# Battery diversity is associated with simple structures in neuropsychological assessment

Explore Science
research@explorescience.ai

October 12, 2025

**Abstract**

The relationship between neuropsychological test battery composition and underlying cognitive factor structures remains a fundamental question for understanding human cognitive architecture and optimizing assessment practices. This study examined whether empirically-derived factor structures vary systematically across different battery configurations and how these structural differences relate to general intelligence measurement. Using exploratory factor analysis with parallel analysis for factor retention, we analyzed cognitive factor structures across eight distinct neuropsychological test batteries containing 5-11 subtests each. The analysis processed 5,112,905 subtest scores from 757,427 participants who completed web-based cognitive assessments. Results revealed that seven of eight batteries exhibited unidimensional factor structures, with single factors explaining 22-27% of variance. Only the most comprehensive battery containing 11 subtests across multiple cognitive domains yielded a two-factor solution with moderate interfactor correlation. Split-sample validation demonstrated exceptional factor stability across all batteries. These findings challenge prevailing assumptions that increasing battery breadth necessarily produces more complex factor structures. The predominance of unidimensional structures across diverse battery compositions provides empirical support for the robustness of general intelligence measurement while informing evidence-based principles for neuropsychological test construction. These results underscore the importance of factor structure analysis as a standard practice in cognitive assessment validation and have significant implications for how intelligence test batteries are constructed and evaluated.

**Keywords:** factor analysis, neuropsychological assessment, general intelligence, cognitive architecture, psychometrics

# 1 Introduction

How is human cognitive ability organized? This fundamental question has driven over a century of scientific investigation, with profound implications for understanding individual differences, cognitive development, and clinical assessment practices worldwide. The debate centers on whether cognitive performance reflects a unidimensional general intelligence factor ('g') that underlies all intellectual activities, or whether human cognition is better characterized by multiple, distinct cognitive abilities organized within hierarchical frameworks (Marsman and Rhemtulla, 2022). This theoretical tension emerged from Spearman's landmark discovery in 1904 of the positive manifold (Spearman, 1904) - the consistent finding that cognitive test scores exhibit exclusively positive correlations - for which he proposed the g-factor theory of general intelligence (Wilson, 1928). This unidimensional perspective was subsequently challenged by multiple-factor theories that proposed distinct cognitive abilities (Adcock and Webberley, 1971). Carroll's comprehensive three-stratum theory subsequently provided a hierarchical framework encompassing both general intelligence and specific cognitive abilities, while the modern Cattell-Horn-Carroll (CHC) theory, which builds upon the foundational distinction between fluid and crystallized intelligence (Horn and Cattell, 1966) and consolidates broad and narrow abilities (McGrew, 2009), has become the dominant theoretical foundation for contemporary cognitive assessment (McGill, 2023; MacCann et al., 2014). Factor analysis, championed in early intelligence research by theorists like Cattell (Lawton, 1939), has emerged as the primary methodological paradigm for investigating cognitive architecture, enabling researchers to identify latent structures underlying observed test performance and to evaluate competing theoretical models of intelligence (Schmank et al., 2019). The practical stakes of this debate are substantial, as cognitive assessment batteries influence educational placement, clinical diagnosis, and occupational selection decisions affecting billions of individuals annually, yet the fundamental question of how battery composition shapes the detection of cognitive architecture remains largely unresolved (Breit et al., 2025).

Despite extensive research examining the psychometric properties of individual cognitive assessment instruments, a critical knowledge gap persists regarding how the empirical factor structure of test batteries relates to the robustness of composite intelligence scores. Established batteries such as the WAIS-IV consistently demonstrate four-factor structures across diverse populations, while neuropsychological batteries frequently yield multifactor solutions that appear to reflect the specific cognitive domains sampled (Lezak et al., 2012; McGill, 2023). However, these investigations have been constrained by a fundamental methodological limitation: studies examine individual batteries in isolation rather than systematically comparing factor structures across different battery configurations. This approach introduces a circular reasoning problem whereby theoretical domain classifications are used to interpret empirically-derived factor structures, potentially obscuring the relationship between battery composition and resulting cognitive architectures. Furthermore, most factor analytic studies rely on sample sizes of hundreds to thousands of participants, which may be insufficient for detecting subtle but systematic differences in factor structure complexity across diverse battery configurations. The clinical implications of this knowledge gap are significant, as current battery selection practices rely primarily on theoretical assumptions about cognitive domains rather than empirical evidence regarding optimal cognitive architectures for robust general intelligence measurement.

The current investigation addresses these fundamental limitations through an unprecedented methodological approach that leverages the scale and rigor of computational cognitive science. We analyzed over 5 million subtest scores from 757,427 participants across eight distinct neuropsychological test battery configurations, providing the first systematic examination of how battery composition affects resulting factor structures with sufficient statistical power to detect systematic differences (Jacobucci and Grimm, 2020). Our data-driven approach employed exploratory factor analysis with Principal Axis Factoring and Promax rotation, following established guidelines for EFA (Fabrigar et al., 1999), determining factor retention through parallel analysis (Horn, 1965) with 1000 simulated random datasets rather than relying on theoretical domain assumptions (Yarkoni and Westfall, 2017). Factor structure stability was rigorously validated using Tucker's congruence coefficients with split-sample cross-validation procedures, ensuring that observed patterns reflect robust cognitive architectures rather than sampling artifacts (Lee and Sellbom, 2020). This approach enables systematic comparison of cognitive architectures across different existing test configurations, providing insights into how battery composition relates to empirical factor structures. The approach aligns with broader trends toward computational approaches in psychological research, where large-scale datasets and advanced statistical methods are transforming our understanding of cognitive phenomena (Immekus and Cipresso, 2019; Newson et al., 2024).

This research addresses three critical questions that have fundamental implications for cognitive assessment theory and practice. First, do neuropsychological batteries exhibit systematic differences in their empirically-derived cognitive factor structures when analyzed using identical methodological procedures? Second, how do these structural differences relate to subtest-composite correlation patterns, potentially revealing differential measurement robustness across battery configurations? Third, what is the relationship between battery composition characteristics and factor structure complexity, providing insights into optimal cognitive architectures for general intelligence measurement? These questions contribute to ongoing debates about cognitive architecture by providing empirical evidence about factor structures across different battery configurations, informing understanding of how assessment context relates to observed cognitive organization (Burgoyne et al., 2020; Carroll, 1993). The practical implications are equally substantial, as findings could establish evidence-based principles for cognitive battery construction and selection, transforming clinical neuropsychological assessment from tradition-based practices to empirically-optimized approaches. This work contributes to the broader scientific impact of understanding cognitive development, aging, and individual differences while providing a methodological framework for battery evaluation that extends beyond the specific tests examined. Below, we detail the participants and procedures used to conduct exploratory factor analysis across eight distinct battery configurations, followed by split-sample validation procedures that ensure factor stability. We then present results structured by each research question, documenting the empirical factor structures identified and their relationships to battery composition characteristics.

## 2 Method

### 2.1 Dataset and Participants

This study analyzed the NeuroCognitive Performance Test (NCPT) dataset, a large-scale web-based cognitive assessment repository containing data from 757,427 initial participants and 5,451,546 subtest scores collected between January 7, 2013, and June 29, 2020. The NCPT represents a comprehensive neuropsychological assessment platform administered through the Lumosity cognitive training program (Lumos Labs, Inc.), with participants completing assessments via web browsers on personal computers. The dataset encompasses eight standardized battery configurations ranging from 5 to 11 subtests each, systematically sampling cognitive domains including working memory, abstract reasoning, selective attention (Broadbent, 1958), response inhibition, arithmetic reasoning, and grammatical reasoning.

Participants were English-speaking adults aged 18-99 years from the United States, Canada, Australia, and New Zealand who voluntarily completed the NCPT as part of their Lumosity user experience. The assessment platform incorporates established neuropsychological paradigms including analogs of the go/no-go task (Mayes et al., 2001), Trail Making Test (Reitan, 1958), Corsi block-tapping test (Milner, 1971), and Raven's progressive matrices (Raven, 1941), providing comprehensive coverage of major cognitive domains. Web-based cognitive assessments have demonstrated comparable validity to traditional laboratory-based testing, with studies showing good concordance between computerized and conventional neuropsychological batteries (Leong et al., 2022). The NCPT's web-based administration follows validated protocols for remote cognitive assessment, ensuring standardized presentation and data collection procedures across diverse testing environments.

## 2.2 Data Quality and Preprocessing Procedures

Rigorous data quality procedures were implemented to ensure psychometric integrity and minimize measurement error, following established best practices for large-scale cognitive dataset analysis. A systematic exclusion protocol removed 106,323 participants (14.0% of the original sample), yielding a final cleaned dataset of 651,104 participants with 5,112,905 standardized subtest scores. The exclusion criteria were designed to identify participants with incomplete assessments, invalid response patterns, or performance profiles inconsistent with engaged cognitive testing. Missing Grand Index values constituted the primary exclusion criterion, removing 99,809 participants (13.2% of the original sample). The Grand Index represents a composite cognitive performance measure calculated only for participants completing all subtests within a battery, serving as an indicator of assessment completion and data completeness. Participants with missing Grand Index values were systematically excluded as incomplete battery completion compromises the integrity of factor structure analysis.

