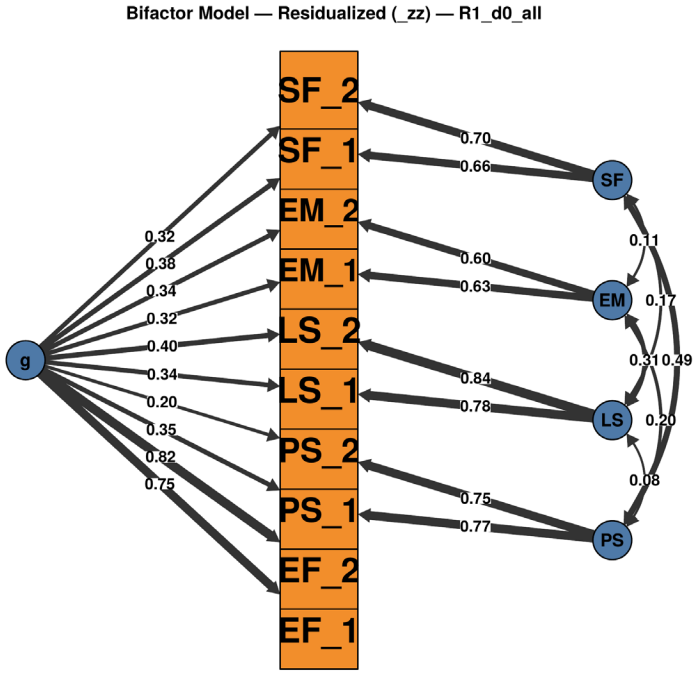

Figure 2B. *Bifactor confirmatory factor analysis path diagram for the full sample (n=1,916) using demographically residualized measures and recovering the same bifactor architecture. EF subclusters retained strong g loadings (0.75, 0.82) with specific factor loadings stable or slightly strengthened relative to the unregressed model. General-factor loadings on non-EF subclusters declined substantially after residualization, indicating that approximately 30–50% of apparent g-saturation reflected shared demographic variance. The PS×LS covariance reversed sign from –0.08 (unregressed) to +0.08 (residualized). (Note: 'EM' in the figure corresponds to 'MEM' in the manuscript.)*

Inter-Factor Covariance Patterns

Inter-factor covariances among the four specific factors changed in informative ways with residualization. PS×SF strengthened from 0.43 (unregressed) to 0.49 (residualized), and LS×MEM strengthened from 0.19 to 0.26—patterns consistent with genuine shared cognitive processes between speech production speed and processing speed and between memory recall and lexical-semantic processing. PS×MEM weakened from 0.31 to 0.16, suggesting that approximately half of the apparent processing speed–memory relationship in unregressed indicators was demographically generated, most plausibly through shared age effects. LS×SF and MEM×SF were small and stable across estimations.

However, a change with residualization occurred for PS×LS, which reversed sign from a significant negative covariance in the unregressed model (–0.08 ($z = -2.40$, $p = .016$) to a significant positive covariance (+0.08, $z = 2.73$, $p = .006$) in the residualized model. This sign reversal may have reflected opposing age effects: processing speed declined with age while crystallized lexical knowledge (vocabulary) improved, generating a spurious negative covariance. After residualization on age, vocabulary, and other demographics, the underlying cognitive relationship was revealed to be slightly positive. This pattern shows the confounding produced by demographic loadings on unregressed latent factors, and illustrates the diagnostic value of comparing factor solutions across unregressed and demographically corrected estimations.

Across both estimations, all six pairwise specific-factor covariances were ≤0.49 in absolute value, well below conventional discriminant validity thresholds. The bifactor solution thus yielded specific factors that were genuinely separable from one another.

Construct Validity: Age and Vocabulary Correlations

Figure 3 presents Pearson correlations of age and vocabulary with bifactor specific-factor scores from the unregressed solution. Each factor displayed a theoretically interpretable demographic signature.

PS exhibited the strongest age-related decline of any factor (r = –0.61, equivalent to 37% of variance explained by age alone) with a modest negative vocabulary association (r = –0.17). This is the classic fluid-ability signature: age-driven decline with limited dependence on crystallized verbal knowledge. The magnitude is consistent with the Salthouse (2010) finding that perceptual/processing speed is the most age-vulnerable cognitive dimension across the adult lifespan.

LS showed the opposite signature: a positive age effect (r = +0.32) and the strongest vocabulary association (r = +0.47). This pattern is consistent with the LS construct as a crystallized-knowledge factor, whose lexical-statistical and content-recall indicators draw heavily on accumulated vocabulary and verbal experience. Older adults performed better on lexical/story processing tasks because vocabulary continues to accumulate across the adult lifespan—an established finding in cognitive aging that the LS factor recovered cleanly.

MEM (EM) showed a moderate age decline (r = –0.33) with little vocabulary association (r = +0.07), consistent with a fluid-style memory capacity construct that is age-vulnerable but largely independent of crystallized knowledge. SF showed a modest age decline (r = –0.21) and no vocabulary association (r = +0.03), indicating that speech production timing is mildly age-sensitive and operates independently of lexical content. The general factor (g) showed a moderate age decline (r = –0.20) and a strong positive vocabulary association (r = +0.51), consistent with the heavy contribution of crystallized verbal knowledge to general cognitive ability. Thus, the differential age and vocabulary signatures across factors—PS as fluid, LS as crystallized, MEM as fluid and vocabulary independent, SF as fluid and slightly age dependent, g as a mixture—provide construct-level evidence that the bifactor solution recovers theoretically distinct cognitive dimensions.

