Rigorous model comparison reveals linear speed-reasoning coupling

Explore Science
research@explorescience.ai

November 17, 2025

Abstract

Processing speed theories propose that cognitive operation rate constrains higher-order reasoning performance through linear coupling, while resource allocation theories predict nonlinear patterns where different cognitive systems become limiting at different performance levels. Despite extensive research linking processing speed and reasoning, the functional form of this relationship remains empirically underspecified, with important theoretical implications for understanding cognitive architecture. We tested competing linear and nonlinear models (exponential, piecewise linear, and smooth cubic spline) using 556,776 participants across three independent cognitive assessment batteries, applying cross-validation procedures and practical significance thresholds to adjudicate between theoretical accounts.Linear models were consistently selected across all twelve battery-outcome-predictor combinations, with nonlinear alternatives showing effectively zero improvement in predictive accuracy. Processing speed explained 7.4-16.0% of variance in reasoning performance, while working memory moderation effects contributed less than 0.08% of additional variance, indicating predominantly additive rather than interactive relationships between these cognitive constructs. However, sensitivity analyses revealed that these conclusions depended critically on Trail Making Test operationalization: log-transformed completion times yielded linear speed-reasoning relationships, whereas raw seconds produced nonlinear patterns with substantial model improvements. Cross-battery analyses further demonstrated that three of four speed-reasoning relationships exhibited battery-specific parameters, indicating limited generalizability across measurement contexts.These findings reveal that theoretical conclusions regarding processing speed's role in reasoning depend fundamentally on measurement operationalization choices rather than reflecting stable cognitive architecture properties. When Trail Making Test times are log-transformed, results support linear processing speed theory, but this support dissolves under alternative operationalizations that produce contradictory functional forms. Transformation dependency represents a critical boundary condition for interpreting speed-cognition relationships and challenges the generalizability of functional form conclusions across different assessment batteries and operationalization conventions.

Keywords: processing speed, fluid intelligence, working memory, cognitive aging, model comparison

1 Introduction

Processing speed - the rate at which elementary cognitive operations execute (Donders, 1969; Kail and Salthouse, 1994) - fundamentally constrains complex cognitive performance across the lifespan (Salthouse, 1996, 2000). Decades of research have established processing speed as a central mechanism underlying age-related cognitive decline (Cerella, 1985; Myerson et al., 1990; Salthouse, 1996; Wei et al., 2024), individual differences in fluid intelligence (Jensen, 1998; Mashburn et al., 2023), and clinical neuropsychological impairment (Wechsler, 1940; Ogden, 2005). Despite extensive documentation that processing speed predicts cognitive abilities, a foundational assumption has never been rigorously tested: the precise mathematical form of speed-cognition relationships. Whether these relationships follow linear or nonlinear functional forms determines theoretical interpretations - distinguishing uniform constraint models from multiple-bottleneck accounts - and affects clinical diagnostic accuracy through normative comparison frameworks (Strauss et al., 1991).

Competing theoretical frameworks make divergent predictions about speed-cognition functional forms. Classical processing speed theory posits that general cognitive slowing affects all mental operations proportionally (Cerella, 1985; Myerson et al., 1990; Salthouse, 1996), predicting monotonic linear relationships where cognitive performance scales uniformly with processing speed across the entire performance distribution. Despite early meta-analytic examination of linear versus nonlinear age effects (Verhaeghen and Salthouse, 1997), systematic empirical tests comparing functional specifications for speed-cognition relationships remain absent from the cognitive literature. Conversely, resource allocation models propose that multiple cognitive systems - including processing speed, working memory capacity (Baddeley and Hitch, 1974; Baddeley, 2012; Cowan, 2001; Conway et al., 2003), and executive attention (Kane and Engle, 2002) - become differentially limiting at different performance levels (Fry and Hale, 1996; Mashburn et al., 2023). These models predict nonlinear relationships featuring threshold effects where sufficiently fast processing enables qualitatively different cognitive strategies, or accelerating returns where speed benefits compound at higher performance levels. Kyllonen and Zu (2016) demonstrated that speed of solving simple items was uncorrelated with probability of solving complex items of the same type, suggesting these dimensions capture distinct cognitive properties. Virtually all published research employs linear regression models or structural equation models with linear path specifications (Wilms and Nielsen, 2015), though this default linearity assumption may reflect accumulated observation that linear specifications adequately capture relationships rather than blind assumption.

Working memory capacity represents a second fundamental constraint on cognitive performance (Salthouse and Babcock, 1991; Unsworth and Engle, 2007) that could create dual-constraint effects indicative of nonlinear resource allocation (Fry and Hale, 1996). The relationship between working memory capacity and fluid intelligence is well-established (Engle et al., 1999; Ackerman et al., 2005; Conway et al., 2003), though whether working memory benefits vary systematically across the processing speed distribution remains untested. Working memory moderation hypotheses further predict task-specific patterns reflecting differential executive control demands and relational complexity across reasoning tasks. Progressive Matrices tasks (Raven et al., 1998) impose high executive control demands through abstract pattern recognition, relational integration across multiple elements (Halford et al., 1998), and systematic hypothesis testing (Carpenter et al., 1990; Miyake et al., 2000; Diamond, 2013). Conversely, Grammatical Reasoning tasks (Baddeley, 1968)

involve lower executive demands through rule application and sequential statement verification, with greater involvement of crystallized knowledge systems (Cattell, 1963; Horn and Cattell, 1966). Differential patterns across tasks would validate resource allocation frameworks, whereas uniform linearity would support general processing speed theory.

To address these theoretical questions, we conducted a preregistered comparison of four functional forms - linear ordinary least squares regression, exponential models, piecewise linear models (Gordon et al., 1984), and smooth cubic splines (de Boor, 1978; Chiang, 2007) - for speed-reasoning relationships across 556,776 participants from three independent cognitive assessment batteries. Our methodological framework integrated cross-validation procedures (Stone, 1974; Efron and Tibshirani, 1994; Nordhausen, 2009; Yarkoni and Westfall, 2017) with enhanced model selection criteria (Akaike, 1974; Guthery et al., 2003) and Benjamini-Hochberg false discovery rate correction (Benjamini and Hochberg, 1995). Core analytical procedures testing functional form specifications and working memory moderation were preregistered before examining speed-reasoning relationships (Nosek et al., 2018), though sensitivity analyses examining alternative Trail Making Test transformations (Reitan, 1958; Tombaugh, 2004) and outlier inclusion effects were designated as exploratory robustness checks to assess the stability of primary findings. Sample sizes ranged from 86,358 to 263,853 participants per battery-outcome-predictor combination, exceeding the prespecified threshold by approximately 100-fold, providing sufficient statistical power (Muller, 1989; Button et al., 2013) to detect small nonlinear improvements.

2 Method

All analyses were preregistered (Wagenmakers et al., 2011; Nosek et al., 2018) prior to examining speed-cognition relationships. This model comparison study examined functional forms across 556,776 participants from three test batteries using stratified cross-validation (Stone, 1974; Kohavi, 1995; Arlot and Celisse, 2010) and comprehensive sensitivity analyses. The de-identified dataset is publicly available.

2.1 Dataset and Sample Characteristics

We analyzed data from the NeuroCognitive Performance Test (NCPT), a web-based cognitive assessment battery administered between January 2013 and June 2020. The NCPT is a self-administered assessment accessed through web browsers on personal computers, comprising multiple brief subtests based on established neuropsychological paradigms. We selected three test batteries (Battery 32, 39, and 50) for analysis based on their large sample sizes and overlapping subtest composition, which enabled cross-battery generalizability testing. The sample included participants from four English-speaking countries: the United States, Canada, Australia, and New Zealand.