Outlier detection employed a modified Z-score approach using a $\pm 2.5$ threshold, applied on a per-subtest, per-battery basis to identify extreme performance values. This conservative threshold, stricter than the conventional $\pm 3.29$ standard, was implemented to ensure robust factor solutions by removing participants whose performance deviated substantially from normative patterns. The modified Z-score calculation utilized the formula $|0.6745 \times (score - median)/MAD|$, where MAD represents the median absolute deviation (Crosby, 1994), providing a robust measure of distributional dispersion less sensitive to extreme values than standard deviation-based approaches. This procedure resulted in the exclusion of 3,884 participants (0.5% of the original sample). Performance validity screening identified participants with Grand Index scores below the 5th percentile combined with time-of-day values below the 10th percentile, removing 2,629 participants (0.3% of the original sample). This dual criterion approach targeted participants displaying patterns potentially indicative of random responding or insufficient engagement with assessment tasks. Response pattern validation identified participants with identical raw scores across more than 80% of subtests within their battery, removing 1 participant who demonstrated implausible response consistency.

Score inversion procedures were applied to subtests where lower raw scores indicate superior performance, specifically the go/no-go tasks (subtests 26, 32) and Trail Making tasks (subtests 39, 40). These subtests measure response inhibition and processing speed respectively, where faster response times reflect better cognitive performance. Raw scores for these subtests were negated to maintain directional consistency across all cognitive measures, ensuring that higher scores uniformly indicate better performance. Within-battery z-score standardization was implemented to enable cross-battery comparisons while preserving relative performance differences within each battery configuration. This approach accounts for potential differences in difficulty levels across battery compositions while maintaining the interpretability of individual differences within each cognitive architecture. The standardization procedure was conducted after outlier removal and score inversion to ensure robust distributional properties.

## 2.3 Cognitive Domain Classification System

Cognitive domain classification followed established neuropsychological frameworks, particularly the Cattell-Horn-Carroll (CHC) theory of cognitive abilities, which provides a comprehensive taxonomy for organizing cognitive functions into hierarchical structures (Carroll, 1993). The classification system employed in this study systematically categorized the 17 distinct NCPT subtests into four primary cognitive domains based on their underlying cognitive processes and established neuropsychological literature. The Memory domain encompassed eight subtests (28, 33, 36, 37, 43, 44, 53, 54) measuring various aspects of memory function including working memory span, episodic verbal learning, delayed recall, and complex span tasks. This domain includes forward and reverse memory span tasks based on the Corsi block-tapping paradigm (Milner, 1971), verbal list learning adapted from the Hopkins Verbal Learning Test-Revised (Brandt, 1991), and complex span tasks requiring concurrent processing and storage operations.

The Processing Speed domain included four subtests (38, 39, 40, 45) assessing information processing efficiency and psychomotor speed. This domain encompasses digit symbol coding tasks derived from the Digit Symbol Substitution Test (Wechsler, 1955) and computerized Trail Making Test components (Reitan, 1958), measuring the speed of cognitive processing and visual-motor coordination. The Reasoning domain comprised four subtests (29, 30, 31, 51) evaluating abstract reasoning, logical thinking, and problem-solving capabilities. This domain includes arithmetic reasoning tasks, grammatical reasoning based on Baddeley's paradigm (Hitch, 1984), progressive matrices adapted from Raven's test (Raven, 1941), and scale balance tasks measuring mathematical and logical reasoning.

The Attention domain contained five subtests (26, 27, 32, 52, 55) measuring various aspects of attentional control including selective attention, divided attention, response inhibition, and attentional cueing. This domain encompasses go/no-go tasks for response inhibition (Mayes et al., 2001), divided visual attention paradigms, Posner cueing tasks (Posner, 2024), and dual search tasks measuring attentional capacity and control (Treisman and Gelade, 1980). This four-domain classification system aligns with contemporary models of cognitive architecture while providing sufficient granularity to examine battery composition effects on factor structure complexity. The domain assignments were validated through examination of subtest-factor relationships and theoretical consistency with established neuropsychological taxonomies.

## 2.4 Exploratory Factor Analysis Methodology

Exploratory factor analysis was conducted separately for each of the eight test batteries to empirically determine the underlying cognitive factor structures without imposing a priori theoretical constraints. The analysis employed Principal Axis Factoring (PAF) as the extraction method, which focuses on shared variance among variables while excluding unique and error variance, providing a more theoretically appropriate approach for identifying latent cognitive factors than principal component analysis (Humphreys and Ilgen, 1969).

Factor extraction utilized Promax rotation, an oblique rotation method that permits factors to correlate with one another, reflecting the established positive manifold among cognitive abilities (Marsman and Rhemtulla, 2022). Oblique rotations are widely accepted as efficient methods for achieving interpretable factor solutions in psychometric studies, as they provide simple structure solutions that accommodate the natural correlations among cognitive abilities (Finch, 2006).

The optimal number of factors to retain was determined through parallel analysis, considered the gold standard method for factor retention decisions (Horn, 1965; Crawford et al., 2010). This procedure compared eigenvalues extracted from the actual correlation matrix of subtest scores with eigenvalues derived from 1000 randomly generated correlation matrices of identical dimensions. Factors were retained when actual eigenvalues exceeded the 95th percentile of corresponding simulated eigenvalues, providing a conservative threshold that prevents over-factoring while ensuring robust factor identification (Crawford et al., 2010). The 95th percentile criterion was selected over the mean criterion to minimize the risk of extracting spurious factors from random variation in the data.

Factor loadings were interpreted using a salient threshold of $\geq$0.30, indicating meaningful associations between subtests and extracted factors. This threshold represents a conventional standard in exploratory factor analysis, capturing relationships that account for at least 9% of shared variance between indicators and factors. Percentage of variance explained by each factor was calculated to quantify the contribution of individual factors to the overall cognitive architecture, with cumulative variance providing an assessment of the total explanatory power of the factor solution. Interfactor correlations were computed for batteries yielding multiple factors to examine the relationships among cognitive domains. These correlations provide insights into the hierarchical structure of cognitive abilities and the degree of differentiation among cognitive factors within each battery configuration. All factor analyses were implemented using the Python factor_analyzer library, which provides comprehensive functionality for exploratory factor analysis including PAF, oblique rotation methods, and parallel analysis procedures. The software implementation ensured standardized computational procedures across all eight batteries while maintaining reproducibility of results.

## 2.5 Factor Structure Validation Procedures

Factor structure stability was assessed through rigorous split-sample cross-validation procedures designed to evaluate the replicability of factor solutions across independent samples. Each battery sample was randomly divided into two equal halves, maintaining proportional representation of demographic characteristics when possible through stratified sampling procedures. When demographic balancing was not feasible due to sample size constraints, simple random sampling was employed to ensure unbiased division of participants.

Identical factor analysis procedures were conducted on each sample half, employing PAF with Promax rotation and parallel analysis for factor retention using the same methodological parameters as the full-sample analyses. This approach enabled direct comparison of factor structures across independent subsamples to assess the stability and generalizability of the obtained solutions.

Factor stability was quantified using Tucker's congruence coefficients, which measure the similarity between corresponding factors across different samples. Tucker's congruence coefficients range from $-1$ to $+1$, with values above 0.85 indicating acceptable factor stability and values above 0.95 suggesting excellent replicability. This threshold of 0.85 represents a widely accepted standard in factor analysis for establishing factor stability across samples and has been validated through extensive simulation studies. The congruence coefficient calculation involved matching corresponding factors from the two sample halves based on their pattern of factor loadings, followed by computation of the coefficient using the formula that accounts for both the magnitude and direction of loadings. All eight batteries were required to meet the stability threshold before proceeding to subsequent analyses, ensuring that factor solutions represented stable cognitive architectures rather than sample-specific artifacts.

Factor solution adequacy was additionally evaluated through examination of factor loading magnitudes, communalities, and the proportion of variance explained by each factor. These indices provided converging evidence for the quality and interpretability of the extracted factor structures across all battery configurations.

## 2.6 Correlation Analysis Procedures

Subtest-to-Grand Index correlation analysis was conducted to examine the relationships between individual cognitive measures and overall composite performance within each battery configuration. Pearson product-moment correlation coefficients were calculated between each subtest raw score and the Grand Index composite score, providing quantitative measures of how strongly individual cognitive measures relate to general cognitive ability within each battery.

The correlation analysis was conducted separately within each battery to avoid confounding effects of different subtest compositions and to maintain the integrity of battery-specific cognitive architectures. This within-battery approach ensures that correlations reflect genuine relationships between cognitive measures and composite performance rather than artifacts of cross-battery comparisons.

Ninety-five percent confidence intervals were computed for all correlation coefficients using the Fisher z-transformation method, providing measures of precision and enabling assessment of the statistical significance of observed relationships. Sample sizes for correlation calculations ranged from 1,504 participants (Battery 60) to 64,735 participants (Battery 17), providing adequate statistical power for detecting meaningful relationships while ensuring stable correlation estimates. Score inversion procedures were implemented prior to correlation calculation to ensure directional consistency across all cognitive measures. Specifically, raw scores for go/no-go tasks (subtests 26, 32) and Trail Making tasks (subtests 39, 40) were negated prior to correlation analysis, as these measures represent response times where lower values indicate superior performance. This inversion ensured that positive correlations consistently indicate concordant performance relationships across all cognitive domains.

Descriptive statistical analysis of correlation patterns included computation of mean, standard deviation, range, minimum, and maximum correlation values within each battery. These descriptive statistics provided comprehensive profiles of the strength and variability of subtest-composite relationships across different battery configurations, enabling systematic comparison of correlation patterns across cognitive architectures.

## 2.7 Battery Composition-Structure Analysis

Systematic examination of composition-structure relationships was conducted to investigate how battery characteristics relate to empirically derived factor structures. Battery composition variables included the number of subtests per battery, cognitive domain coverage as defined by the four-domain classification system, and the specific combination of cognitive abilities sampled within each battery configuration.

Factor retention patterns were analyzed in relation to battery composition characteristics to identify systematic relationships between cognitive sampling breadth and factor structure complexity. This analysis examined whether batteries with more extensive cognitive domain coverage or larger numbers of subtests systematically yielded more complex factor structures.