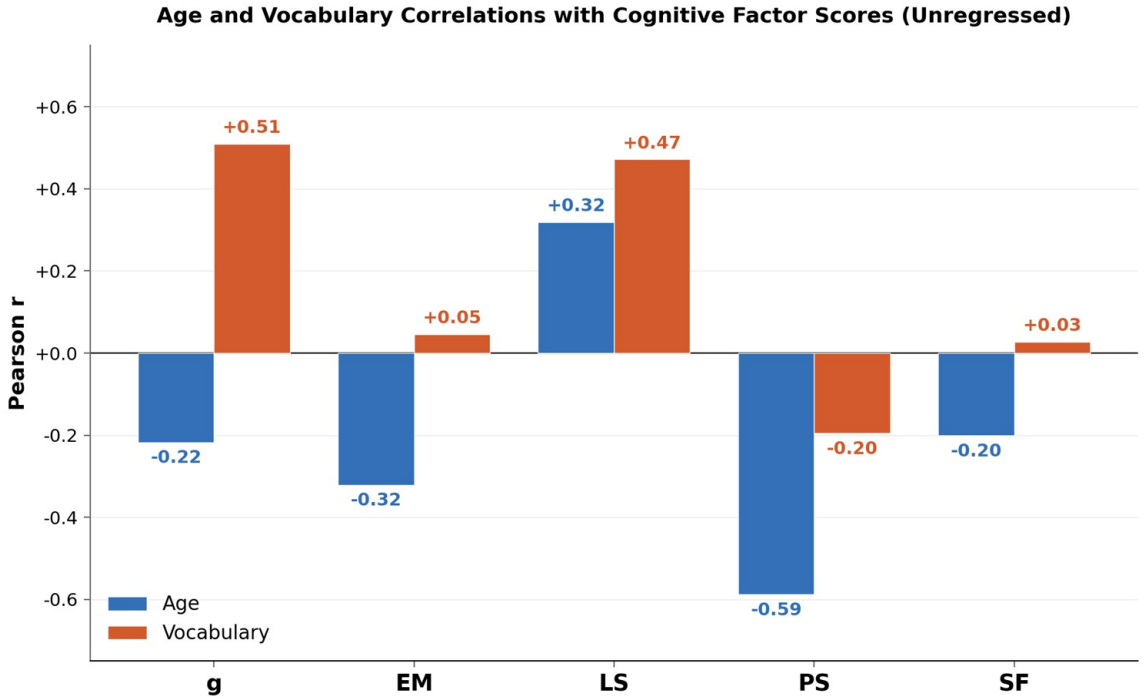

Figure 3. *Pearson correlations of age (blue) and vocabulary (orange) with bifactor specific-factor scores from the unregressed CFA solution (n=1,916). Each factor shows a theoretically interpretable demographic signature: PS exhibits the classic fluid pattern with strong age decline and weak vocabulary association (likely mediated by the increase in vocabulary with age); LS exhibits the classic crystallized pattern with positive age and strong vocabulary effects; MEM (EM) shows fluid-style age decline without vocabulary association; SF shows modest age decline without vocabulary effect. The general factor shows a mixed signature with strong vocabulary loading. Factor scores from the residualized solution are by construction uncorrelated with these demographic predictors and therefore not displayed.*

Robustness: Initial Subsample with Expanded Battery

To assess whether the bifactor structure was robust to indicator composition, we re-estimated the model in the initially enrolled cohort subsample (n=1,031) using expanded subclusters that incorporated nine additional cognitive tests. New indicators were folded into existing subclusters by measurement type and theoretical content. For example, finger tapping rate, simple and choice reaction times, and mental rotation speed and

reaction times were assigned to PS parcels; face-name binding accuracy, figure drawing recall, and short-form BAVLT recall accuracy were assigned to MEM parcels; continuous picture naming speech rate and SLB timing measures in the additional logical memory story were assigned to SF; and SLB lexical-statistical features and keyword recall from the read-aloud logical memory tasks were assigned to LS.

The bifactor architecture replicated under expanded indicator coverage. Figure 4A presents the unregressed bifactor with more extensive indicators and Figure 4B presents the residualized model. Fit indices are summarized in Table 2 alongside full-sample estimations.

Table 2. Bifactor CFA fit indices across estimations and samples.

Index	Unregressed (n=1,916)	Residualized (n=1,916)	Unregressed (n=1,031)	Residualized (n=1,031)
χ^2 (df=25)	260.58	101.43	108.66	82.97
Robust CFI	0.976	0.986	0.987	0.983
Robust TLI	0.956	0.976	0.976	0.970
Robust RMSEA	0.076	0.043	0.061	0.050
RMSEA 90% CI	[0.068, 0.084]	[0.035, 0.052]	[0.049, 0.072]	[0.039, 0.063]
SRMR	0.027	0.022	0.021	0.027
AIC	43,403	40,252	22,402	20,819

Note. *Fit indices are robust (Satorra-Bentler scaled) versions estimated with MLR. CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; AIC = Akaike Information Criterion; _z = unregressed; _zz = residualized via the C-model.*

EF parcels continued to load near-exclusively on g across both expanded indicator estimations (residualized λ = 0.78 and 0.75). Specific factor loadings remained robust, particularly in the residualized model: PS = 0.76, LS = 0.66, MEM = 0.55, and SF = 0.71. Notably, SF specific loadings strengthened relative to the original parcel structure (0.66 in the full sample vs. 0.71 with expanded indicators), reflecting additional speech-rate indicators from the added speech tasks. Several inter-factor covariances strengthened with denser sampling; most notably LS×MEM (0.31 vs. 0.46 in residualized scores) so that shared variance between narrative recall and pure encoding-retrieval operations became more visible with more indicators per domain. The PS×LS demographic-confound result also replicated: the negative raw covariance attenuated with denser sampling (–0.08 vs. +0.06 unregressed) and was positive under demographic residualization (+0.08 vs. +0.10, both non-significant), supporting the interpretation of opposing age effects on processing speed and crystallized lexical knowledge.

Residualized fit in the expanded indicator model was comparable to the full-sample model (robust RMSEA 0.043 vs. 0.050; robust CFI 0.986 vs. 0.983), despite the smaller sample size and the inclusion of indicators with heterogeneous demographic loadings (notably reaction-time indicators with strong age effects and narrative recall indicators with strong vocabulary effects). The narrowed gap between raw and residualized fit (ΔRMSEA: 0.033 vs. 0.011) suggests that denser indicator coverage partially compensated for demographic confounding, even without explicit residualization. Overall, the expanded indicator analysis confirmed that the bifactor architecture was robust to substantial expansion of the indicator space and demonstrated that specific factor loadings strengthened as additional indicators were added to subparcels.