The initial dataset contained 641,216 unique participants aged 18-99 years who met dataset-level inclusion criteria: English as preferred language, reported demographic information (age, country, language), first NCPT assessment only, no pauses exceeding 24 hours between subtests, and Trail Making Test completion under 15 minutes. Following preliminary exclusion of 67,607 participants missing both reasoning outcome measures, the initial analytical sample comprised 573,609 participants (89.5% retention). To comply with Safe Harbor de-identification standards (Department of Health and Human Services, 2002), participants over age 89 were coded as 90

years.

The three batteries contributed differentially to the sample: Battery 32 included 90,789 participants, Battery 39 included 209,573 participants, and Battery 50 included 273,247 participants. All analytical combinations substantially exceeded the preregistered minimum sample size requirement of $N \geq 500$ participants per battery-outcome-predictor combination, with final sample sizes ranging from 86,358 to 263,846 after applying analysis-specific exclusion criteria. This threshold was established to ensure stable cross-validation (Varma and Simon, 2006) with nested hyperparameter tuning, reliable nonlinear model fitting, and adequate statistical power (Muller, 1989) for detecting meaningful RMSE improvements of 2% or greater.

2.2 Cognitive Measures and Operationalization

2.2.1 Reasoning Outcomes

The NCPT battery included two reasoning measures assessing distinct cognitive domains. Progressive Matrices (subtest 31) assessed abstract reasoning and fluid intelligence (Cattell, 1963; Horn and Cattell, 1966) through a computerized adaptation of Raven's Progressive Matrices (Raven et al., 1998), a well-validated measure of non-verbal reasoning that does not rely on crystallized knowledge. The subtest presented participants with a 3×3 grid containing abstract patterns in eight squares, with the lower-right square empty. Participants selected which of six possible choices best completed the pattern. The assessment comprised 17 trials with progressively increasing difficulty, automatically terminating after three consecutive incorrect responses. The raw score reflected the total number of correct trials. This measure has been extensively validated as an assessment of fluid intelligence (Carpenter et al., 1990), showing expected age-related decline patterns (Lindenberger and Baltes, 1994; Salthouse, 2009) and strong psychometric properties.

Grammatical Reasoning (subtest 30) evaluated logical reasoning under time pressure through a task inspired by Baddeley's grammatical reasoning paradigm (Baddeley, 1968). On each trial, participants viewed a blue square and blue triangle displayed side-by-side, with a logical statement presented below (e.g., "The triangle is to the left of the square"). Participants judged whether each statement was true or false, with negations included in 50% of trials to increase task difficulty. The 45-second time limit emphasized rapid logical inference. The raw score was calculated as the number of correct responses minus incorrect responses, with a floor of zero to prevent negative scores. This scoring approach penalized impulsive responding while rewarding both speed and accuracy.

2.2.2 Processing Speed Predictors

Processing speed was assessed through two complementary measures (Donders, 1969; Sternberg, 1966; Kail and Salthouse, 1994; Faust et al., 1999; Jensen, 2006) targeting different aspects of cognitive tempo. Digit Symbol Coding (subtests 38 and 45) employed a symbol-digit substitution paradigm derived from the Digit Symbol Substitution Test (Wechsler, 1955), a standard component of Wechsler intelligence scales. Participants viewed a legend mapping numbers 1-9 to abstract symbols, displayed continuously throughout the task. On each trial, participants saw one symbol and entered the corresponding number as quickly as possible. The 90-second time limit yielded a raw score reflecting the number of correct completions. Two versions of this subtest (38 and 45) were administered across different batteries but had no substantive differences in task structure or

difficulty. This measure assesses perceptual-motor speed, visual search efficiency, and working memory access for symbol-number associations (Joy, 2004; Hoyer et al., 2004; Jung, 2005).

Trail Making Test Part A (subtest 39) provided a computerized adaptation of the Trail Making Test (Reitan, 1958; Tombaugh, 2004; Bowie and Harvey, 2006), a widely used neuropsychological measure of visual scanning speed and psychomotor processing (S'anchez-Cubillo et al., 2009). Participants clicked sequentially on blue circles numbered 1-24 arranged in one of six possible random layouts. The task provided immediate error feedback through an 'X' mark when participants clicked an incorrect number, requiring them to return to the previous circle before continuing. The raw score represented completion time in seconds, with longer times indicating slower processing speed.

A critical methodological decision concerned the transformation of Trail Making A completion times. Raw completion times exhibit substantial positive skew characteristic of reaction time distributions (Ratcliff, 1993), potentially violating parametric statistical assumptions and introducing heteroscedasticity. Following established practices in response time modeling (Luce, 1986; Ratcliff, 1993; Whelan, 2008), we applied a logarithmic transformation to normalize the distribution. Specifically, raw completion times were log-transformed, then negated to preserve intuitive directionality (higher scores indicating faster performance), and finally standardized within each battery. This transformation sequence served multiple purposes: the logarithmic function normalized right-skewed distributions while potentially reflecting Weber-Fechner principles (Fechner, 1889) in temporal perception, negation maintained interpretability, and within-battery standardization accounted for potential battery-specific administration contexts or population differences. This transformation choice was consequential, as sensitivity analyses revealed that alternative operationalizations (raw seconds versus log-transformed) produced fundamentally different functional form selections (see Sensitivity Analyses section). The preference for log-transformation over alternatives (inverse transformation, Box-Cox; Box and Cox 1964; Osborne 2010) was based on its theoretical grounding in psychophysical principles of temporal perception and its widespread use in mental chronometry research.

2.2.3 Working Memory Predictors

Working memory capacity was assessed through forward and reverse spatial span tasks (Baddeley and Hitch, 1974; Cowan, 2001; Conway et al., 2005) based on the Corsi block-tapping test (Corsi, 1972; Milner, 1971; Vandierendonck et al., 2004), a gold-standard measure of visuospatial working memory originally developed by Corsi in 1972. Forward Memory Span (subtests 28 and 43) presented participants with 10 blue circles positioned at random non-overlapping spatial locations. Following a brief delay, a subset of circles was sequentially highlighted in orange. Participants reproduced the sequence by clicking circles in the same order presented. Two versions existed across batteries: subtest 28 highlighted each circle for 500 ms with 500 ms inter-stimulus intervals, while subtest 43 used 333 ms presentation and delay durations. In both versions, sequence length (span) increased systematically - subtest 28 increased span by one after every two trials starting from span 3, while subtest 43 increased after every three trials. Assessment terminated when participants responded incorrectly on two trials at the same span level. Scoring differed between versions: subtest 28 recorded the maximum span level for which at least one trial was answered correctly, whereas subtest 43 tallied the total number of correct trials across all span levels.

Reverse Memory Span (subtests 33 and 44) employed identical stimuli and spatial layouts but required participants to reproduce sequences in reverse order - clicking circles from last-to-first rather than first-to-last. This reverse-order requirement substantially increased executive demands (Miyake et al., 2000; Diamond, 2013), as participants must maintain spatial information while simultaneously manipulating its sequential organization. Version parameters matched their forward span counterparts (subtest 33 paired with 28; subtest 44 paired with 43), with scoring calculated identically: maximum span (subtest 33) or total correct trials (subtest 44). The distinction between forward and reverse span enabled examination of differential working memory components, with forward span primarily assessing storage capacity and reverse span tapping both storage and executive control processes.