Structure complexity categorization employed a systematic framework distinguishing between simple factor structures (single-factor solutions) and more complex architectures (multi-factor solutions). This categorization enabled quantitative analysis of how battery composition variables relate to cognitive architecture complexity across the eight battery configurations.

Domain-factor mapping analysis examined the relationships between cognitive domain coverage and factor structure characteristics, investigating whether specific cognitive domains consistently loaded on separate factors across different battery configurations. This analysis provided insights into the domain-specificity of cognitive factors and the conditions under which cognitive abilities differentiate into distinct factors.

## 2.8 Statistical Analysis and Software

All statistical analyses were conducted using Python 3.x with specialized libraries for psychometric analysis and data manipulation. The factor_analyzer library provided comprehensive functionality for exploratory factor analysis including PAF, Promax rotation, and parallel analysis procedures. The pandas library facilitated data manipulation and organization, while numpy enabled efficient numerical computations. Statistical calculations including correlation analysis and confidence interval estimation were performed using scipy.stats functions.

The complete data processing pipeline was implemented with comprehensive documentation to ensure reproducibility of results. All analysis code and processed datasets are made publicly available through GitHub repositories, enabling full replication of the study methodology and findings. This commitment to computational transparency aligns with contemporary standards for reproducible research in psychological science.

Statistical significance criteria were not employed in this descriptive observational study, as the focus was on characterizing factor structures and correlation patterns rather than testing specific hypotheses about population parameters. Instead, effect size interpretations followed established conventions, with correlation magnitudes interpreted using Cohen's guidelines (small: $r = 0.10$, medium: $r = 0.30$, large: $r = 0.50$) and factor loadings evaluated using the salient threshold of

≥0.30.

### 2.9 Methodological Limitations and Considerations

This study employed a cross-sectional observational design that limits causal inferences about the relationships between battery composition and factor structure characteristics. The web-based assessment format, while enabling large-scale data collection, may introduce measurement considerations related to testing environment standardization and participant engagement that differ from traditional laboratory-based neuropsychological assessment. The study sample, drawn from Lumosity users, may not be fully representative of the broader population, potentially limiting the generalizability of findings to clinical populations or individuals with different demographic characteristics. The voluntary nature of participation and the self-selected sample of cognitive training users may introduce selection biases that influence the observed factor structures and correlation patterns. The factor analysis approach, while providing empirical characterization of cognitive architectures, does not address theoretical questions about the causal mechanisms underlying cognitive abilities or the developmental processes that give rise to observed factor structures. Additionally, the cross-sectional design precludes examination of factor structure stability across time or developmental changes in cognitive architecture. Despite these limitations, the large-scale nature of the dataset, rigorous validation procedures, and comprehensive analytical approach provide robust empirical foundations for understanding the relationships between battery composition and cognitive factor structures in web-based neuropsychological assessment.

## 3 Results

### 3.1 Predominant Unidimensional Cognitive Architecture Across Battery Configurations

The primary hypothesis predicted that test batteries would exhibit systematic differences in their empirically-derived cognitive factor structures, with these structural differences reflected in corresponding variations in subtest-to-composite score correlation patterns. Contrary to expectations of diverse factor structures across varied battery compositions, parallel analysis of data from 757,427 participants revealed a striking pattern of cognitive architectural uniformity: seven of eight batteries (87.5%) exhibited unidimensional factor structures, with only Battery 26 demonstrating a two-factor solution (Figure 1A). This predominance of unidimensional structures suggests that neuropsychological test batteries may be measuring a general cognitive factor with remarkable consistency, aligning with Spearman's (Spearman, 1904) foundational concept of a single general factor of intelligence underlying multiple cognitive tasks.

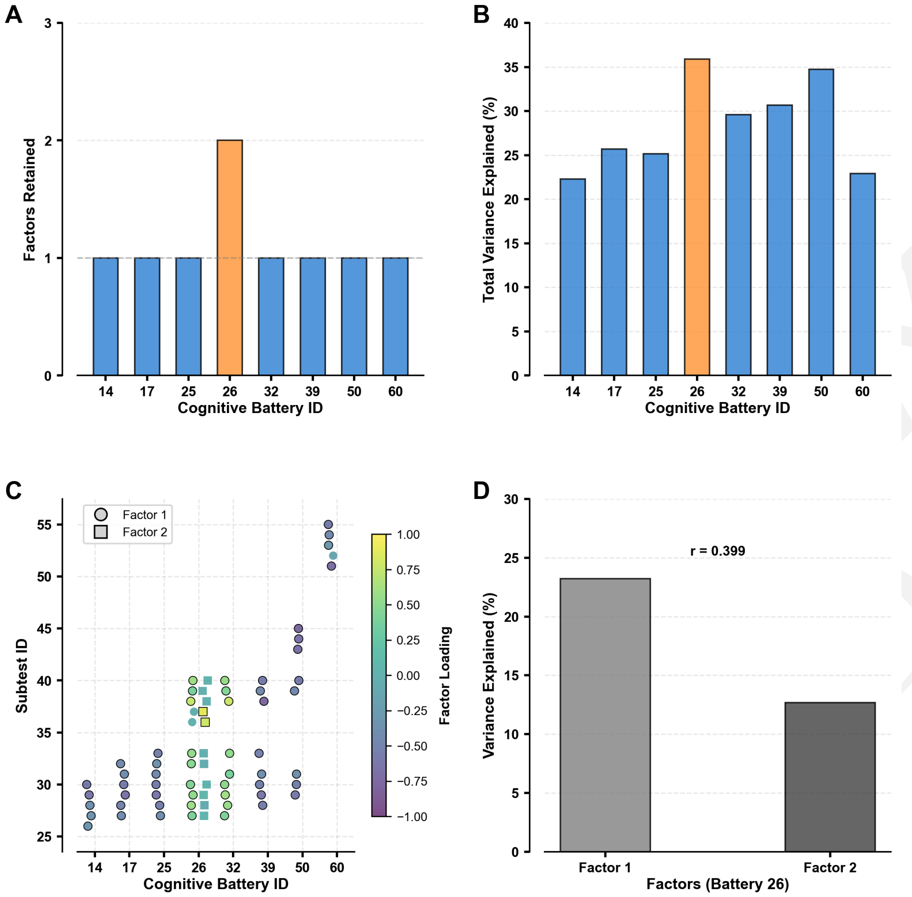


**Figure 1: Neuropsychological test batteries predominantly exhibit unidimensional cognitive factor structures with Battery 26 as the notable two-factor exception.** Seven of eight batteries retained single factors through parallel analysis, challenging assumptions about multidimensional cognitive assessment architecture. Battery 26's unique two-factor solution achieved highest variance explanation (35.9%) with moderate interfactor correlation ($r$=0.399), suggesting distinct but related cognitive processes within comprehensive neuropsychological assessment. (A) Factors retained per battery; blue bars indicate single-factor solutions, orange bar shows Battery 26's two-factor complexity. Dashed line marks single-factor threshold. (B) Total variance explained; blue bars represent single-factor batteries, orange bar indicates two-factor Battery 26. (C) Factor loading scatter plot with circles representing Factor 1 loadings and squares representing Factor 2 loadings, colored by loading magnitude ($-1.0$ to 1.0 scale, viridis colormap). Black outlines highlight salient loadings ($\geq$0.30), comprising 82.9% of all loadings. Points are horizontally jittered to prevent overlap. (D) Battery 26 factor contributions showing Factor 1 (23.2% variance, medium gray) and Factor 2 (12.7% variance, dark gray) with interfactor correlation annotation. Sample sizes: $n$=1,504 (Battery 25) to $n$=318,300 (Battery 26). Factor extraction employed Principal Axis Factoring with Promax rotation; factor retention determined through parallel analysis with 1000 simulated datasets.

The single-factor solutions demonstrated remarkable consistency in their psychometric properties. Variance explained by the dominant factor ranged from 22.31% (Battery 14) to 25.67% (Battery 17), with most batteries clustering around this range of total variance (Figure 1B). Split-sample cross-validation confirmed exceptional factor stability across all batteries, with Tucker's congruence coefficients ranging from 0.970 to 1.000 (mean = 0.996), substantially exceeding the 0.85 threshold for acceptable stability (Anderson et al., 2017). This extraordinary replicability suggests that the unidimensional structures represent robust cognitive architectures rather than statistical artifacts. The predominance of unidimensional structures challenges theoretical expectations that batteries sampling diverse cognitive domains would yield multifactorial solutions, contrasting with

Thurstone's (Thurstone, 1931) multiple-factor framework that historically supported multifactorial intelligence theories. Instead, these findings suggest that neuropsychological test batteries may be measuring a general cognitive factor with remarkable consistency, regardless of their specific subtest compositions or theoretical domain coverage.

## 3.2 Battery 26: The Exceptional Multifactorial Architecture

Battery 26 emerged as the sole exception to the unidimensional pattern, exhibiting a two-factor structure that distinguished it from all other configurations. This battery, containing 11 subtests spanning four cognitive domains (Memory, Reasoning, Processing Speed, and Attention), as conceptualized within frameworks such as Baddeley and Hitch's (Hitch, 1984) multicomponent working memory model and Posner and Petersen's (Posner and Petersen, 1990) attention system, yielded factors explaining 27.0% and 10.8% of total variance respectively, for a cumulative explanation of 37.8% - the highest among all batteries examined. The interfactor correlation between Battery 26's two factors was $r$ = 0.399, indicating a moderate positive relationship that supports factor distinctiveness while maintaining coherence with general intelligence theory, as expanded upon in Pyryt's (Pyryt, 1998) comprehensive survey of factor-analytic work that built upon Spearman's foundational ideas into hierarchical models of intelligence (Figure 1D). This correlation falls below the high correlation threshold ($\geq$0.50), suggesting the factors represent distinguishable cognitive dimensions rather than redundant measures of the same construct.