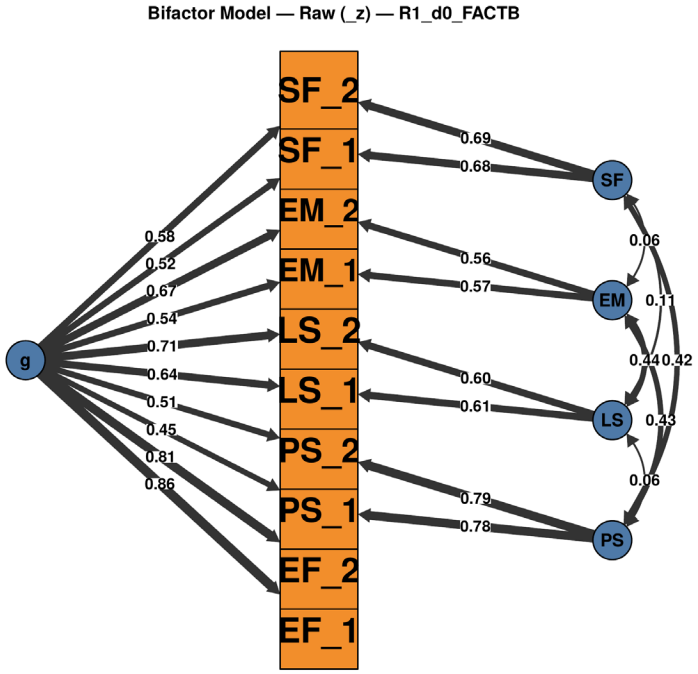

Figure 4A. *Bifactor CFA path diagram for the expanded indicator subsample (n=1,031) with expanded indicator coverage, using demographically unregressed indicators. Nine additional cognitive tests were folded into existing parcels by measurement type. The bifactor architecture replicates: EF parcels load strongly on g (0.86, 0.81), and four orthogonal specific factors were recovered.*

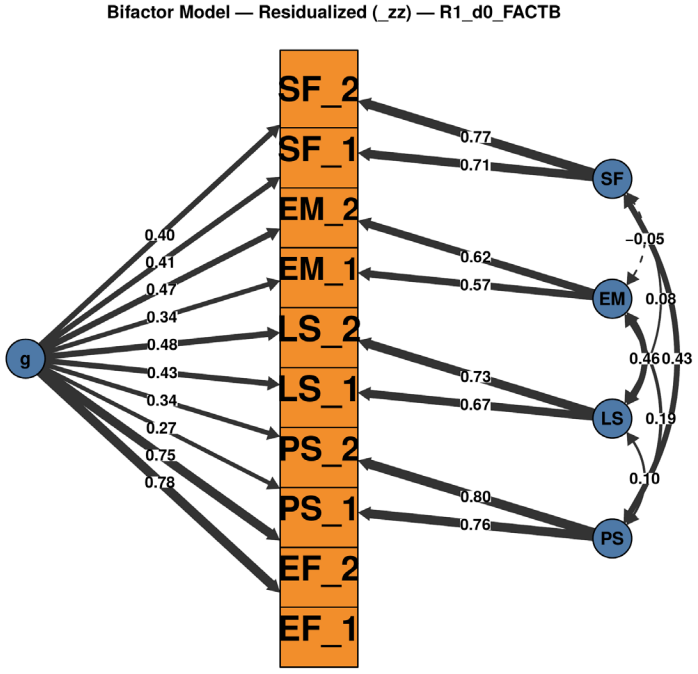

Figure 4B. *Bifactor CFA path diagram for the expanded indicator subsample (n=1,031) using demographically residualized indicators. The bifactor architecture is preserved; specific factor loadings strengthen for SF (driven by the addition of speech-rate indicators from the expanded battery), and the LS×MEM covariance increases substantially relative to the original parcel specification, suggesting genuine shared variance between narrative recall and encoding-retrieval operations made more visible with denser sampling.*

Model-Based Reliability and Dimensionality

Table 3 shows model-based reliability and dimensionality indices from the standardized loadings of each estimation (Rodriguez, Reise, & Haviland, 2016). These included coefficient omega (ω) and omega-hierarchical for the general factor (ωH), omega-hierarchical-subscale for each specific factor (ωHS), construct replicability (H), the explained common variance of the general factor (ECV), and the proportion of uncontaminated correlations (PUC). Because each specific factor was defined by two parcels, ωHS and H values represent conservative lower bounds.

Table 3. Model-based reliability and dimensionality indices for the bifactor model across estimations and samples.

Index	Unregressed (Full, n=1,916)	Residualized (Full, n=1,916)	Unregressed (more indicators, n=1,031)	Residualized (more indicators, n=1,031)
ECV (general)	0.49	0.34	0.54	0.38
PUC	0.91	0.91	0.91	0.91
ω (total)	0.94	0.88	0.95	0.89
ωH (general)	0.77	0.60	0.81	0.65
H (general)	0.88	0.81	0.90	0.82
ωHS — PS	0.68	0.70	0.67	0.71
ωHS — MEM	0.42	0.51	0.38	0.47
ωHS — LS	0.55	0.73	0.40	0.58
ωHS — SF	0.49	0.58	0.53	0.64
H — PS	0.76	0.73	0.76	0.75
H — MEM	0.51	0.55	0.49	0.53
H — LS	0.68	0.80	0.54	0.66
H — SF	0.57	0.64	0.64	0.71

Note. *ECV = explained common variance of the general factor; PUC = proportion of uncontaminated correlations; ω = coefficient omega; ωH = omega-hierarchical (general factor); ωHS = omega-hierarchical-subscale (specific factor); H = construct replicability. Specific factors are defined by two parcels each; ωHS and H are therefore conservative lower bounds. _z = unregressed; _zz = C-model residualized.*

The general factor was well-determined and replicable across all four estimations (H = 0.81–0.90; ωH = 0.60–0.81). Demographic residualization reduced the general factor's explained common variance (ECV: 0.49 vs. 0.34 in the full sample; 0.54 vs. 0.38 with expanded indicators) and its hierarchical reliability (ωH: 0.77 vs. 0.60 and 0.81 vs. 0.65, respectively), reflecting the reallocation of demographically shared g variance. In every estimation, ECV fell in the low-to-moderate range while PUC remained high (0.91), a combination that indicates a genuinely multidimensional structure for which a specific factor solution is warranted.