2.2.4 Within-Battery Standardization Approach

All cognitive measures were z-standardized within each battery to account for potential version differences, administration contexts, and population variations across the three test batteries. For each measure, we calculated z-scores using the formula $z = (X - \mu_{\text{battery}})/\sigma_{\text{battery}}$, where X represents the individual raw score, μ_{battery} is the battery-specific mean, and σ_{battery} is the battery-specific standard deviation. This approach yielded standardized variables with mean $= 0$ and standard deviation $= 1$ within each battery, enabling meaningful cross-battery comparisons while preserving individual differences in performance.

The standardized variables created were: reasoning_matrices_z and reasoning_grammar_z for reasoning outcomes; ds_speed_z (harmonizing subtests 38 and 45) for Digit Symbol processing speed; trailsA_speed_z (the primary log-transformed, negated, standardized operationalization) for Trail Making A processing speed; wm_fwd_z (harmonizing subtests 28 and 43) and wm_rev_z (harmonizing subtests 33 and 44) for working memory measures. For measures with multiple subtest versions (Digit Symbol, Forward Span, Reverse Span), we first computed z-scores separately for each subtest version within each battery, then averaged the standardized scores for participants who completed both versions, or used the single available standardized score for participants who completed only one version. Missing values were preserved as NaN throughout all transformations to maintain transparency (Schafer and Graham, 2002) in subsequent analyses and enable analysis-specific sample size reporting.

2.2.5 Alternative Processing Speed Operationalizations

To assess the robustness of findings to Trail Making Test transformation choices, we created two alternative operationalizations for sensitivity analyses. The first alternative (trailsA_speed_neg) applied negative raw seconds: raw completion times were multiplied by -1, then z-standardized within each battery. This transformation preserved the original scale properties while reversing direction for interpretability. The second alternative (trailsA_speed_inv) used negative inverse time: raw completion times were transformed as $-1/\text{seconds}$, then z-standardized within each battery. This reciprocal transformation is common in reaction time research and emphasizes differences at faster speeds. These alternative operationalizations enabled systematic examination of whether log-transformation critically influenced functional form conclusions.

2.3 Outlier Detection and Data Quality Control

We implemented an outlier removal approach (David and Tukey, 1977; Mosteller and Tukey, 1977) without winsorization, motivated by the theoretical concern that extreme values could drive apparent nonlinearities through leverage effects (Dixon, 1953) rather than reflecting genuine population-level functional relationships. This strategy prioritized examination of typical speed-cognition relationships within the normal performance range. Any participant flagged on any exclusion criterion was removed entirely from the analytical sample.

2.3.1 Cognitive Performance Outliers

Cognitive performance outliers were identified using an absolute z-score threshold of 3.0 standard deviations on six standardized variables: ds_speed_z, trailsA_speed_z, wm_fwd_z, wm_rev_z, reasoning_grammar_z, and reasoning_matrices_z. This threshold corresponds to approximately the 99.7th percentile of a normal distribution, identifying only the most extreme performers while retaining participants with naturally high or low but plausible performance levels. The criterion was applied within each battery to respect battery-specific distributions. This procedure flagged 14,178 participants across all criteria and batteries. Notably, zero participants were flagged on reasoning_matrices_z, indicating no extreme values beyond ± 3 standard deviations on this measure across any battery.

2.3.2 Trail Making A Robust Outlier Detection

We supplemented standard z-score outlier detection with a robust method specifically for Trail Making A raw completion times, given this measure's particular susceptibility to extreme values. The procedure employed median absolute deviation (MAD), a robust scale estimator resistant to outlier influence (Huber, 1981; Hoaglin and Iglewicz, 1987; Leys et al., 2013; Wilcox, 2012). For each battery, we calculated the median and MAD of log-transformed Trail Making A times. The robust z-score was computed as $z_{\text{robust}} = (\log(s39) - \text{median}_{\log \text{time}})/(1.4826 \times \text{MAD}_{\log \text{time}})$, where the constant 1.4826 represents the scale factor making MAD a consistent estimator of standard deviation for normally distributed data (Rousseeuw and Croux, 1993). Participants with $|z_{\text{robust}}| > 3.0$ were flagged for exclusion. This robust approach identified 10,132 participants (Battery 32: 1,403; Battery 39: 3,561; Battery 50: 5,168), capturing extreme completion times that might reflect recording errors, task misunderstanding, or non-compliance rather than genuine individual differences in processing speed.

2.3.3 Feasibility Thresholds

Two additional criteria targeted implausible Trail Making A completion times. A minimum threshold flagged times under 10 seconds as likely recording errors, given that sequentially clicking 24 circles with visual search and motor execution demands requires substantial time even under optimal conditions. This criterion identified 24 participants. A maximum threshold flagged times exceeding 300 seconds (5 minutes) as indicating non-compliance or task abandonment, given that typical completion times fall well below this duration. This criterion identified 1 participant. These thresholds balanced inclusivity with data quality, removing only clearly implausible performances.

2.3.4 Additional Quality Control

Three additional procedures ensured data integrity. First, duplicate removal identified any identical user_id, test_run_id, and battery_id combinations, detecting zero duplicates. Second, invariant responding detection flagged participants who obtained identical raw scores across all completed subtests within a battery, a pattern suggesting invalid responding; zero cases were detected. Third, participants missing both reasoning outcome measures (reasoning_grammar_z and reasoning_matrices_z) were excluded to ensure meaningful model fitting, though this criterion removed no additional participants beyond the preliminary exclusion of 67,607 participants.

2.3.5 Final Exclusion Summary

The comprehensive outlier detection pipeline flagged 16,833 participants (2.9% of the initial 573,609) for exclusion, yielding a final analytical sample of 556,776 participants with 97.1% overall retention. An exclusion log documented the specific number of participants removed for each criterion within each battery, providing complete transparency and traceability. All outlier flags were preserved in the dataset through an outlier_any boolean variable, enabling future researchers to examine sensitivity to exclusion decisions.

2.4 Demographic Covariate Selection via Lasso Regularization

To prevent arbitrary demographic adjustment decisions while controlling for potential confounds (Schmidt and Hunter, 1998), we implemented data-driven covariate selection using Lasso (Least Absolute Shrinkage and Selection Operator) regularization (Tibshirani, 1996) with cross-validation. This approach systematically identifies relevant demographic predictors while promoting parsimony through L1 penalization (Tibshirani, 1996; Zou and Hastie, 2005), which shrinks irrelevant coefficients to exactly zero. Separate covariate selection was performed for each reasoning outcome to accommodate outcome-specific demographic relationships, and this process occurred before any hypothesis testing to maintain preregistration integrity.

2.4.1 Candidate Covariate Set and Encoding

The candidate covariate set included five demographic variables known to influence cognitive performance (Neisser et al., 1996): age (continuous), time_of_day (continuous hours when test was initiated), gender (categorical), country (categorical), and education_level (categorical). Continuous covariates were standardized within battery_id groups to z-scores, accounting for potential battery-specific demographic distributions. Categorical variables were one-hot encoded using the most frequent category across the entire dataset as the reference level: female gender (contrasted with male), United States (contrasted with Australia, Canada, New Zealand), and college degree education level 4 (contrasted with levels 1=some high school, 2=high school diploma/GED, 3=some college, 5=professional degree, 6=master's degree, 7=Ph.D., 8=associate's degree, and 99=other). This encoding scheme produced 14 total candidate predictors: 2 continuous standardized variables and 12 dummy-coded categorical variables.