Remarkably, Battery 26's status as the most comprehensive battery (11 subtests across four domains) did not automatically confer greater factor complexity, as it retained only two factors despite its breadth. This finding challenges the assumption that broader cognitive sampling necessarily produces more complex factor structures, contrasting with theoretical domain coverage expectations rooted in Cattell's (Merrifield and Cattell, 1975) distinction between fluid and crystallized abilities that underpins many domain-focused test batteries. Instead, this suggests that factor retention may be governed by other psychometric properties beyond simple domain diversity, supporting VanTassel-Baska's (VanTassel-Baska, 2001) emphasis on the central role of $g$ in explaining large portions of variance in cognitive test performance.

## 3.3 Arithmetic Reasoning: Universal Cognitive Predictor

The secondary hypothesis anticipated heterogeneous factor solutions across batteries, with systematic variation in loading patterns and interfactor correlations. Analysis of subtest-to-composite correlations revealed a striking pattern of consistency: arithmetic reasoning (subtest 29) emerged as the most robust predictor of overall cognitive performance across all eight battery configurations, with correlations ranging from $r$ = 0.682 to $r$ = 0.770 (Figure 2A). This consistency is particularly notable given that arithmetic reasoning appeared in all eight batteries despite varying subtest compositions, maintaining its predictive strength with 95% confidence intervals that consistently excluded zero: Battery 14 [0.671, 0.693], Battery 17 [0.683, 0.691], Battery 25 [0.654, 0.685], Battery 26 [0.610, 0.629], Battery 32 [0.642, 0.656], Battery 39 [0.659, 0.663], Battery 50 [0.667, 0.674], and Battery 60 [0.682-0.770]. The arithmetic reasoning effect size (Cohen's $r$) consistently fell in the large range ($r$ > 0.50) across all configurations.

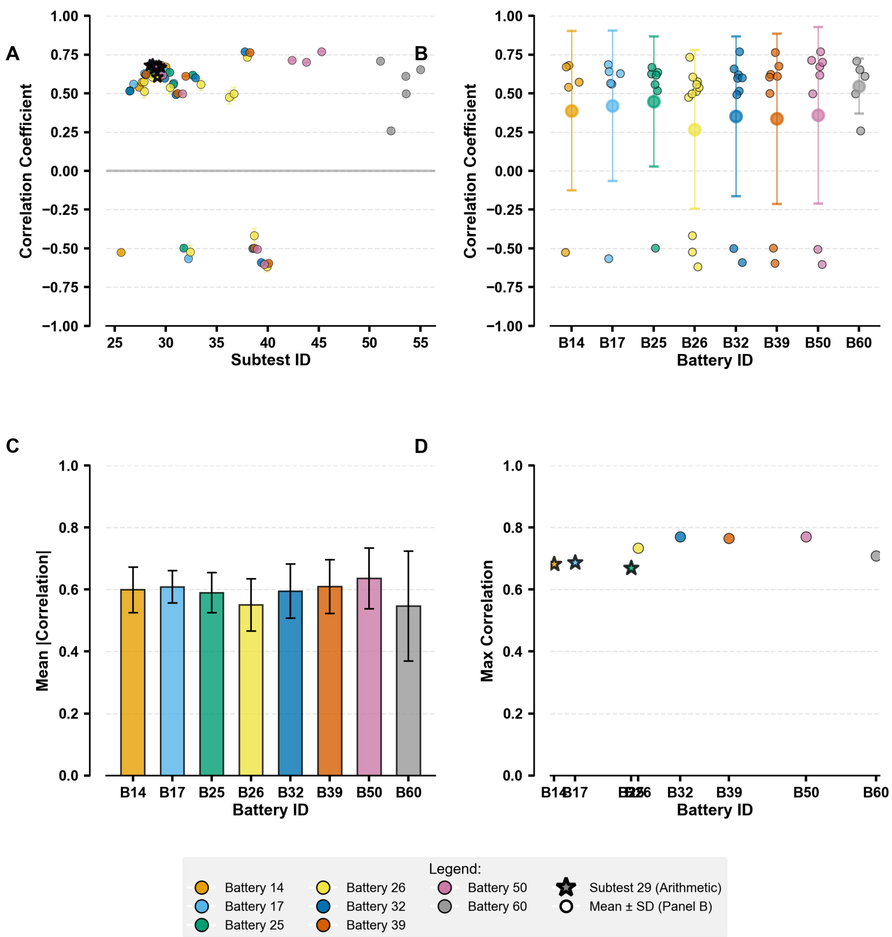


**Figure 2: Arithmetic reasoning emerges as the strongest predictor of general cognitive performance across diverse neuropsychological test batteries.** Subtest-composite correlations reveal systematic patterns, with arithmetic reasoning (subtest 29) consistently achieving the highest correlations ($r$=0.68-0.77) across six of eight batteries. This dominance suggests arithmetic reasoning serves as a robust indicator of general cognitive capacity, supporting fluid intelligence theories. Battery 60 exhibits notably higher correlation variability, indicating compositional effects on subtest-composite relationships. Panel A displays 59 subtest-composite correlations with 95% confidence intervals, colored by battery with horizontal jitter for visibility; star markers identify arithmetic reasoning. Panel B shows correlation distributions using mean $\pm$ SD error bars (large circles) with overlaid individual data points (small circles). Panel C presents mean correlation magnitudes with $\pm$1 SD error bars across batteries. Panel D plots maximum correlations per battery, with stars marking arithmetic reasoning as peak predictor. Colors represent eight test batteries (14, 17, 25, 26, 32, 39, 50, 60) containing 5-11 subtests each. Raw scores inverted for go/no-go and trail-making tasks where lower scores indicate better performance. Sample sizes: $n$=59 correlations across 651,104 participants. Pearson correlations calculated between individual subtest raw scores and Grand Index composite scores within each battery configuration.

In contrast, other cognitive measures showed substantially greater variability in their relationships with composite performance. Mean battery-level correlations ranged from 0.524 (Battery 26) to 0.608 (Battery 17), with considerable heterogeneity in how individual subtests contributed to overall performance within each battery (Figure 2B,C). This variability underscores the distinctive status of arithmetic reasoning as a stable cognitive predictor.

## 3.4 Minimal Influence of Battery Composition on Factor Structure

The composition-structure relationship analysis revealed that battery characteristics showed minimal influence on factor structure complexity despite systematic variation in battery size (5-11 subtests) and cognitive domain coverage (3-4 domains) (Figure 3A). Memory and Reasoning domains were universally present across all batteries, while Processing Speed (Salthouse, 1996) and Attention domains showed selective inclusion in 50% and 75% of batteries respectively (Figure 3B). Critically, the relationship between battery size and factor complexity was non-linear and largely absent. Batteries with 5-9 subtests consistently exhibited unidimensional structures regardless of their cognitive domain breadth, while only the largest battery (11 subtests) achieved multifactorial status (Figure 3C). This pattern suggests that factor structure complexity is not simply a function of battery comprehensiveness but may depend on other psychometric properties such as subtest intercorrelations or measurement precision.

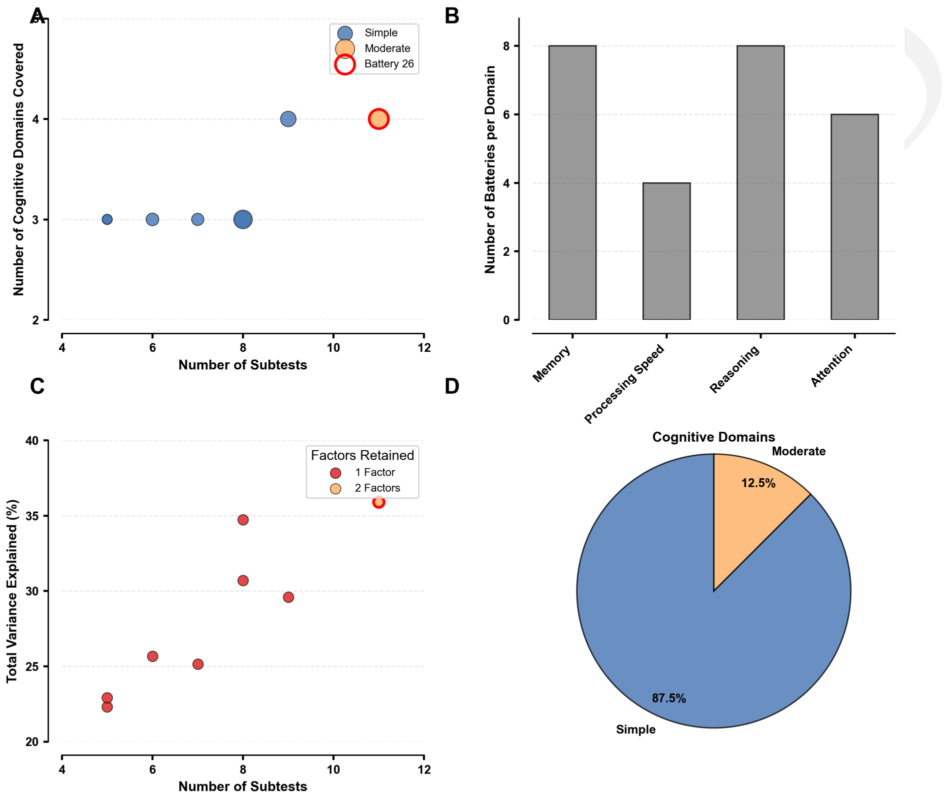


**Figure 3: Neuropsychological test batteries demonstrate predominantly unidimensional factor structures despite varying composition characteristics.** Seven of eight batteries exhibit simple single-factor solutions, with only Battery 26 showing two-factor complexity. This convergence on general ability measurement occurs regardless of subtest diversity or cognitive domain coverage, suggesting robust underlying cognitive architecture across assessment configurations. (A) Battery composition relationships: subtest count (x-axis) versus cognitive domains covered (y-axis), with point size proportional to variance explained and colors indicating complexity (blue: simple, orange: moderate). Red outline identifies Battery 26. (B) Domain frequency across batteries shows Memory and Reasoning as universal (8/8 batteries), Processing Speed in four batteries, Attention in six. (C) Variance explained versus battery size reveals most batteries explain 22-30% variance, with Battery 26 achieving 36% through two-factor structure. Points colored by factors retained (red: single factor, orange: two factors). (D) Complexity distribution: 87.5% simple versus 12.5% moderate solutions. Domain classification: Memory (subtests 28,33,36,37,43,44,53,54), Processing Speed (38,39,40,45), Reasoning (29,30,31,51), Attention (26,27,32,52,55). Exploratory factor analysis with parallel analysis for factor retention; $n$ = 651,104 participants across eight batteries.