The specific factors showed a consistent pattern across samples. Demographic residualization increased ωHS for every specific factor in both samples, indicating that the regressed scores sharpened domain-specific measurement by removing demographic confounding. Processing speed (PS) was the most reliably determined specific factor across all estimations (ωHS = 0.67–0.71; H = 0.73–0.76) and was essentially invariant to both score correction and battery expansion. Lexical-statistical/story processing (LS) was well-determined in the full residualized model (ωHS = 0.73; H = 0.80) but weaker in the expanded battery (ωHS = 0.58; H = 0.66), consistent with dilution of lexical-statistical coherence when additional indicators were folded into parcels. Semantic fluency (SF) was moderate and strengthened under both residualization and indicator expansion (H up to 0.71). Memory (MEM) was the least reliable specific factor in all estimations (ωHS = 0.38–0.51; H = 0.49–0.55), indicating that MEM-specific factor scores—the variance remaining after the general factor was partialed out—should be interpreted with greater caution than those of PS, LS, or SF.

Cohort Invariance

Metric invariance in cohort waves differing in age and enrollment source held in both estimations: constraining factor loadings to equality across initial and later cohorts did not degrade fit relative to the configural model (_z:

scaled $\Delta\chi^2(5) = 1.00$, $p = .96$; _zz: scaled $\Delta\chi^2(5) = -21.95$, $p = 1.00$). The factor structure thus carried equivalent meaning in both groups. Factor stability was reflected in Tucker's congruence coefficient (φ) between every pair of solutions which was 0.96–0.99 for g-loadings and 0.99 for specific-factor loadings**.**

Scalar invariance was not supported in either estimation: Adding intercept-equality constraints produced significant deterioration in fit (unregressed: scaled $\Delta\chi^2(5) = 127.23$, $p = 9.2 \times 10^{-26}$; residualized: scaled $\Delta\chi^2(5) = 53.43$, $p = 2.7 \times 10^{-10}$). Indicator intercepts therefore differed across cohorts even at equal latent ability, consistent with the substantial age and wave differences between groups. Critically, demographic residualization reduced the scalar misfit by 2.4-fold ($\Delta\chi^2$ from 127.23 to 53.43) and lowered the scalar-model AIC from 42,777 to 40,054, indicating that the bulk of the cross-cohort intercept differences in raw scores reflected demographic structure (principally age) that residualization removed.

Discussion

We estimated a bifactor confirmatory factor analytic model on 107 process-level indicators from a comprehensive computerized neuropsychological battery in 1,916 community-dwelling adults. The model recovered a five-construct cognitive architecture: a general factor (g) that absorbed almost all EF-specific variance, plus four separable specific factors representing Processing Speed (PS), Memory (MEM), Lexical/Story processing (LS), and Speech Fluency (SF). The architecture replicated under demographic residualization—indeed, residualized fit was substantially better than unregressed fit by every criterion—and replicated again in a subsample with substantial expansion of the indicator space .

Specific factors in bifactor models should not be interpreted as entirely independent neurocognitive systems, but rather as residual covariance structures after partitioning general variance. Nevertheless, specific factors were genuinely separable, with all pairwise covariances ≤0.49 in absolute value, well below conventional discriminant validity thresholds. The differential demographic signatures of factor scores—PS as fluid, LS as crystallized, MEM as fluid-and-vocabulary-independent, SF as fluid but somewhat age-sensitive —provided construct-level validation that the recovered factors index theoretically distinct cognitive dimensions rather than differential age-related decline.

An alternative interpretation is that the recovered factors partially reflect shared measurement modality. However, several observations argue against this explanation: (1) indicators were drawn from multiple tasks within each domain, and included mixed measurement modalities. For example, LS included word recalled as well as lexical measures, and SF included response latencies in Continuous Picture Naming as well as SLB measures of speech and articulation rate.(2) factor structures replicated across battery expansion, and (3) construct-validity patterns with age and vocabulary followed theoretically predicted fluid/crystallized dissociations.

EF Loaded Exclusively on g

A strong empirical finding in our model is the absorption of executive function indicators into the general factor. EF parcels showed standardized g loadings of 0.82–0.83 in unregressed estimation and 0.75–0.82 after demographic correction; attempts to model EF-specific variance as an additional specific factor resulted in Heywood cases. This result was robust across sample size, parcel composition, demographic correction, and indicator expansion. It is consistent with the unity-of-EF tradition in cognitive psychology, which holds that a Common Executive Function factor accounts for substantial variance in ostensibly distinct executive operations such as inhibition, shifting, and updating (Friedman & Miyake, 2017; Miyake & Friedman, 2012) and characterizes executive control as a domain-general process (Salthouse, 2010). It is also consistent with the observation in older-adult factor analyses that executive control measures load strongly on g and show limited specific-factor identifiability after g is extracted. For example, the Matusz et al. (2025) UDS-3 model attempted to fit a higher-order g model, but, like us, encountered a Heywood case when attempting to modal an independent EF factor. Our bifactor parameterization therefore handled the g-saturation differently: by partitioning shared variance into a single general factor that included EF and constraining specific factors to be orthogonal to g, so that g absorbed EF-variance directly rather than redistributing it through high inter-factor correlations.

LS and SF Are Separable Latent Factors

The bifactor model identified LS and SF as distinct latent factors. Inter-factor covariance between LS and SF was small (0.09 unregressed, 0.14 residualized in the full sample), and both factors showed substantial specific-factor loadings on their respective subcluster pairs. This separability formalizes in psychometric structure a distinction between lexical content and articulatory-phonological timing proposed in models of spoken language production (Levelt, 1989; Dell, 1986), and clinically observed in the dissociation between semantic dementia (lexical-semantic loss with preserved articulation) and non-fluent variant primary progressive aphasia (articulation slowing with preserved lexical access; Mesulam et al., 2014; Hodges & Patterson, 2007). Conventional cognitive batteries cannot separate these constructs because they score language at the level of total accuracy. The CCAB scores both: lexical content and diversity features (Brunet's index, Honoré's statistic, word entropy, content keyword retrieval) and temporal speech features (speaking rate, pause ratio, articulation rate, inter-word intervals). Estimated jointly within a bifactor framework, these indicator classes recovered separable LS and SF latent factors.

It is worth emphasizing that we do not claim to have discovered new cognitive constructs. The lexical-versus-articulatory distinction is well-established in psycholinguistics and clinical neuropsychology while the present work revealed that distinction using factor analysis in an older-adult population. To our knowledge, no prior factor analysis of an older-adult cognitive battery has identified speech fluency as a separable latent factor distinct from a generic Language factor. The Matusz et al. (2025) UDS-3 analysis treats fluency as a function of the speed/executive factor (because phonemic fluency total counts are speed-dominated) and language as a single factor, anchored by Multilingual Naming Test totals. The NIH Toolbox Cognition Battery does not include speech fluency or lexical-statistical scoring at all. The Salthouse Virginia Cognitive Aging Project battery scores vocabulary as a single crystallized factor without process-level decomposition. Thus, the CCAB's separation of LS and SF likely reflects broader measurement coverage made possible by SLB analysis rather than theoretical innovation.