2.4.2 LassoCV Implementation and Model Selection

Covariate selection employed scikit-learn's (Pedregosa et al., 2022) LassoCV implementation (Friedman et al., 2010) with stratified 10-fold cross-validation. Stratification was based on battery_id to ensure balanced representation of all three batteries within each fold, maintaining the integrity of battery-specific distributions during validation. The regularization parameter α was selected using the one-standard-error rule (Gordon et al., 1984; Nordhausen, 2009), a principled approach that selects the most parsimonious model (largest α) whose cross-validated performance falls within one standard error of the best-performing model's performance. This rule implements a bias-variance tradeoff favoring simpler models unless complexity demonstrably improves prediction, reducing overfitting risk. Cross-validation performance was assessed using negative mean squared error as the loss metric. The algorithm employed parallel processing with 2 workers to enhance computational efficiency.

2.4.3 Progressive Matrices Covariate Selection

For Progressive Matrices (y_mat_z), listwise deletion (Lenth, 2004; Enders, 2010) removed 5,173 participants (0.93%) with missing outcome or demographic data, yielding a complete-case sample of 551,603 participants. The LassoCV procedure selected $\alpha = 0.002247$ under the one-standard-error rule, achieving cross-validated performance of -0.934117 ± 0.001560. Eleven covariates received non-zero coefficients: age_std ($\beta = -0.207$, indicating age-related decline; Schaie 1994; Salthouse 1996), gender_m ($\beta = 0.010$, small male advantage; Brannon 2012; Hyde 2005), country_AU ($\beta = 0.019$, slight Australian advantage relative to US), and eight education level indicators (β ranging from -0.310 for high school diploma to $+0.123$ for Ph.D.; Ceci 1991). Notably, time_of_day_std and country indicators for Canada and New Zealand were not retained (coefficients shrunk to zero), suggesting minimal systematic relationships with matrices performance after accounting for other demographics.

2.4.4 Grammatical Reasoning Covariate Selection

For Grammatical Reasoning (y_gram_z), only 6 participants (0.001%) were excluded for missing data, producing a complete-case sample of 556,770 participants. The selected regularization parameter was $\alpha = 0.002032$, achieving cross-validated performance of -0.884783 ± 0.001101. Thirteen covariates received non-zero coefficients, representing a slightly more complex demographic structure than matrices. Continuous predictors included age_std ($\beta = -0.243$, stronger age effect than matrices; Salthouse 1996) and time_of_day_std ($\beta = -0.017$, small decrements for later testing), suggesting vulnerability to fatigue or circadian factors (May et al., 1993; Schmidt-Reinwald et al., 1999). Gender showed a larger effect (gender_m $\beta = 0.020$; Brannon 2012; Hyde 2005) than for matrices. Country effects differed notably from matrices: country_AU showed negative association ($\beta = -0.045$), possibly reflecting population-level differences (Flynn, 1987; Lynn and Vanhanen, 2002), and country_CA emerged ($\beta = -0.108$), both indicating poorer performance relative to US participants. Education effects (β ranging from -0.285 for high school diploma to $+0.079$ for Ph.D.; Ceci 1991) paralleled matrices patterns but with different magnitudes.

2.4.5 Cross-Task Comparison and Implementation

The grammatical reasoning model retained more covariates (13 versus 11) and showed better cross-validated performance, suggesting more systematic demographic relationships. Key differences included stronger age effects for grammatical reasoning, emergence of time-of-day effects only for grammatical reasoning, and divergent country effects (positive for Australian participants on matrices, negative for Australian and Canadian participants on grammatical reasoning). These outcome-specific covariate sets were applied consistently in all subsequent regression models for their respective outcomes, ensuring proper demographic adjustment without arbitrary variable selection or overfitting to sample-specific patterns. The Lasso regularization framework provided an empirically grounded, replicable approach to covariate selection that balanced predictive accuracy with model parsimony.

2.5 Functional Form Model Comparison and Cross-Validation

For each combination of battery (32, 39, 50), outcome (Progressive Matrices, Grammatical Reasoning), and speed predictor (Digit Symbol, Trail Making A), we compared four functional forms: linear ordinary least squares regression, exponential models using the form $y = \beta_0 + \beta_1 \times \exp(\beta_2 \times \text{speed}) + \Sigma\beta_j \times \text{covariate}_j + \varepsilon$ fitted via nonlinear least squares, piecewise linear models with single breakpoint optimization, and smooth cubic splines with tuned degrees of freedom. Each analytical combination employed stratified 10-fold cross-validation (Efron and Tibshirani, 1994; Stone, 1974) with stratification based on outcome tertiles computed within each specific battery-outcome-predictor analysis subset to ensure balanced representation of performance levels within each fold. The random seed was set to 42 for complete reproducibility across all analyses.

2.5.1 Model Fitting and Hyperparameter Tuning

Within each of the 10 outer training folds, four models were fitted and tuned. Linear models included the speed predictor and previously selected covariates. Exponential models employed robust starting values to ensure convergence. Piecewise linear models optimized breakpoint location via inner 5-fold cross-validation, testing breakpoints from the 10th to 90th percentiles of the speed distribution in 10-percentile increments. Smooth cubic spline models tuned degrees of freedom from the set $\{3, 4, 5, 6\}$ via inner 5-fold cross-validation to prevent overfitting. For each tuned model, performance was evaluated on the corresponding outer test fold, calculating root mean squared error (RMSE), mean absolute error (MAE), and R-squared. Performance metrics were aggregated across the 10 outer folds by computing means and standard deviations.

2.5.2 Model Selection Criteria

Model selection employed the one-standard-error rule (Gordon et al., 1984; Nordhausen, 2009) combined with a practical significance threshold. The one-standard-error rule was implemented by calculating the threshold as $\text{CV-RMSE}_{\text{linear}} + \text{SE}_{\text{linear}}$, where $\text{SE}_{\text{linear}}$ is the standard deviation of cross-validated RMSE across the 10 folds for the linear model. A nonlinear model was selected only if its mean CV-RMSE fell below this threshold. Additionally, a practical significance criterion required nonlinear models to demonstrate RMSE improvements $\geq 1\%$ relative to the linear model, calculated as $100 \times (\text{RMSE}_{\text{linear}} - \text{RMSE}_{\text{selected}})/\text{RMSE}_{\text{linear}}$. If multiple nonlinear models satisfied both criteria, the model with the lowest mean CV-RMSE was selected. This dual-threshold approach

balanced statistical significance with practical importance, favoring parsimonious linear models unless nonlinear alternatives demonstrated both statistical superiority and meaningful predictive improvements.

2.6 Working Memory Moderation Analysis

For each battery-outcome-speed predictor combination meeting the sample size threshold, we tested working memory moderation within the best-fitting functional forms identified through cross-validation. The analysis examined whether Forward Span (wm_fwd_z) and Reverse Span (wm_rev_z) exhibited differential predictive relationships with reasoning performance across the processing speed distribution. Complete-case analysis required non-missing values for the outcome, speed predictor, both working memory measures, and previously selected covariates for that outcome.

2.6.1 Significance Testing and Effect Estimation

For significance testing, we fitted models using the single best-fitting functional form (the selected model from cross-validation results) including main effects for the speed predictor (using the selected functional form), wm_fwd_z, wm_rev_z, selected covariates, and two-way interactions between speed terms and each working memory variable. Statistical significance was assessed via partial F-tests comparing models with and without interaction terms, calculating partial eta-squared (η_p^2) effect sizes as $\eta_p^2 = \text{SS}_{\text{interaction}}/(\text{SS}_{\text{interaction}} + \text{SS}_{\text{error}})$. For effect estimation in cases where multiple functional forms were plausible (CV-RMSE within 1 standard error of optimal), marginal effects were computed as AIC-weighted averages across all plausible models, where AIC weights were calculated as $w_i = \exp(-0.5 \times \Delta\text{AIC}_i)/\Sigma \exp(-0.5 \times \Delta\text{AIC}_j)$, with ΔAIC representing the difference in AIC from the best model. This model averaging approach properly accounted for model selection uncertainty when interpreting moderation effects.