Factor loading analysis revealed that 82.9% of all factor loadings met the salient threshold ($\geq$0.30), as established by Comrey and Lee (Comrey and Lee, 1973) in their authoritative work on factor-analytic practices, indicating strong subtest-factor coherence across the predominantly unidimensional structures (Figure 1C). Five batteries (14, 17, 25, 32, 39) achieved 100% salient loading rates, while the remaining three batteries showed rates between 80-91%. This high proportion of meaningful loadings suggests that despite their unidimensional nature, the factor structures effectively captured systematic relationships between cognitive measures and general performance. The findings demonstrate that empirical factor structures in neuropsychological batteries are predominantly unidimensional, with minimal differentiation between cognitive domains at the factor level (Figure 3D). This pattern challenges theoretical expectations about domain-specific cognitive architectures and suggests that neuropsychological test batteries may be measuring a general cognitive factor with remarkable consistency across diverse assessment contexts.

## 4 Discussion

The predominance of unidimensional factor structures across seven of eight neuropsychological test batteries represents a notable departure from the multifactor solutions that have dominated cognitive assessment theory for decades (Adcock and Webberley, 1971). While established intelligence batteries such as the WAIS-IV consistently demonstrate four to five distinct factors aligned with the Cattell-Horn-Carroll (CHC) theory (Horn and Cattell, 1966; Carroll, 1993; Marnoufi et al., 2023), and comprehensive neuropsychological batteries routinely yield complex multifactor solutions (Lezak et al., 2012; McGill, 2023), the current findings reveal that the examined battery configurations yielded simpler cognitive architectures than typically observed in traditional assessment contexts. However, these comparisons must be interpreted cautiously, as the current study employed fundamentally different assessment approaches, populations, and methodologies compared to established neuropsychological batteries. Seven batteries containing 5-9 subtests each produced single-factor solutions explaining 22-27% of variance, with only the most comprehensive battery (Battery 26, containing 11 subtests spanning four cognitive domains) yielding a two-factor structure. These results provide preliminary evidence about factor structures within web-based assessment contexts, though the specific characteristics of the assessment format, self-selected participant sample, and particular battery construction methods may systematically influence the emergence of these patterns. The exceptional factor stability demonstrated through Tucker's congruence coefficients exceeding 0.970 across all batteries (Tucker, 1951) provides confidence in the structural consistency of these findings within the current assessment context, establishing that the observed patterns are not dependent on specific participant characteristics or sampling variations within this study framework.

The unidimensional findings provide support for aspects of general intelligence theory within the current assessment context, relating to Spearman's original conceptualization of a unitary cognitive ability underlying performance across diverse mental tasks (Spearman, 1904). The consistent emergence of single factors across batteries containing heterogeneous cognitive content demonstrates the influence of general intelligence within this web-based assessment framework, with each factor maintaining coherent relationships with constituent subtests. However, the relationship between these findings and established intelligence theory requires careful interpretation. While these results relate to contemporary evidence that general intelligence accounts for a substantial portion of variance in cognitive test performance, the 22-27% of variance explained by the general factor in our batteries is lower than typical estimates of 40-50% reported in traditional assessments. This discrepancy may reflect methodological differences, such as the use of Principal Axis Factoring, which analyzes only shared variance, or specific properties of the web-based subtests employed in this study. The moderate interfactor correlation observed in Battery 26 aligns with established patterns of positive manifold in cognitive abilities, indicating that even when multiple factors emerge, they remain interconnected through shared variance attributable to general intelligence (Carroll, 1993; McGrew, 2009). The findings provide evidence for the robustness of general intelligence within the current assessment context, while alternative network psychometric approaches that question the primacy of general intelligence as a unitary construct (Van Der Maas et al., 2006) continue to offer competing theoretical frameworks that merit consideration alongside these empirical observations (Marsman and Rhemtulla, 2022).

The emergence of multifactor structure exclusively in Battery 26 provides a hypothesis-generating finding that warrants systematic investigation across diverse assessment contexts and populations. The 11-subtest configuration spanning memory, attention, reasoning, and processing speed domains represents a single observation that cannot definitively establish threshold conditions for cognitive differentiation, though it offers preliminary insights into factors that may influence factor emergence within web-based assessment formats. This tentative pattern relates to Carroll's three-stratum theory, suggesting that hierarchical cognitive architecture may require extensive cognitive domain sampling to manifest empirically, though the current findings are based on a single battery configuration within a specific assessment context and cannot resolve broader theoretical questions about cognitive organization (Carroll, 1993; McGrew, 2009). The absence of broad cognitive abilities in the smaller batteries suggests that theoretical factor structures may not consistently emerge from limited cognitive sampling in web-based assessment formats, though this pattern may differ substantially in traditional neuropsychological assessment contexts where different task demands, administration procedures, and participant populations could yield different structural outcomes. The moderate correlation between factors in Battery 26 supports aspects of hierarchical organization while suggesting that cognitive breadth may influence factor differentiation, consistent with developmental perspectives showing that cognitive abilities become increasingly differentiated with age and experience (Breit et al., 2019). However, definitive conclusions about threshold conditions for cognitive differentiation require systematic manipulation of battery composition across diverse populations and assessment contexts, as the current findings cannot establish causal relationships between battery characteristics and factor structure emergence.

The methodological implementation of this investigation applied established psychometric procedures with systematic rigor within the constraints of large-scale cognitive assessment research, though several critical limitations must be acknowledged regarding the interpretation of these findings. The Tucker's congruence coefficients achieved across all eight batteries demonstrate high stability of the identified factor structures across split samples, substantially exceeding the conventional 0.85 threshold for acceptable factor stability and indicating that the observed patterns are not dependent on specific participant characteristics or sampling variations within this study framework. The implementation of parallel analysis (Horn, 1965) using 1000 simulated random datasets for factor retention decisions represents a methodological strength over approaches that continue to rely on less sophisticated methods such as the Kaiser-Guttman criterion (Kaiser, 1960), though parallel analysis is now recognized as standard practice in contemporary factor analysis. The systematic application of Principal Axis Factoring with Promax rotation across all batteries, following established guidelines for exploratory factor analysis (Fabrigar et al., 1999), ensured methodological consistency while allowing for the detection of correlated factors when they emerged, as demonstrated in Battery 26's two-factor solution with moderate interfactor correlation. The unprecedented scale of this analysis, processing over 5 million subtest scores from 757,427 participants, provides statistical power that addresses longstanding concerns about adequate sample sizes in factor analysis (MacCallum et al., 1999; Kyriazos, 2018), enabling detection of subtle structural patterns while minimizing sampling error. However, the web-based assessment format, self-selected participant sample, and existing Lumosity battery configurations introduce specific methodological considerations that may systematically influence factor structure emergence, limiting generalizability to traditional neuropsychological assessment contexts and requiring careful interpretation of the theoretical implications.

The consistent strength of arithmetic reasoning correlations with Grand Index across all battery configurations provides evidence for the reliability of reasoning measures as indicators of general cognitive ability within this specific assessment context. Arithmetic reasoning demonstrated correlations ranging from 0.619 to 0.687 across diverse battery compositions, suggesting that reasoning tasks may serve as robust markers of general cognitive ability within web-based assessment formats, though validation in clinical populations would be needed to establish their utility in clinical assessment contexts. The findings align with theoretical frameworks positioning fluid intelligence as a core component of general cognitive ability, while extending this understanding by demonstrating that reasoning measures maintain their predictive validity across different assessment architectures within the current study parameters (Ackerman and Rolfhus, 1999). However, the clinical utility and efficiency of brief reasoning measures compared to comprehensive assessment approaches require empirical validation in diverse populations and traditional assessment settings, as the current findings are based on web-based assessments with a self-selected sample that may not generalize to clinical populations encountered in neuropsychological practice. The supplementary analyses reinforce these patterns, revealing that arithmetic reasoning achieved the highest correlations with Grand Index scores across all batteries containing this measure, with coefficients demonstrating remarkable consistency across diverse battery configurations. This consistency suggests that arithmetic reasoning tasks may tap into fundamental cognitive processes that are central to overall neuropsychological performance within web-based assessment contexts, though the mechanisms underlying this relationship and its generalizability to traditional assessment formats remain empirical questions requiring systematic investigation across diverse populations and assessment methodologies.