Demographic Residualization Improves Rather Than Degrades Fit

A central methodological finding of the present work is that demographic residualization at the indicator level improved bifactor model fit: robust RMSEA dropped from 0.076 to 0.043, robust CFI rose from 0.976 to 0.986, SRMR dropped from 0.027 to 0.022, and AIC declined by approximately 3,151 points—a substantial preference for the residualized model on information-theoretic grounds. This finding runs counter to the naive expectation that demographic residualization would degrade fit by removing substantively meaningful variance. Instead, the bifactor architecture was more cleanly recovered after demographic variance was removed.

This pattern may arise because demographic variables load heterogeneously across cognitive domains. Age strongly degrades PS and improves LS (through age-related increases in crystallized intelligence). Vocabulary, and to a lesser extent education, affected both LS and g/EF strongly, with smaller effects on PS and SF. Sex influences MEM and LS indicators but not PS or SF. These heterogeneous loadings generate demographic covariance among indicators that model structure does not capture—covariance that drives misfit. Removing demographic influences removes this excess covariance, clarifying the latent cognitive architecture. The general factor loadings on non-EF parcels decline by 20-40% after residualization (e.g., PS g loadings: 0.33–0.53 vs. 0.20–0.35), indicating that a substantial fraction of apparent g-saturation in raw indicators reflected shared demographic rather than cognitive variance.

This finding is novel relative to the existing factor analytic literature on cognitive aging batteries. Kiselica et al. (2020) extracted factor scores from unregressed CFA and applied demographic correction downstream as a regression-based scoring patch, justified by the failure of the UDS-3 indicators to achieve scalar invariance across age and sex groups. Matusz et al. (2025) tested measurement invariance across demographic groups using multi-group CFA on Blom-transformed but demographically unregressed indicators. Neither approach directly compares factor solutions estimated on the same indicators with and without demographic residualization. Our results confirm that the bifactor structure survives demographic residualization and demonstrate that residualization can clarify latent cognitive architecture otherwise distorted by demographic influences on unregressed scores.

One example of demographic influences altering interpretations can be seen in PS×LS covariance, significantly negative for unregressed scores ($r = –0.08$, $z = –2.40$, $p = .016$) and significantly positive (+0.08, z

= 2.73, p = .006) after demographic correction. This is a textbook instance of confounding in latent-variable modeling, and its diagnostic value depends on having both estimations available. A single-estimation analysis on raw indicators would interpret the PS×LS = –0.08 as a modest negative cognitive relationship between the constructs; a single-estimation analysis on residualized indicators would interpret the PS×LS = +0.08 as a modest positive relationship. The dual estimation makes both readings available and reveals that the apparent negative relationship is generated by demographic structure rather than by cognitive process. This methodological approach generalizes: any pairwise factor relationship that changes substantially across the two estimations is a candidate demographic confound, and any relationship that is stable across the two estimations can be interpreted as a residual cognitive relationship relatively independent of demographic influences.

The model-based reliability indices showed that the recovered specific factors were sufficiently determined to support the bifactor solution for several reasons. First, the general factor was strongly replicable (H = 0.81–0.90), so general-cognition scores are robust. Second, the conjunction of low-to-moderate ECV with high PUC across all estimations provided quantitative evidence that the battery was multidimensional, in contrast to higher-order solutions of batteries in which the general factor approaches or exceeds unity in relation to domain factors. Third, demographic residualization improved specific-factor reliability uniformly (ωHS increased for every specific factor), reinforcing the suggestion that demographic influences obscured latent structure in unregressed scores. While the reliance on two parcels per specific factor renders the reported ωHS and H values conservative lower bounds, the MEM specific factor remained the least well-determined in every estimation, so domain-specific memory interpretations warrant additional caution and corroboration.

Comparison with Prior CFA Solutions

Comparison with prior factor analytic solutions for AD-focused neuropsychological batteries highlights both the contributions and limitations of the present work. Kiselica et al. (2020) reported a higher-order factor model with five lower-order factors and found good fit (CFI=0.962, RMSEA=0.054), but with g-loadings of 1.08 on the Speed/Executive factor and 0.97 on the Language factor. Matusz et al. (2025) extended this analysis to 29,462 NACC participants spanning the cognitive continuum and obtained a four-factor correlated solution, with several cross-factor correlations approaching unity (Speed/Executive vs. Attention r = 0.94 in NACC; Speed/Executive vs. Language r = 0.89; Language vs. Memory r = 0.83). Their higher-order model failed due to a Heywood case in the speed/executive factor.

The CCAB bifactor solution differs from these UDS-3 solutions in several ways. First, our maximum inter-factor covariance was 0.49 (residualized), providing clear support for factor independence. Second, the models recovered separable LS and SF factors, reflecting the CCAB's expanded indicator coverage compared to UDS language measures. Third, we did not recover a separate visuospatial factor, in part because visuospatial indicators in this analysis were limited to figure drawing copy and recall scores and design fluency, and were absorbed into EF and MEM parcels rather than constituting a separate domain.

A second body of bifactor work bears more directly on our findings. The Wechsler Adult Intelligence Scale—Fourth Edition (Wechsler, 2008) has been the canonical reference for bifactor modeling of adult intelligence since Gignac and Watkins (2013) demonstrated that a bifactor parameterization fits the WAIS-IV normative correlation matrices better than the higher-order and oblique-correlated alternatives endorsed in the WAIS manual, an effect consistent across all standardization age groups (see also Benson, Hulac, & Kranzler, 2010; Reise, 2012, for the general bifactor framework). The WAIS-IV bifactor literature parallels our findings in three respects. First, the general factor in the WAIS-IV bifactor solution is heavily saturated, with model-based reliability (ωh) for the full-scale composite high and ωs values for index-specific composites uniformly modest—mirroring our finding that EF parcels load entirely on g and that non-EF specific factors carry substantial but bounded residual variance after g extraction. Second, Processing Speed emerged as the most identifiable specific factor in the WAIS-IV bifactor, paralleling the strong, stable PS specific factor we recovered (λ = 0.77 across estimations). Third, the Perceptual Reasoning Index, the WAIS-IV analog of visuospatial ability, is well identified in the WAIS-IV bifactor because the Block Design, Matrix Reasoning, and Visual Puzzles subtests provide three dedicated indicators. The CCAB's currently sparse visuoconstructive sampling—anchored primarily by figure drawing copy and recall scores and design fluency—did not support an analogous separable factor. The convergence of WAIS and CCAB models on a strong g and a robust Processing Speed specific factor reinforces the interpretation that our results are not artifacts of the CCAB's

process-level scoring; the divergence on LS-versus-SF separability and on visuoconstructive coverage reflects the differential indicator densities of the two batteries