2.6.2 Marginal Effects Calculation

To interpret significant moderation effects, we calculated marginal effects of a $+1$ standard deviation change in working memory on reasoning outcomes at four specific levels of the speed predictor distribution: 25th percentile, 50th percentile (median), 75th percentile, and 95th percentile. Marginal effects were computed using the delta method to obtain standard errors for each estimate. We implemented a 2000-iteration bootstrap procedure by resampling participants with replacement to generate bias-corrected 95% confidence intervals for the marginal effect estimates at each speed percentile. For model-averaged cases, marginal effects and confidence intervals incorporated both parameter uncertainty within models and model selection uncertainty across plausible functional forms.

3 Results

3.1 Linear Speed-Cognition Relationships Across All Test Batteries

Across 12 battery-outcome-predictor combinations spanning 86,358 to 263,846 participants per analysis, rigorous cross-validation procedures (Stone, 1974; Arlot and Celisse, 2010; Geisser, 1975) with enhanced model selection criteria (Akaike, 1974; Stone, 1977) consistently favored

linear functional forms over nonlinear alternatives. Model selection evaluated both cross-validated prediction error and information criteria, employing the one-standard-error rule (Efron, 1983; Gordon et al., 1984) combined with a practical significance threshold requiring root mean squared error improvements $\geq 1\%$. Despite systematic evaluation of exponential models, piecewise linear models with optimized breakpoints (Muggeo, 2003), and smooth cubic spline models (Brown et al., 1991; Eilers and Marx, 1996; Wood, 2011), no combination satisfied the preregistered criteria for selecting nonlinear models. Nonlinear models achieved only marginal improvements when observed, with root mean squared error reductions of 0.03% to 0.13% - values 15 to 67 times smaller than the 2% practical significance threshold and failing to exceed linear model performance by one standard error. Table 1 presents comprehensive cross-validation results for all functional forms across battery-outcome-predictor combinations.

Table 1: Linear models universally outperform nonlinear alternatives across all speed-reasoning relationships in three cognitive test batteries. Cross-validated model comparison (N=556,776 unique participants after outlier exclusion using $|z| > 3.0$ threshold) testing four functional forms: linear ordinary least squares (OLS), exponential nonlinear least squares, piecewise linear with optimized breakpoint, and smooth cubic splines with tuned degrees of freedom. Section 1 presents linear model performance; Section 2 presents best nonlinear model performance. Rows show 12 battery-outcome-predictor combinations from NeuroCognitive Performance Test batteries 32, 39, and 50, grouped by battery and ordered by outcome (Progressive Matrices, then Grammatical Reasoning) and speed predictor (Digit Symbol, then Trail Making A). N indicates complete cases after listwise deletion for each specific analysis. Linear RMSE (root mean squared error; Mean ± SD format) shows 10-fold stratified cross-validation results with reasoning outcome tertile stratification (random seed=42). Linear R^2 shows coefficient of determination from cross-validation. Best Nonlinear RMSE and Best Nonlinear R^2 represent optimal performance (minimum RMSE, maximum R^2) among exponential, piecewise, and spline alternatives. Effective DF (degrees of freedom) differs by outcome due to LassoCV covariate selection: 13 parameters for Progressive Matrices (11 covariates + speed + intercept), 15 for Grammatical Reasoning (13 covariates including time_of_day and country_CA + speed + intercept). All linear models selected via enhanced 1-standard-error rule requiring nonlinear models achieve CV-RMSE improvement ≥1 SE or ≥1% practical threshold - neither criterion met in any comparison. Total observations across all 12 analyses: 2,216,740.

Section 1: Linear Model Performance

Battery	Outcome	Speed Predictor	N	Linear RMSE	Linear R^2
32	Progressive Matrices	Digit Symbol	86,358	0.9490 ± 0.0057	0.094
		Trail Making A	86,358	0.9557 ± 0.0061	0.081
	Grammatical Reasoning	Digit Symbol	88,606	0.9004 ± 0.0087	0.158
		Trail Making A	88,606	0.9189 ± 0.0085	0.123
39	Progressive Matrices	Digit Symbol	203,156	0.9526 ± 0.0039	0.087
		Trail Making A	203,156	0.9597 ± 0.0033	0.074
	Grammatical Reasoning	Digit Symbol	204,317	0.9051 ± 0.0039	0.154
		Trail Making A	204,317	0.9234 ± 0.0043	0.119
50	Progressive Matrices	Digit Symbol	262,088	0.9523 ± 0.0031	0.087
		Trail Making A	262,087	0.9591 ± 0.0031	0.074
	Grammatical Reasoning	Digit Symbol	263,846	0.8927 ± 0.0038	0.160
		Trail Making A	263,845	0.9118 ± 0.0030	0.124

Section 2: Nonlinear Model Performance

Battery	Outcome	Speed Predictor	Best Nonlinear RMSE	Best Nonlinear R^2	Effective DF
32	Progressive Matrices	Digit Symbol	0.9486	0.095	13
		Trail Making A	0.9554	0.082	13
	Grammatical Reasoning	Digit Symbol	0.9004	0.158	15
		Trail Making A	0.9179	0.125	15
39	Progressive Matrices	Digit Symbol	0.9519	0.089	13
		Trail Making A	0.9592	0.075	13
	Grammatical Reasoning	Digit Symbol	0.9048	0.154	15
		Trail Making A	0.9225	0.121	15
50	Progressive Matrices	Digit Symbol	0.9511	0.089	13
		Trail Making A	0.9587	0.074	13
	Grammatical Reasoning	Digit Symbol	0.8921	0.161	15
		Trail Making A	0.9111	0.125	15

Note: Total N = 2,216,740 across all 12 analyses. All selected models were linear.
1-SE rule applied with ≥1% improvement threshold. Spline models converged to linear performance.

The linear models demonstrated consistent predictive accuracy across all combinations, with cross-validated root mean squared error ranging from 0.8927 to 0.9597 and explained variance (R^2) ranging from 0.074 to 0.160 (Figure 1A,B). Cross-validation standard deviations were uniformly

small (0.0030 to 0.0087), indicating stable performance estimates with minimal fold-to-fold variability. Smooth cubic spline models systematically failed to converge across all 12 combinations and all 10 cross-validation folds despite comprehensive hyperparameter tuning (degrees of freedom tested: 3, 4, 5, 6) and robust starting values (Wood, 2011). When optimization algorithms did reach convergence criteria, the fitted splines reduced to approximately linear functions, producing root mean squared error values nearly identical to linear models (differences <0.0001), providing strong evidence that the underlying relationships lack the curvature that spline basis functions could capture. Exponential and piecewise linear models converged successfully but provided no meaningful improvement over linear specifications. The absence of cases requiring permutation testing (Winkler et al., 2014) for nonlinear model significance represents itself a substantive null finding: speed-reasoning relationships are adequately characterized by linear functions across diverse cognitive assessment contexts and large samples.