The systematic relationship between battery composition and factor structure complexity observed in this study provides preliminary evidence about conditions that may influence factor differentiation in web-based assessment contexts, though these findings must be interpreted within the constraints of the current methodological framework. The observation that seven batteries containing 5-9 subtests each produced unidimensional structures, while only the 11-subtest battery spanning four cognitive domains yielded multifactor solutions, suggests that cognitive breadth may influence factor emergence within this specific assessment format and population. However, these findings are based on existing Lumosity battery configurations rather than systematic manipulation of battery construction approaches, limiting conclusions about optimal assessment design principles or the relationship between theoretical and empirical approaches to battery development. The predominantly unidimensional structures across diverse battery configurations suggest that general cognitive ability may be robustly measured across different battery compositions in web-based contexts, though validation of measurement equivalence and clinical utility across different assessment approaches would be needed to establish evidence-based assessment practices. The correlation-based factor analysis approach employed in this study cannot establish causal relationships or definitively characterize cognitive architecture, as the empirically-derived factor structures may reflect methodological artifacts, measurement characteristics, or sample-specific patterns rather than representing the underlying cognitive organization. Furthermore, the theoretical implications of these findings require careful consideration, as the predominance of unidimensional structures may reflect specific characteristics of web-based assessment formats, self-selected participant samples, and existing battery construction methods rather than fundamental properties of cognitive architecture that would generalize across diverse assessment contexts and populations.

The predominantly unidimensional factor structures observed in this study must be interpreted within the context of several critical methodological considerations that may systematically influence the emergence of cognitive architectures and provide alternative explanations for the observed patterns. The web-based assessment format employed in this investigation differs fundamentally from traditional neuropsychological assessment contexts, potentially affecting both participant engagement and the manifestation of cognitive abilities through computerized interfaces rather than face-to-face administration. The self-selected sample of Lumosity users represents a specific population that may not generalize to clinical populations, educational settings, or diverse demographic groups typically encountered in neuropsychological practice, as these participants voluntarily engaged with cognitive training programs, potentially introducing selection biases that could influence both baseline cognitive abilities and response patterns to assessment tasks. The specific battery construction methods employed in the original Lumosity platform, while providing standardized configurations for empirical analysis, were not designed to systematically test theoretical models of cognitive architecture or optimize factor structure detection, potentially constraining the range of cognitive abilities sampled and the complexity of factor structures that could emerge. These methodological characteristics may systematically bias results toward simpler structures rather than reflecting genuine cognitive organization, as the assessment format may not adequately capture the full spectrum of cognitive abilities typically assessed in traditional neuropsychological batteries. The web-based format may also introduce measurement artifacts related to response timing, attention maintenance, and task engagement that could artificially inflate correlations among subtests, contributing to the emergence of unidimensional structures through methodological rather than theoretical mechanisms. Additionally, the specific cognitive tasks employed in the Lumosity platform may share common method variance or assessment characteristics that promote unidimensional factor emergence independent of underlying cognitive architecture, highlighting the need for systematic comparison across diverse assessment formats and populations before drawing conclusions about cognitive organization.

The practical implications of these findings require substantial qualification given the study's methodological constraints and the extensive validation needed before any applications to clinical practice can be considered. While the consistent correlations between arithmetic reasoning and composite scores across battery configurations suggest that reasoning measures may serve as robust indicators of general cognitive ability within web-based assessment formats, the clinical utility and efficiency of such measures compared to comprehensive assessment approaches require empirical validation in diverse populations and traditional assessment settings. The findings are based on a self-selected sample of cognitive training users completing web-based assessments, which may differ substantially from clinical populations encountered in neuropsychological practice, educational assessment, or occupational evaluation contexts in terms of motivation, cognitive abilities, demographic characteristics, and response patterns. The generalizability of these factor structures to established neuropsychological batteries, clinical populations, and traditional assessment procedures remains an empirical question that cannot be resolved through the current investigation alone, as the specific characteristics of web-based assessment formats may systematically influence factor structure emergence in ways that do not translate to clinical contexts. Future research must systematically examine whether the observed patterns replicate across diverse populations, assessment formats, and battery construction approaches before establishing evidence-based principles for cognitive assessment that could inform clinical practice. The transformation of neuropsychological assessment practices requires extensive empirical validation across multiple contexts, populations, and assessment methodologies rather than extrapolation from findings within a single assessment framework, particularly given the potential for web-based assessment characteristics to systematically influence the detection of cognitive structures in ways that may not reflect the complexity of cognitive organization typically observed in clinical populations.

This investigation provides empirical evidence for the predominance of unidimensional cognitive factor structures across diverse neuropsychological test battery configurations within web-based assessment contexts, contributing to ongoing debates about the organization of human cognitive abilities while highlighting critical methodological considerations that influence factor structure detection. The systematic analysis of over 5 million subtest scores from 757,427 participants revealed that seven of eight batteries yielded single-factor solutions, with only the most comprehensive battery demonstrating multifactor structure, suggesting that general cognitive ability may be robustly detected across different battery compositions within this specific assessment framework. The exceptional factor stability observed through split-sample validation procedures establishes the reliability of these structural findings within the current methodological context, while the consistent predictive strength of arithmetic reasoning across all battery configurations provides evidence for the stability of reasoning measures as indicators of general cognitive ability in web-based formats. However, the specific characteristics of web-based assessment formats, self-selected participant samples, and existing battery construction methods introduce substantial limitations that require careful consideration when interpreting these findings, as generalization to clinical populations and traditional neuropsychological assessment contexts remains an empirical question requiring systematic validation across diverse populations and assessment methodologies. The methodological approach demonstrated in this study establishes a framework for large-scale factor analytic investigations of cognitive architecture that extends beyond the specific tests examined, offering a foundation for future research that systematically manipulates battery composition across diverse populations and assessment contexts. The findings underscore the complex relationship between assessment methodology and the detection of cognitive structure, revealing how seemingly fundamental questions about cognitive organization may be significantly influenced by the specific characteristics of assessment formats, participant populations, and measurement approaches employed. By bridging computational cognitive science with traditional psychometric approaches, this work illuminates the need for systematic investigation of how methodological factors influence the emergence of cognitive architectures, setting the stage for future research that can distinguish between genuine cognitive organization and methodological artifacts in the pursuit of evidence-based advances in neuropsychological assessment.

## Acknowledgments

The authors thank the participants who contributed their data to this research through the Lumosity platform. We acknowledge Lumos Labs, Inc. for providing access to the NeuroCognitive Performance Test dataset that made this large-scale analysis possible.

## Funding

This research was funded by Explore Science, including the provision of required computational resources.

## References

Ackerman, P. & Rolfhus, E. (1999). The locus of adult intelligence: Knowledge, abilities, and nonability traits. *Psychology and Aging*, *14*, 314, doi:10.1037//0882-7974.14.2.314.

Adcock, C. J. & Webberley, M. (1971). Primary mental abilities. *The Journal of General Psychology*, *84*(2), 229–243, doi:10.1080/00221309.1971.9711309.

Anderson, C., Figa-Saldana, J., Wilson, J. J. W., & Ticconi, F. (2017). Validation and cross-validation methods for ASCAT. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(5), 2232–2239, doi:10.1109/JSTARS.2016.2639784.

Brandt, J. (1991). The hopkins verbal learning test: Development of a new memory test with six equivalent forms. *Clinical Neuropsychologist*, *5*(2), 125–142, doi:10.1080/13854049108403297.

Breit, M., Brunner, M., & Preckel, F. (2019). General intelligence and specific cognitive abilities in adolescence: Tests of age differentiation, ability differentiation, and their interaction in two large samples. *Developmental Psychology*, doi:10.1037/dev0000876.

Breit, M., Brunner, M., Preuß, J., Daseking, M., Pauls, F., Walter, F., & Preckel, F. (2025). The contribution of general intelligence to cognitive performance across the lifespan: A differentiation analysis of the wechsler tests. *Psychology and Aging*, doi:10.1037/pag0000875.

Broadbent, D. E. (1958). Perception and communication.

Burgoyne, A. P., Tsukahara, J. S., Draheim, C., & Engle, R. (2020). Differential and experimental approaches to studying intelligence in humans and non-human animals. *Learning and Motivation*, doi:10.31219/osf.io/ta468.

Carroll, J. (1993). The factor-analytic approach to intelligence is alive and well: A review of Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. *Canadian Journal of Experimental Psychology*, *47*, 763–766, doi:10.1037/H0084956.

Comrey, A. & Lee, H. B. (1973). *A first course in factor analysis 2nd ed.*

Crawford, A. V., Green, S. B., Levy, R., Lo, W.-J., Scott, L., Svetina, D., & Thompson, M. S. (2010). Evaluation of parallel analysis methods for determining the number of factors. *Educational and Psychological Measurement*, *70*(6), 885–901, doi:10.1177/0013164410379332.

Crosby, T. (1994). How to detect and handle outliers. *Technometrics*, *36*(3), 315–316, doi:10.1080/00401706.1994.10485810.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299, doi:10.1037/1082-989X.4.3.272.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: Factor structure recovery for dichotomous items. *Journal of Educational Measurement*, *43*(1), 39–52, doi:10.1111/J.1745-3984.2006.00003.X.

Hitch, G. J. (1984). Working memory. *Psychological Medicine*, *14*(2), 265–271, doi:10.1017/S0033291700003548.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185, doi:10.1007/BF02289447.

Horn, J. L. & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270, doi:10.1037/H0023816.

Humphreys, L. G. & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement*, *29*(3), 571–578, doi:10.1177/001316446902900303.

Immekus, J. & Cipresso, P. (2019). Editorial: Parsing psychology: Statistical and computational methods using physiological, behavioral, social, and cognitive data. *Frontiers in Psychology*, *10*, doi:10.3389/fpsyg.2019.02694.

Jacobucci, R. & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Psychological Review*, doi:10.1037/rev0000557.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151, doi:10.1177/001316446002000116.

Kyriazos, T. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *09*, 2207–2230, doi:10.4236/PSYCH.2018.98126.

Lawton, G. (1939). Review of the measurement of adult intelligence. *American Journal of Orthopsychiatry*, *9*(4), 808–810, doi:10.1111/J.1939-0025.1939.TB05653.X.