Three differences are equally notable. The WAIS-IV bifactor is estimated on 10–15 conventional summary subtest scores; the CCAB bifactor model was estimated on 107-153 process-level indicators distributed across temporal, kinematic, acoustic, and lexical-statistical measurement modalities. The WAIS-IV bifactor does not address demographic confounding because its indicators are pre-normed; the dual-estimation contrast between unregressed and regressed solutions that we report is therefore not available in the WAIS-IV literature. And the WAIS-IV bifactor does not separate language into lexical-content and articulation-timing factors because the Vocabulary, Similarities, and Information subtests are accuracy-only summary scores. .

We note that no prior CFA in this literature has compared raw and demographically residualized estimations on the same bifactor specification. The dual-estimation design and the resulting structural-robustness finding represent a methodological contribution that extends the existing UDS-3, NIH Toolbox, and WAIS-IV literatures in a way that does not require new data collection but rather a different analytic approach to existing data. We anticipate that comparable dual-estimation analyses on UDS-3, NIH Toolbox, or WAIS-IV indicators would reveal similar patterns of demographic confounding and similar improvements in fit under residualization.

Factor Models in Existing Clinical Practice

A natural question is whether the factor-analytic results reported here translate into scores that clinicians can use when interpreting cognitive assessments. Commercial neuropsychological scoring software—including the Wechsler Adult Intelligence Scale—Fourth Edition (Wechsler, 2008)—reports subtest scaled scores and conventional Index composites (Verbal Comprehension, Perceptual Reasoning, Working Memory, Processing Speed) plus a Full Scale IQ. These composites are sum-score aggregates, not factor scores derived from a fitted CFA, and reflect correlated-factors rather than bifactor parameterization. The bifactor literature on the WAIS-IV has consistently demonstrated that a bifactor specification fits the standardization data better than correlated-factors alternatives (Gignac & Watkins, 2013) and replicates in clinical neuropsychological samples (Nelson, Canivez, & Watkins, 2013; Collinson et al., 2017), but bifactor specific-factor scores are not extracted, normed, or returned to clinicians by any commercial scoring system. The field's interpretive practice has remained anchored to the four-index correlated-factors model that the WAIS manual endorses, even where the manual's own data favor a bifactor parameterization.

Two exceptions are worth noting. First, Kiselica et al. (2020) constructed a higher-order factor solution for the NACC UDS-3 (g plus five lower-order factors: processing speed/executive, visual, attention, language, memory) and made a freely available Excel calculator for demographically adjusted factor scores. The calculator is used in research and is available to clinicians who download it, although there is no centralized evidence on routine clinical adoption. Second, the UDS3-EF (Staffaroni et al., 2021) is a factor-derived executive function composite developed in 3,507 NACC controls using item-response theory and validated as a cognitive endpoint for clinical trials, where it required approximately 40% of the sample size of the next-best individual test to detect change. UDS3-EF is a single composite intended for trial endpoints rather than profile reporting, but it represents an instance of a CFA-derived score replacing single-test scores in a high-stakes evaluation context. Neither the Kiselica calculator nor the UDS3-EF uses a bifactor parameterization; both are correlated-factors or higher-order specifications applied to summary subtest scores. The NIH Toolbox Cognition Battery distributes Crystallized, Fluid, and Total Cognition composites (Mungas et al., 2014) are factor-informed in derivation, but again are correlated-factors composites rather than bifactor specific-factor scores.

Against this backdrop, the CCAB is, to our knowledge, the first cognitive assessment battery in which (a) a bifactor CFA is fit to process-level indicators rather than to conventional summary scores, (b) bifactor specific-factor scores are returned alongside raw and demographically corrected cluster scores, and (c) the resulting profile included clinically meaningful dissociations (LS versus SF; selective MEM impairment against preserved g) that conventional commercial scoring cannot provide. The case is not that bifactor modeling is new—it has been the recommended psychometric approach for the Wechsler scales for over a decade—but that it has not been implemented in a clinical reporting pipeline. The contribution of the present work is therefore both psychometric (recovering a five-factor architecture with separable LS and SF) and translational (constructing a reporting pipeline that delivers bifactor profile scores to clinicians who can use them). This pipeline may assist

in identifying domain-specific cognitive deficits, particularly in older patients, where dedifferentiation progressively conflates domain-specific operations with general cognitive ability and renders the bifactor decomposition increasingly informative relative to conventional domain composites (Woods et al., 2026).

Factor analysis in clinical trials

Factor analytic interpretation of multi-site cognitive data depends on the indicator-level measurements being free of administration-source variance. In manual neuropsychological administration, this condition is not met. For example, different examiners deliver word lists at different speaking rates — a variable that affects episodic memory encoding (Wingfield, Tun, & McCoy, 2005). Different examiners also have different speech intensities and different prosodic patterns, both of which affect connected-speech processing, particularly in older adults whose peripheral hearing loss compromises speech perception. And different examiners apply scoring rubrics with variable strictness, particularly on tasks where partial-credit or qualitative judgments are required (verbal fluency, narrative recall, drawing reproduction). These differences appear in normative data as additional indicator-level variance and so confound multi-site applications of single-site norms. The CCAB administers every test through an identical digital interface across all sites: stimulus presentation rate, speech intensity is adjusted for hearing loss, while prosody, response timing, and automated scoring rules are identical across participants and testing locations. Thus, the bifactor architecture recovered in the parent analysis reflects cognitive variance unconfounded by administration heterogeneity.