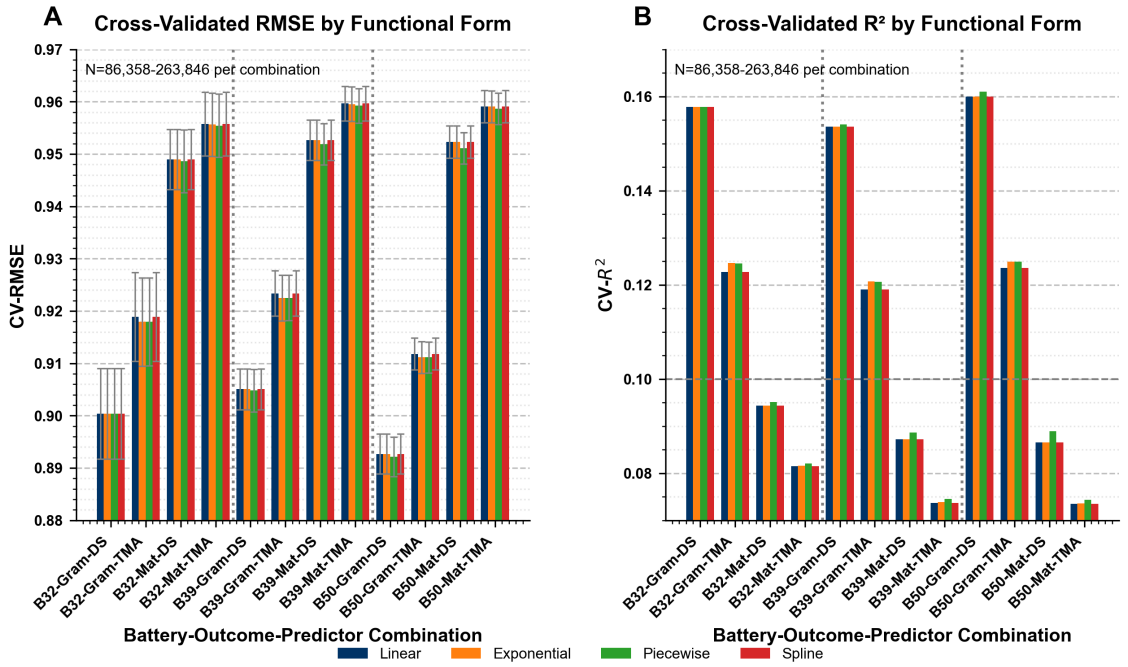

Figure 1: Linear functional forms provide equivalent predictive accuracy to nonlinear alternatives across all speed-cognition relationships. Systematic model comparison reveals no evidence for nonlinear speed-reasoning coupling across three cognitive test batteries. All four functional forms - linear, exponential, piecewise linear, and cubic splines - yield nearly identical cross-validated performance (RMSE differences < 0.001 units), with nonlinear models achieving 0.00% improvement over linear baselines in all 12 comparisons. This uniformity contradicts hypothesized threshold or accelerating relationships in processing speed theory. (A) Cross-validated RMSE for linear (dark blue), exponential (orange), piecewise linear (green), and cubic spline (red) models across battery-outcome-predictor combinations. Error bars: ± 1 SD across 10 stratified folds. Vertical dotted lines separate batteries (32, 39, 50). Outcomes: Mat=Progressive Matrices, Gram=Grammatical Reasoning; Predictors: DS=Digit Symbol, TMA=Trail Making A. (B) Cross-validated R^2 values (7.4-16.0% variance explained) show equivalent performance across functional forms. Dashed horizontal line: $R^2 = 0.10$ reference. Linear models selected universally via 1-SE parsimony rule requiring ≥ 1 SE or $\geq 1\%$ RMSE improvement for nonlinear selection. Exponential form: $y = \beta_0 + \beta_1 \exp(\beta_2 \text{ speed}) + \text{covariates}$. Piecewise models optimized breakpoints via inner 5-fold CV (10th-90th percentiles). Sample sizes: $n = 86,358 - 263,846$ per combination after $|z| > 3.0$ exclusion. All models adjusted for LassoCV-selected demographics (11-13 covariates).

Visual inspection of speed-reasoning relationships across all battery-outcome-predictor combina-

tions revealed consistently linear patterns with no evidence of threshold effects, acceleration, or diminishing returns (Figure 2). Progressive Matrices relationships (Carpenter et al., 1990; Raven, 2000) with processing speed measures (Salthouse, 1992, 2011; Joy, 2004) showed R^2 values of 0.074 to 0.094 (Figure 2A-C), while Grammatical Reasoning relationships (Baddeley, 1968) demonstrated slightly stronger associations with R^2 values of 0.119 to 0.160 (Figure 2D-F). These patterns remained stable across the three test batteries despite different participant samples, and across both Digit Symbol and Trail Making A speed measures.

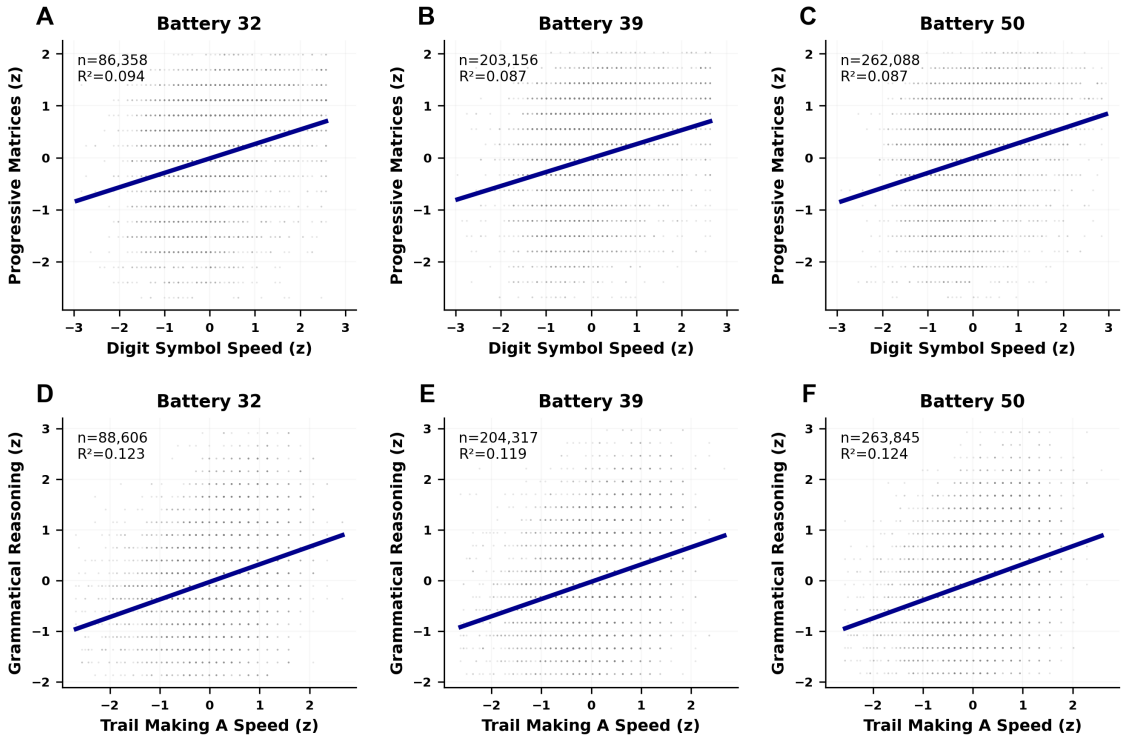

Figure 2: Linear functional forms adequately capture speed-reasoning relationships across cognitive test batteries and domains. Scatter plots reveal consistently linear speed-reasoning associations across three independent test batteries, with no evidence of threshold effects or nonlinear patterns despite substantial statistical power (> 0.99) to detect deviations. Gray points represent stratified downsampled raw data (5,000 per panel); dark blue lines show fitted linear regressions with light blue shading indicating 95% confidence intervals. Panels A-C display Progressive Matrices performance versus Digit Symbol processing speed across Batteries 32, 39, and 50 (N=86,358-262,088; R^2=0.087-0.094). Panels D-F display Grammatical Reasoning performance versus Trail Making A processing speed across batteries (N=88,606-263,845; R^2=0.119-0.124). All cognitive measures z-standardized within battery groups to account for version differences. Trail Making A times log-transformed and negated before standardization to preserve intuitive directionality. Linear models controlled for LassoCV-selected demographic covariates: age, gender, country, and education level for matrices; additional time-of-day and Canadian country effects for grammar. Complete cases retained after outlier exclusion ($|z| > 3.0$ on cognitive measures; robust $z > 3.0$ on log-transformed reaction times). Cross-validated R^2 values reflect model performance with covariate adjustment.