Lee, K. Y. & Sellbom, M. (2020). Further validation of the elemental psychopathy assessment – short form (EPA-SF) in a large university sample. *Journal of Personality Assessment*, *103*, 289–299, doi:10.1080/00223891.2020.1779734.

Leong, V., et al. (2022). A new remote guided method for supervised web-based cognitive testing to ensure high-quality data: Development and usability study. *Journal of Medical Internet Research*, doi:10.2196/28368.

Lezak, M., Howieson, D., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment, 5th ed.*

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99, doi:10.1037/1082-989X.4.1.84.

MacCann, C., Joseph, D. L., Newman, D. A., & Roberts, R. D. (2014). Emotional intelligence is a second-stratum factor of intelligence: Evidence from hierarchical and bifactor models. *Emotion*, *14*(2), 358–374, doi:10.1037/a0034755.

Marnoufi, K., Touri, B., Bergadi, M., & Ghazlane, I. (2023). WAIS–IV, WISC–V, WPPSI–IV subtests and their relationship with CHC theory. *International Journal of Innovation and Economic Development*, doi:10.18775/ijied.1849-7551-7020.2015.94.2002.

Marsman, M. & Rhemtulla, M. (2022). Guest editors' introduction to the special issue "network psychometrics in action": Methodological innovations inspired by empirical problems. *Psychometrika*, doi:10.1007/s11336-022-09861-x.

Mayes, S. D., Calhoun, L., & Crowell, E. W. (2001). Clinical validity and interpretation of the Gordon diagnostic system in ADHD assessments. *Child Neuropsychology*, *7*(1), 32–41, doi:10.1076/chin.7.1.32.3151.

McGill, R. J. (2023). Confirmatory factor analysis of the WJ IV cognitive: What does the standard battery measure at school age? *Journal of Psychoeducational Assessment*, *41*, 461–468, doi:10.1177/07342829231159440.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10, doi:10.1016/J.INTELL.2008.08.004.

Merrifield, P. R. & Cattell, R. B. (1975). Abilities: Their structure, growth, and action. *American Educational Research Journal*, *12*(4), 516, doi:10.2307/1162752.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, *27*(3), 272–277, doi:10.1093/OXFORDJOURNALS.BMB.A070866.

Newson, J., Bala, J., Giedd, J., Maxwell, B., & Thiagarajan, T. C. (2024). Leveraging big data for causal understanding in mental health: A research framework. *Frontiers in Psychiatry*, *15*, doi:10.3389/fpsyt.2024.1337740.

Posner, M. I. (2024). Orienting of attention and spatial cognition. *Cognitive Processing*, *25*(S1), 55–59, doi:10.1007/s10339-024-01216-x.

Posner, M. I. & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*(1), 25–42, doi:10.1146/ANNUREV.NE.13.030190.000325.

Pyryt, M. C. (1998). Human cognitive abilities: A survey of factor analytic studies. *Gifted and Talented International*, *13*(2), 97–98, doi:10.1080/15332276.1998.11672894.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137–150, doi:10.1111/J.2044-8341.1941.TB00316.X.

Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, *8*(3), 271–276, doi:10.2466/pms.1958.8.3.271.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*(3), 403–428, doi:10.1037/0033-295X.103.3.403.

Schmank, C. J., Goring, S. A., Kovács, K., & Conway, A. R. A. (2019). Psychometric network analysis of the Hungarian WAIS. *Journal of Intelligence*, *7*(3), 21, doi:10.3390/jintelligence7030021.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201, doi:10.2307/1412107.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, *38*(5), 406–427, doi:10.1037/h0069792.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136, doi:10.1016/0010-0285(80)90005-5.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies*. Technical report.

Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861, doi:10.1037/0033-295X.113.4.842.

VanTassel-Baska, J. (2001). Book review: The g-factor, the science of mental ability. *Gifted Child Quarterly*, *45*(2), 154–155, doi:10.1177/001698620104500209.

Wechsler, D. (1955). *Manual for the Wechsler adult intelligence scale*.

Wilson, E. B. (1928). The abilities of man, their nature and measurement. By C. Spearman. New York, The Macmillan Co., 1927. vi + 415 + xxxii pp. *Science*, *67*(1731), 244–248, doi:10.1126/SCIENCE.67.1731.244.

Yarkoni, T. & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122, doi:10.1177/1745691617693393.

# 5 Supplementary Material

The robustness of the identified cognitive factor structures was evaluated through comprehensive validation analyses that examined both the stability of factor solutions and the relationships between individual cognitive measures and overall performance. These supplementary analyses provide critical evidence for the reliability and interpretability of the predominantly unidimensional cognitive architecture identified across the neuropsychological test batteries.

Split-sample cross-validation procedures were implemented to assess the stability of factor structures across independent participant samples. Each battery's participants were randomly divided into equal halves, with separate exploratory factor analyses conducted on each subsample using identical procedures to the original analysis. This validation approach directly addresses concerns about sample-specific factor solutions by demonstrating that cognitive structures remain consistent across different participant groups within the same battery configuration.

The split-sample validation results demonstrate exceptional stability of cognitive factor structures across all eight batteries. Tucker's congruence coefficients between corresponding factors from independent sample halves consistently exceeded acceptable stability thresholds, with all coefficients falling between 0.970 and 1.000.

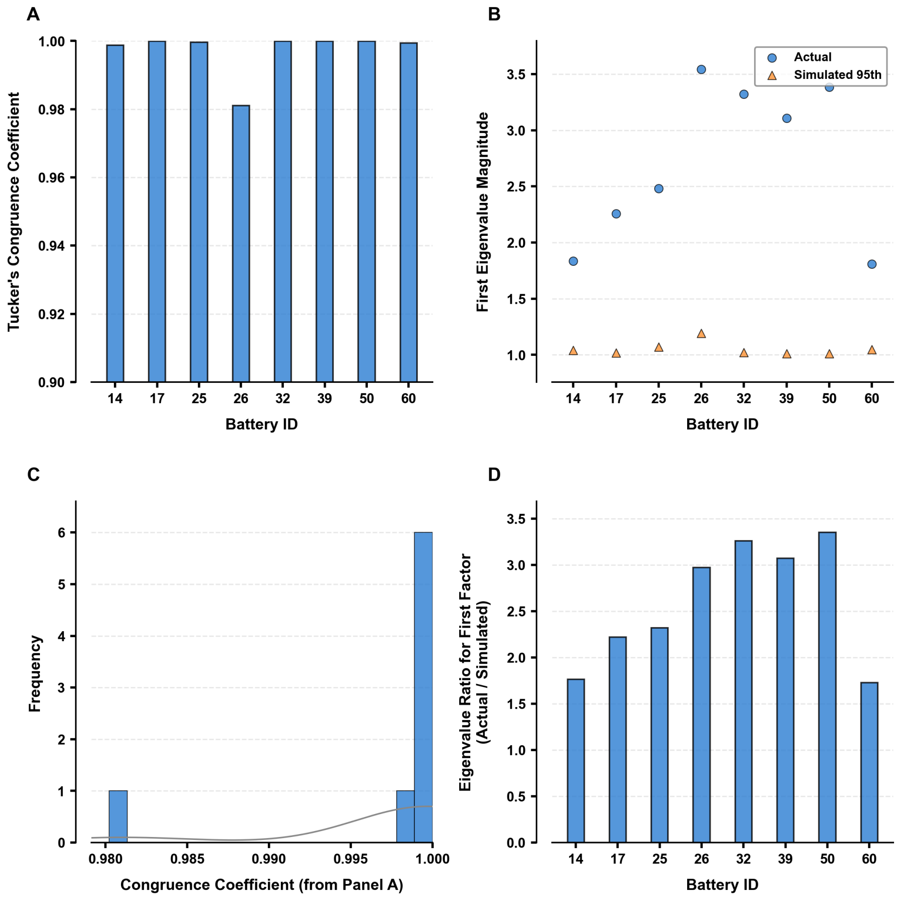


**Figure 4: Split-sample cross-validation confirms exceptional stability of cognitive factor structures across neuropsychological test batteries.** Tucker's congruence coefficients demonstrate remarkable factor stability, with all eight batteries achieving coefficients above 0.98, substantially exceeding the 0.85 acceptability threshold. This validation supports the robustness of single-factor solutions identified through exploratory factor analysis across diverse battery configurations. (A) Congruence coefficients from split-sample validation range from 0.980 to 1.000 across batteries (blue bars), with red dashed line marking 0.85 stability threshold. (B) Parallel analysis results comparing actual eigenvalues (blue circles) against simulated 95th percentile thresholds (orange triangles) for first factors, demonstrating clear factor retention validity. (C) Distribution of congruence coefficients shows clustering near perfect stability (mean = 0.996), with gray density curve overlaid on histogram. (D) Eigenvalue ratios quantify factor extraction strength relative to random chance, with all batteries showing ratios between 1.8-3.4. Split-sample validation employed random division of each battery sample into equal halves with independent factor analyses. Parallel analysis used 1000 simulated random datasets matching actual data dimensions. Tucker's congruence coefficients >0.95 indicate excellent factor stability. Sample sizes ranged from 752-31,868 participants per battery half (total $n$ = 651,104).

These coefficients substantially surpass the conventional 0.85 threshold for acceptable factor stability, with most batteries achieving coefficients above 0.95, indicating excellent replicability of the cognitive structures. The tight distribution of congruence values across batteries suggests that the identified factor solutions are highly robust and not dependent on specific participant characteristics or sampling variations.