Limitations

The CCAB normative sample on which the demographic-residualization model (C-model) and bifactor parameters were trained reflects Northern California recruitment. While the sample is demographically diverse (37% White, 23% Black, 18% Asian, 22% other race/ethnicity in the full sample; 26% with post-college-level educational attainment), it is concentrated in the San Francisco Bay Area and surrounding metropolitan region. The demographic-correction equations that anchor individual interpretation and the factor scores may not generalize unchanged to other regions.

Six limitations qualify the present findings. First, the bifactor specification with EF subsumed in g was empirically motivated by the observed covariance structure but is not the only plausible parameterization. Alternative models—including higher-order g structures, correlated-factors solutions, and exploratory structural equation models (ESEM)—may capture aspects of the data differently. In particular, bifactor models are known to preferentially absorb variance from highly complex or cognitively demanding tasks into the general factor, potentially reducing the residual variance available for executive-specific constructs. Accordingly, the present results should not be interpreted as demonstrating that executive function is identical to g, but rather that executive indicators in this dataset showed high g saturation within the present parameterization.

Second, the present analyses relied on parcel-level indicators rather than modeling all 107 individual variables simultaneously. Parceling was necessary to maintain model identifiability and stable estimation given the large indicator space and the complexity of bifactor estimation. However, parceling can simplify covariance structure, suppress cross-loadings, and potentially exaggerate apparent factor separability. Although parcel construction was based on random within-domain partitioning and replicated across alternative indicator sets and samples, the recovered latent structure should therefore be interpreted primarily as evidence for robust domain-level covariance organization rather than definitive proof of the dimensionality of individual cognitive indicators.

Third, some portion of the recovered factor structure may reflect shared measurement modality in addition to shared cognitive process. For example, lexical and language measures clustered separately from temporal speech measures, and timing-based indicators clustered separately from accuracy-based indicators. This raises the possibility that portions of the LS, SF, and PS factors partly reflect methodological covariance associated with scoring modality rather than purely separable neurocognitive systems. Several observations argue against a purely methodological interpretation—including replication across tasks, theoretically coherent age and vocabulary associations, and stability under battery expansion—but the possibility of residual method variance cannot be excluded.

Fourth, demographic residualization substantially improved model fit, but this should not be interpreted as revealing “true” latent cognitive architecture. Demographic variables such as age, education, vocabulary, race/ethnicity, and computer familiarity are integral components of real-world cognitive structure and may contribute meaningful covariance among cognitive measures. Residualized and unresidualized solutions therefore answer different questions: the raw models characterize covariance structure in the observed population, whereas the residualized models characterize covariance after removal of selected demographic effects. Although the broad bifactor architecture was stable across both estimations, demographic correction altered several inter-factor relationships substantially, most notably the PS×LS covariance.

Finally, the present study establishes psychometric structure rather than clinical utility. Although the recovered factors showed theoretically coherent demographic signatures and appeared interpretable their clinical significance remains to be established at the individual-patient level. Future work will need to evaluate relationships with longitudinal decline, neurodegenerative phenotypes, neuroimaging and fluid biomarkers, functional outcomes, and differential diagnostic sensitivity across neurological and psychiatric conditions.

Implications and Future Directions

The findings reported here have several implications for cognitive assessment research. First, the resolution of separable LS and SF factors suggesting that a single Language composite may include separable components. Conditions that selectively affect lexical access (semantic dementia, AD) versus articulation (PPA-G, Parkinson's-related speech changes) may yield different signal-to-noise properties on a separated versus combined composite. Second, the demographic-residualization finding suggests that indicator-level demographic correction should be considered in factor analytic work, particularly when the goal is to characterize underlying cognitive architecture in individual patients rather than population-level comparisons. This dual-estimation methodology generalizes beyond the CCAB and could be applied productively to existing data from the UDS-3, NIH Toolbox, WAIS-IV, or other major batteries to test whether their reported factor structures are demographically robust.

Conclusion

Comprehensive computerized neuropsychological assessment generated 107 indicators and recovered a five-construct cognitive architecture comprising a general factor with strong executive function saturation and four separable specific factors: processing speed, memory, lexical/story processing, and speech fluency. The architecture was robust to demographic residualization, indeed, demographic correction at the indicator level substantially improved model fit, and replicated under expanded indicator coverage in a robustness subsample. Lexical/Story and Speech Fluency emerged as distinct latent factors, formalizing a psychometric content-versus-timing dissociation in language production that has long been theoretically established but inaccessible to traditional neuropsychological batteries. The factor solution recovered theoretically distinct cognitive dimensions with appropriate fluid-versus-crystallized signatures, and yields specific factors that were separable by conventional discriminant validity criteria. Comprehensive scoring, made feasible by computerized administration and comprehensive measurement, expanded the quantifiable cognitive phenotype and revealed factor structures not typically observed in traditional neuropsychological batteries.

Declarations

Funding. This work was supported by the National Institute on Aging (NIA) Small Business Innovation Research grant R44AG097322. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests. All authors are employees of Neurobehavioral Systems, Inc., the developer of the California Cognitive Assessment Battery.

Ethics approval and consent to participate. All study procedures were approved by the WCG Institutional Review Board (WIRB protocol 20201196). All participants provided written informed consent prior to participation, in accordance with the Declaration of Helsinki.

Clinical trial registration. ClinicalTrials.gov identifier NCT04800588.

Data availability. The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions. All authors contributed to the study conception, design, analysis, and interpretation, and all authors read and approved the final manuscript.

References

Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment, 22*(1), 121–130. https://doi.org/10.1037/a0017767

Bondi, M. W., Edmonds, E. C., Jak, A. J., Clark, L. R., Delano-Wood, L., McDonald, C. R., Nation, D. A., Libon, D. J., Au, R., Galasko, D., & Salmon, D. P. (2014). Neuropsychological criteria for mild cognitive impairment improves diagnostic precision, biomarker associations, and progression rates. *Journal of Alzheimer's Disease, 42*(1), 275–289. https://doi.org/10.3233/JAD-140276

Collinson, R., Evans, S., Wheeler, M., Brechin, D., Moffitt, J., Hill, G., & Muncer, S. (2017). Confirmatory factor analysis of WAIS-IV in a clinical sample: Examining a bi-factor model. *Journal of Intelligence, 5*(1), 2. https://doi.org/10.3390/jintelligence5010002