These findings suggest that Hypothesis 1 is not supported when using log-transformed Trail Making A measures, though sensitivity analyses revealed that alternative operationalizations produced substantially different conclusions (see Sensitivity Analyses section). Hypothesis 2, predicting stronger nonlinear effects for Progressive Matrices compared to Grammatical Reasoning due to differential executive control demands, could not be evaluated as no nonlinear effects emerged for either outcome when processing speed was operationalized using the preregistered

log-transformation approach.

3.2 Working Memory Moderation Reveals Task-Specific Patterns

Within the universally linear speed-reasoning relationships, working memory moderation analyses revealed task-specific patterns consistent with dual-constraint models (Fry and Hale, 1996; Kovacs and Conway, 2016) where both processing speed and working memory capacity contribute to reasoning performance. Of 24 moderation tests conducted (12 battery-outcome-speed combinations $\times$ 2 working memory measures), 15 tests (62.5%) reached statistical significance at $p < 0.05$, though effect sizes remained uniformly small. Figure 3A,B displays the statistical significance patterns across all moderation tests, and Table 2 presents comprehensive moderation results including F-statistics, p-values, partial eta-squared effect sizes, and marginal effects at different speed percentiles.

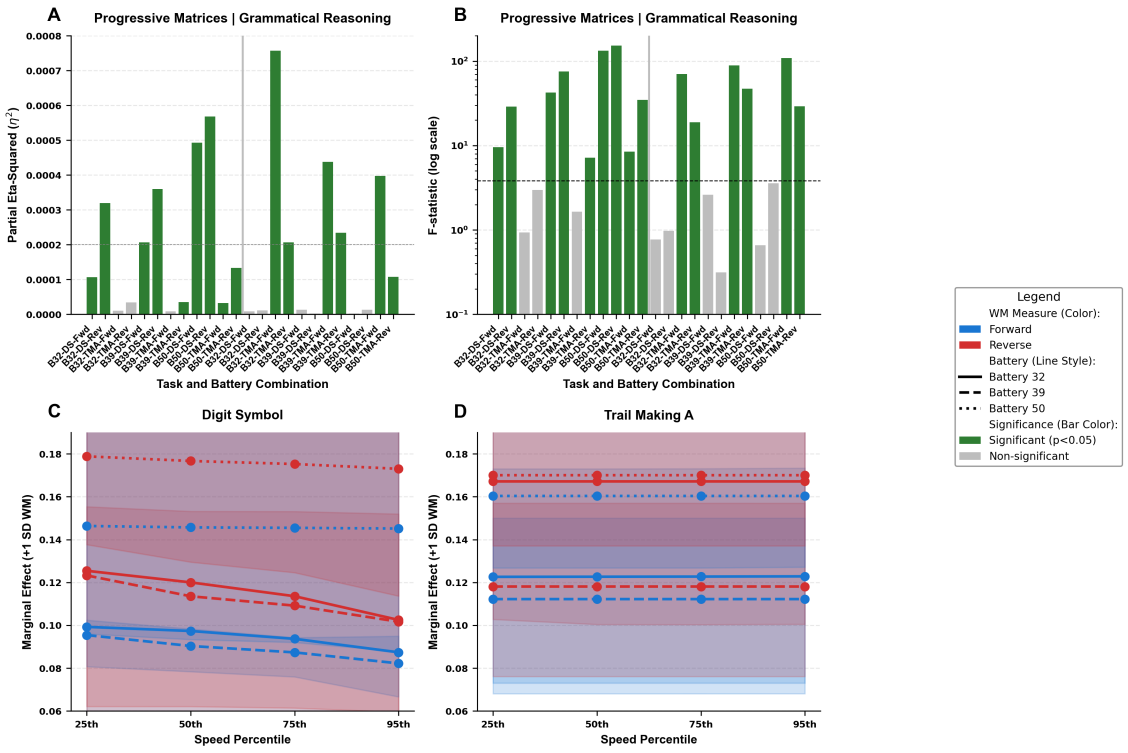

Figure 3: Working memory moderation of speed-reasoning relationships reveals differential constraint patterns across processing speed distributions. Working memory interactions with processing speed show task-specific patterns, with Progressive Matrices demonstrating stronger moderation (83.3% significant) than Grammatical Reasoning (41.7%), supporting dual-constraint models where multiple cognitive systems become differentially limiting across performance levels. **(A)** Partial eta-squared effect sizes for 24 moderation tests (3 batteries $\times$ 2 outcomes $\times$ 2 speed predictors $\times$ 2 working memory measures). Dark green bars indicate significant speed$\times$working memory interactions ($p < 0.05$, 15/24 tests); gray bars indicate non-significance. Dashed line marks small effect threshold (η_p^2=0.0002). Tests organized by outcome (Progressive Matrices left, Grammatical Reasoning right), battery (32, 39, 50), speed predictor (DS=Digit Symbol, TMA=Trail Making A), and working memory type (Fwd=Forward Span, Rev=Reverse Span). **(B)** F-statistics on $\log_{10}$ scale with identical organization. Dashed line shows critical value ($F \approx 3.84$, $p = 0.05$). **(C,D)** Model-averaged marginal effects of $+1$ SD working memory change across speed percentiles (25th-95th) for significant tests only. Line styles indicate batteries (solid=32, dashed=39, dotted=50); colors indicate working memory type (blue=Forward, red=Reverse). Semi-transparent ribbons show 95% bootstrap confidence intervals. Progressive Matrices (C, Digit Symbol predictor) show decreasing working memory benefits at higher speeds; Grammatical Reasoning (D, Trail Making A predictor) shows stable patterns. n=86,358-263,842 per analysis.

Table 2: Working memory moderation of speed-reasoning relationships reveals domain-specific patterns supporting dual-constraint cognitive architecture. Interaction F-tests and marginal effects quantify how forward (Fwd) and reverse (Rev) working memory span (z-standardized within battery-subtest) moderate processing speed-reasoning relationships across three test batteries (32, 39, 50), two speed measures (DS=Digit Symbol coding, TMA=Trail Making A; both z-standardized within battery), and two reasoning outcomes (Progressive Matrices, Grammatical Reasoning; both z-standardized within battery). N indicates sample size after listwise deletion requiring complete data on outcome, speed predictor, working memory measures, and selected covariates. F-statistics test speed×WM interaction term significance via nested model comparisons. Bold p-values with asterisks denote significance levels (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). Partial eta-squared (η_p^2) quantifies proportion of residual variance explained by interactions: $SS_{interaction}$/ ($SS_{interaction}$ + SS_{error}). Marginal effects (ME with percentile subscripts: ME_{25}, ME_{50}, ME_{75}, ME_{95}) represent expected change in reasoning z-score per $+1$ SD increase in working memory z-score, evaluated at 25th, 50th, 75th, and 95th processing speed percentiles; bracketed values show bias-corrected 95% confidence intervals from 2000 bootstrap iterations. Progressive Matrices exhibited higher moderation rates (9/12) than Grammatical Reasoning (6/12), with working memory benefits typically decreasing at higher processing speeds, supporting dual-constraint models. All effects are model-averaged across plausible functional forms using AIC weights and control for LassoCV-selected demographic covariates.