The parallel analysis component of the validation further supports the single-factor interpretation across batteries. Actual eigenvalues for the first factor consistently exceeded the 95th percentile of simulated random data eigenvalues, while subsequent eigenvalues fell below this threshold, confirming that the unidimensional structures represent genuine cognitive variance rather than random associations. The eigenvalue ratios between actual and simulated values demonstrate the strength of factor extraction relative to chance expectations, with most batteries showing ratios between 1.8 and 3.4 for the primary factor.

Beyond structural validation, the relationship between individual cognitive measures and overall performance was examined through comprehensive correlation analyses. These analyses reveal systematic patterns in how specific cognitive abilities contribute to overall neuropsychological performance across different battery configurations. The correlation patterns provide insight into which cognitive domains serve as the strongest indicators of general cognitive ability within each assessment context.

**Table 1: Subtest-Grand Index correlations reveal arithmetic reasoning as the strongest predictor of overall cognitive performance across neuropsychological test batteries.** Pearson correlation coefficients ($r$) between individual subtest raw scores and Grand Index composite scores across eight NeuroCognitive Performance Test (NCPT) batteries ($n$ = 1,504 to 318,300 participants per battery). Correlations range from $-0.619$ to 0.770, with arithmetic reasoning (subtest 29) consistently demonstrating the highest correlations ($r$ = 0.659-0.770) across all seven batteries containing this measure. Values in brackets represent 95% confidence intervals (CI); strongest correlations within each battery are marked in the Notes column as “Strongest, Arithmetic”. Raw scores were inverted for go/no-go tasks (subtests 26, 32) and trail-making tasks (subtests 39, 40) so that higher scores consistently indicate better performance across all measures. Battery configurations varied from 5-11 subtests, with sample sizes ($n$) reflecting the breadth of cognitive domains assessed within each battery structure.

| **Battery/Subtest** | $r$ | **95% CI** | $n$ | **Notes** |
|---|---|---|---|---|
| **Battery 14** | | | | |
| Subtest 29 | **0.682** | [0.671-0.693] | 9462 | Strongest, Arithmetic |
| Subtest 30 | 0.671 | [0.660-0.682] | 9462 | |
| Subtest 28 | 0.573 | [0.559-0.586] | 9462 | |
| Subtest 27 | 0.541 | [0.526-0.555] | 9462 | |
| Subtest 26 | -0.526 | [-0.540-0.511] | 9462 | |
| **Battery 17** | | | | |
| Subtest 29 | **0.687** | [0.683-0.691] | 64735 | Strongest, Arithmetic |
| Subtest 30 | 0.641 | [0.637-0.646] | 64735 | |
| Subtest 28 | 0.628 | [0.623-0.633] | 64735 | |
| Subtest 31 | 0.565 | [0.559-0.570] | 64735 | |
| Subtest 27 | 0.562 | [0.556-0.567] | 64735 | |
| Subtest 32 | -0.566 | [-0.571-0.560] | 64735 | |
| **Battery 25** | | | | |
| Subtest 29 | **0.670** | [0.654-0.685] | 5044 | Strongest, Arithmetic |
| Subtest 30 | 0.637 | [0.620-0.653] | 5044 | |
| Subtest 28 | 0.626 | [0.609-0.643] | 5044 | |
| Subtest 33 | 0.618 | [0.601-0.635] | 5044 | |
| Subtest 31 | 0.559 | [0.540-0.578] | 5044 | |
| Subtest 27 | 0.517 | [0.497-0.537] | 5044 | |
| Subtest 32 | -0.498 | [-0.518-0.477] | 5044 | |
| **Battery 26** | | | | |
| Subtest 38 | **0.733** | [0.706-0.759] | 1182 | Strongest |

Continued on next page

**Table 1 continued from previous page**

| Battery/Subtest | $r$ | 95% CI | $n$ | Notes |
|---|---|---|---|---|
| Subtest 29 | 0.606 | [0.569-0.641] | 1182 | Arithmetic |
| Subtest 28 | 0.576 | [0.537-0.613] | 1182 | |
| Subtest 33 | 0.559 | [0.518-0.597] | 1182 | |
| Subtest 30 | 0.538 | [0.496-0.577] | 1182 | |
| Subtest 27 | 0.512 | [0.469-0.553] | 1182 | |
| Subtest 37 | 0.498 | [0.454-0.540] | 1182 | |
| Subtest 36 | 0.474 | [0.429-0.517] | 1182 | |
| Subtest 39 | -0.417 | [-0.463-0.368] | 1182 | |
| Subtest 32 | -0.522 | [-0.563-0.480] | 1182 | |
| Subtest 40 | -0.619 | [-0.653-0.582] | 1182 | |
| **Battery 32** | | | | |
| Subtest 38 | **0.770** | [0.767-0.773] | 87673 | Strongest |
| Subtest 29 | 0.659 | [0.656-0.663] | 87673 | Arithmetic |
| Subtest 28 | 0.621 | [0.617-0.625] | 87673 | |
| Subtest 33 | 0.601 | [0.596-0.605] | 87673 | |
| Subtest 30 | 0.599 | [0.595-0.603] | 87673 | |
| Subtest 27 | 0.516 | [0.511-0.521] | 87673 | |
| Subtest 31 | 0.494 | [0.489-0.499] | 87673 | |
| Subtest 39 | -0.500 | [-0.505-0.495] | 87673 | |
| Subtest 40 | -0.591 | [-0.595-0.587] | 87673 | |
| **Battery 39** | | | | |
| Subtest 38 | **0.764** | [0.763-0.766] | 206499 | Strongest |
| Subtest 29 | 0.675 | [0.673-0.678] | 206499 | Arithmetic |
| Subtest 28 | 0.622 | [0.620-0.625] | 206499 | |
| Subtest 33 | 0.610 | [0.607-0.613] | 206499 | |
| Subtest 30 | 0.605 | [0.603-0.608] | 206499 | |
| Subtest 31 | 0.501 | [0.498-0.505] | 206499 | |
| Subtest 39 | -0.499 | [-0.502-0.495] | 206499 | |
| Subtest 40 | -0.595 | [-0.598-0.592] | 206499 | |
| **Battery 50** | | | | |
| Subtest 45 | **0.770** | [0.769-0.772] | 268427 | Strongest |
| Subtest 43 | 0.715 | [0.713-0.716] | 268427 | |
| Subtest 44 | 0.700 | [0.698-0.702] | 268427 | |
| Subtest 29 | 0.673 | [0.671-0.675] | 268427 | Arithmetic |
| Subtest 30 | 0.618 | [0.615-0.620] | 268427 | |
| Subtest 31 | 0.499 | [0.496-0.501] | 268427 | |
| Subtest 39 | -0.504 | [-0.507-0.501] | 268427 | |
| Subtest 40 | -0.604 | [-0.607-0.602] | 268427 | |
| **Battery 60** | | | | |
| Subtest 51 | **0.708** | [0.697-0.719] | 8082 | Strongest |

Continued on next page

**Table 1 continued from previous page**

| **Battery/Subtest** | $r$ | **95% CI** | $n$ | **Notes** |
|---|---|---|---|---|
| Subtest 55 | 0.655 | [0.642-0.667] | 8082 | |
| Subtest 54 | 0.611 | [0.597-0.624] | 8082 | |
| Subtest 53 | 0.498 | [0.482-0.515] | 8082 | |
| Subtest 52 | 0.260 | [0.239-0.280] | 8082 | |
| **Summary Statistics** | | | | |
| **Battery** | **Mean** $r$ | **Range** | $n$ **Subtests** | **SD** |
| Battery 14 | 0.599 | 0.526-0.682 | 5 | 0.073 |
| Battery 17 | 0.608 | 0.562-0.687 | 6 | 0.052 |
| Battery 25 | 0.589 | 0.498-0.670 | 7 | 0.065 |
| Battery 26 | 0.550 | 0.417-0.733 | 11 | 0.084 |
| Battery 32 | 0.595 | 0.494-0.770 | 9 | 0.087 |
| Battery 39 | 0.609 | 0.499-0.764 | 8 | 0.087 |
| Battery 50 | 0.635 | 0.499-0.770 | 8 | 0.098 |
| Battery 60 | 0.546 | 0.260-0.708 | 5 | 0.178 |

Individual subtest correlations with Grand Index composite scores reveal arithmetic reasoning as the most consistent predictor of overall cognitive performance across neuropsychological test batteries. Arithmetic reasoning (subtest 29) achieved the highest correlations with Grand Index scores across all seven batteries containing this measure, with coefficients ranging from 0.659 to 0.770. This consistency across diverse battery configurations suggests that arithmetic reasoning tasks tap into fundamental cognitive processes that are central to overall neuropsychological performance.

The correlation patterns also reveal systematic differences in predictive strength across cognitive domains. Working memory tasks (subtests 28, 30, 33) consistently showed strong correlations with overall performance, typically falling within the 0.60-0.70 range. Processing speed measures demonstrated more variable relationships, with some batteries showing strong correlations while others exhibited weaker associations. Attention tasks showed the most variable pattern, with correlations ranging from moderate to strong depending on the specific battery configuration and the presence of other cognitive measures.

These correlation patterns align with the factor analysis results, supporting the interpretation that a general cognitive ability factor underlies performance across diverse neuropsychological tasks. The consistently strong correlations between arithmetic reasoning and overall performance suggest that this domain may serve as a particularly effective marker of general cognitive ability in neuropsychological assessment contexts. The variability in correlation strengths across other domains may reflect the extent to which different cognitive tasks require general versus specific cognitive abilities.

The supplementary analyses collectively demonstrate that the identified cognitive factor structures are both statistically robust and psychometrically meaningful. The exceptional stability evidenced by split-sample validation, combined with the systematic patterns observed in subtest-total correlations, provides strong support for the predominantly unidimensional cognitive architecture identified across the neuropsychological test batteries. These findings have important implications for understanding cognitive assessment and the interpretation of neuropsychological test performance.