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Gaynor, L. S., Lopez, F. V., Van Hulle, C. A., Li, C., Vasunilashorn, S. M., Andrews, S. J., Simone, S. M., & Mungas, D. M. (2025). Measurement equivalence of the UDS version 2.0 and 3.0 neuropsychological batteries. *Alzheimer's & Dementia, 21*(9), e70720. https://doi.org/10.1002/alz.70720

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research, 48*(5), 639–662. https://doi.org/10.1080/00273171.2013.804398

Gross, A. L., Khobragade, P. Y., Meijer, E., & Saxton, J. A. (2020). Measurement and structure of cognition in the Longitudinal Aging Study in India—Diagnostic Assessment of Dementia. *Innovation in Aging, 4*(Suppl 1), 690. https://doi.org/10.1093/geroni/igaa057.2402

Hackett, K., Giovannetti, T., et al. (2024). Psychometric properties of the NIH Toolbox Cognition Battery composites in older adults at risk for Alzheimer's disease and related dementias: A systematic review. *Alzheimer's & Dementia*. https://doi.org/10.1002/alz.13863

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: A unique clinicopathological syndrome. *The Lancet Neurology, 6*(11), 1004–1014. https://doi.org/10.1016/S1474-4422(07)70266-1

Kiselica, A. M., Webber, T. A., & Benge, J. F. (2020). The Uniform Dataset 3.0 neuropsychological battery: Factor structure, invariance testing, and demographically adjusted factor score calculation. *Journal of the International Neuropsychological Society, 26*(6), 576–586. https://doi.org/10.1017/S135561772000003X

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Mayer, M. (2023). *missRanger: Fast imputation of missing values* [R package version 2.4.0]. https://CRAN.R-project.org/package=missRanger

Matusz, E. F., Fiala, J., Kiselica, A. M., Rosselli, M., Armstrong, M. J., Holgerson, A. A., Levy, S.-A., Arias, F., Vélez-Uribe, I., Duara, R., Curiel Cid, R. E., Loewenstein, D. A., Smith, G. E., Marsiske, M., & Asken, B. M. (2025). Cognitive factor structure of the NACC UDS-3 neuropsychological battery across ethno-racial, linguistic, and cognitive status groups. *The Clinical Neuropsychologist*. Advance online publication. https://doi.org/10.1080/13854046.2025.2576154

Mesulam, M.-M., Wieneke, C., Thompson, C., Rogalski, E., & Weintraub, S. (2014). Quantitative classification of primary progressive aphasia at early and mild impairment stages. *Brain, 137*(4), 1176–1192. https://doi.org/10.1093/brain/awu024

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Mungas, D., Heaton, R., Tulsky, D., Zelazo, P. D., Slotkin, J., Blitz, D., Lai, J.-S., & Gershon, R. (2014). Factor structure, convergent validity, and discriminant validity of the NIH Toolbox Cognitive Health Battery (NIHTB-CHB) in adults. *Journal of the International Neuropsychological Society, 20*(6), 579–587. https://doi.org/10.1017/S1355617714000307

Nelson, J. M., Canivez, G. L., & Watkins, M. W. (2013). Structural and incremental validity of the Wechsler Adult Intelligence Scale—Fourth Edition with a clinical sample. *Psychological Assessment, 25*(2), 618–630. https://doi.org/10.1037/a0032086

Nelson, H. E. (1982). *National Adult Reading Test (NART): Test manual*. NFER-Nelson.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Roberts, R., & Knopman, D. S. (2013). Classification and epidemiology of MCI. *Clinics in Geriatric Medicine, 29*(4), 753–772. https://doi.org/10.1016/j.cger.2013.07.003

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Rose, S., Gergoire, A., Pal, S., et al. (2025). Evaluating the factor structure and construct validity of the NIH Toolbox in older adults, with a focus on cognitive normalcy and amnestic mild cognitive impairment: Considerations for diversity, including insights from persons over 85 years of age and Black older Americans. *Journal of the International Neuropsychological Society, 31*, 53–58. https://doi.org/10.1017/S1355617724000626

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

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society, 16*(5), 754–760. https://doi.org/10.1017/S1355617710000706

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*(4), 507–514. https://doi.org/10.1007/BF02296192

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Staffaroni, A. M., Asken, B. M., Casaletto, K. B., Fonseca, C., You, M., Rosen, H. J., Boxer, A. L., Elahi, F. M., Kornak, J., Mungas, D., & Kramer, J. H. (2021). Development and validation of the Uniform Data Set (v3.0) executive function composite score (UDS3-EF). *Alzheimer's & Dementia, 17*(4), 574–583. https://doi.org/10.1002/alz.12214

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics, 28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Vannini, P., Hanseeuw, B., Munro, C. E., Amariglio, R. E., Marshall, G. A., Rentz, D. M., Pascual-Leone, A., Johnson, K. A., & Sperling, R. A. (2017). Anosognosia for memory deficits in mild cognitive impairment: Insight into the neural mechanism using functional and molecular imaging. *Neuropsychologia, 99*, 343–349. https://doi.org/10.1016/j.neuropsychologia.2017.04.002

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV)*. The Psychological Corporation.

Woods, D. L., Pebler, P., Johnson, D. K., Herron, T., Hall, K., Blank, M., Geraci, K., Williams, G., Chok, J., Lwi, S., Curran, B., Schendel, K., Spinelli, M., & Baldo, J. (2024). The California Cognitive Assessment Battery (CCAB). *Frontiers in Human Neuroscience, 17*, 1305529. https://doi.org/10.3389/fnhum.2023.1305529

Woods, D. L., Hall, K., Jaramillo, I., Blank, M., Geraci, K., Pebler, P., Cole, M., & Johnson, D. K. (2026). *CCAB factor analysis reveals age-related cognitive dedifferentiation* [CCAB Technical Report]. Neurobehavioral Systems, Inc. www.ccabresearch.com

Weintraub, S., Besser, L., Dodge, H. H., Teylan, M., Ferris, S., Goldstein, F. C., Giordani, B., Kramer, J., Loewenstein, D., Marson, D., Mungas, D., Salmon, D., Welsh-Bohmer, K., Zhou, X.-H., Shirk, S. D., Atri, A., Kukull, W. A., Phelps, C., & Morris, J. C. (2018). Version 3 of the Alzheimer Disease Centers' neuropsychological test battery in the Uniform Data Set (UDS). *Alzheimer Disease & Associated Disorders, 32*(1), 10–17. https://doi.org/10.1097/WAD.0000000000000223