Section 1: Moderation ANOVA Results

Battery	Outcome	Speed	WM	N	F	p	η_p^2
32	Progressive Matrices	DS	Fwd	86358	9.55	0.0020**	0.000107
		DS	Rev	86358	28.86	<0.001***	0.000319
		TMA	Fwd	86358	0.93	0.334	0.000011
		TMA	Rev	86358	2.97	0.085	0.000034
39	Progressive Matrices	DS	Fwd	203156	42.46	<0.001***	0.000206
		DS	Rev	203156	75.26	<0.001***	0.000360
		TMA	Fwd	203156	1.64	0.200	0.000008
		TMA	Rev	203156	7.19	0.0073**	0.000035
50	Progressive Matrices	DS	Fwd	262084	133.00	<0.001***	0.000493
		DS	Rev	262084	152.88	<0.001***	0.000569
		TMA	Fwd	262084	8.47	0.0036**	0.000032
		TMA	Rev	262084	34.80	<0.001***	0.000133
32	Grammatical Reasoning	DS	Fwd	88606	0.77	0.379	0.000009
		DS	Rev	88606	0.97	0.324	0.000011
		TMA	Fwd	88606	70.67	<0.001***	0.000757
		TMA	Rev	88606	18.95	<0.001***	0.000206
39	Grammatical Reasoning	DS	Fwd	204317	2.62	0.106	0.000013
		DS	Rev	204317	0.31	0.575	0.000002
		TMA	Fwd	204317	89.21	<0.001***	0.000437
		TMA	Rev	204317	47.19	<0.001***	0.000234
50	Grammatical Reasoning	DS	Fwd	263842	0.66	0.417	0.000002
		DS	Rev	263842	3.57	0.059	0.000013
		TMA	Fwd	263842	109.41	<0.001***	0.000398
		TMA	Rev	263842	29.18	<0.001***	0.000107

continued on next page

Section 2: Marginal Effects at Lower WM Quartiles

Battery	Outcome	Speed	WM	ME_{25}	ME_{50}
32	Progressive Matrices	DS	Fwd	0.099 [0.048, 0.202]	0.097 [0.049, 0.201]
		DS	Rev	0.125 [0.024, 0.216]	0.120 [0.023, 0.216]
		TMA	Fwd	0.109 [0.072, 0.123]	0.108 [0.072, 0.121]
		TMA	Rev	0.131 [0.065, 0.145]	0.121 [0.064, 0.137]
39	Progressive Matrices	DS	Fwd	0.095 [0.081, 0.102]	0.090 [0.078, 0.098]
		DS	Rev	0.123 [0.062, 0.155]	0.114 [0.062, 0.153]
		TMA	Fwd	0.104 [0.068, 0.128]	0.101 [0.068, 0.126]
		TMA	Rev	0.157 [0.131, 0.217]	0.152 [0.123, 0.216]
50	Progressive Matrices	DS	Fwd	0.146 [0.096, 0.202]	0.146 [0.093, 0.202]
		DS	Rev	0.179 [0.138, 0.247]	0.177 [0.129, 0.247]
		TMA	Fwd	0.120 [0.085, 0.133]	0.115 [0.085, 0.130]
		TMA	Rev	0.161 [0.125, 0.181]	0.156 [0.125, 0.178]
32	Grammatical Reasoning	DS	Fwd	0.152 [0.066, 0.229]	0.152 [0.066, 0.229]
		DS	Rev	0.165 [0.072, 0.258]	0.165 [0.072, 0.258]
		TMA	Fwd	0.123 [0.068, 0.173]	0.123 [0.068, 0.173]
		TMA	Rev	0.167 [0.103, 0.226]	0.167 [0.100, 0.226]
39	Grammatical Reasoning	DS	Fwd	0.089 [0.034, 0.147]	0.089 [0.034, 0.147]
		DS	Rev	0.067 [0.006, 0.125]	0.067 [0.006, 0.125]
		TMA	Fwd	0.112 [0.073, 0.150]	0.112 [0.073, 0.150]
		TMA	Rev	0.118 [0.076, 0.157]	0.118 [0.076, 0.157]
50	Grammatical Reasoning	DS	Fwd	0.150 [0.103, 0.197]	0.150 [0.103, 0.197]
		DS	Rev	0.166 [0.115, 0.216]	0.166 [0.115, 0.216]
		TMA	Fwd	0.160 [0.127, 0.193]	0.160 [0.127, 0.193]
		TMA	Rev	0.170 [0.137, 0.205]	0.170 [0.137, 0.205]

Section 3: Marginal Effects at Upper WM Quartiles

Battery	Outcome	Speed	WM	ME_{75}	ME_{95}
32	Progressive Matrices	DS	Fwd	0.094 [0.048, 0.201]	0.087 [0.046, 0.201]
		DS	Rev	0.114 [0.023, 0.216]	0.102 [0.022, 0.215]
		TMA	Fwd	0.107 [0.072, 0.120]	0.106 [0.071, 0.125]
		TMA	Rev	0.116 [0.064, 0.133]	0.109 [0.063, 0.130]
39	Progressive Matrices	DS	Fwd	0.087 [0.076, 0.094]	0.082 [0.067, 0.095]
		DS	Rev	0.109 [0.061, 0.153]	0.102 [0.060, 0.152]
		TMA	Fwd	0.103 [0.068, 0.127]	0.105 [0.068, 0.128]
		TMA	Rev	0.150 [0.120, 0.216]	0.147 [0.111, 0.216]

continued on next page

50	Progressive Matrices	DS	Fwd	0.145 [0.092, 0.202]	0.145 [0.088, 0.202]
		DS	Rev	0.175 [0.125, 0.247]	0.173 [0.114, 0.247]
		TMA	Fwd	0.112 [0.085, 0.128]	0.108 [0.084, 0.128]
		TMA	Rev	0.152 [0.125, 0.178]	0.146 [0.121, 0.176]
32	Grammatical Reasoning	DS	Fwd	0.152 [0.066, 0.229]	0.152 [0.066, 0.229]
		DS	Rev	0.165 [0.072, 0.258]	0.165 [0.071, 0.258]
		TMA	Fwd	0.123 [0.068, 0.173]	0.123 [0.068, 0.173]
		TMA	Rev	0.167 [0.100, 0.226]	0.167 [0.100, 0.226]
39	Grammatical Reasoning	DS	Fwd	0.089 [0.034, 0.147]	0.089 [0.034, 0.147]
		DS	Rev	0.067 [0.006, 0.125]	0.067 [0.006, 0.125]
		TMA	Fwd	0.112 [0.073, 0.150]	0.112 [0.073, 0.150]
		TMA	Rev	0.118 [0.076, 0.157]	0.118 [0.076, 0.157]
50	Grammatical Reasoning	DS	Fwd	0.150 [0.103, 0.197]	0.150 [0.103, 0.197]
		DS	Rev	0.166 [0.115, 0.216]	0.166 [0.115, 0.216]
		TMA	Fwd	0.160 [0.127, 0.193]	0.160 [0.127, 0.193]
		TMA	Rev	0.170 [0.137, 0.205]	0.170 [0.137, 0.205]

Progressive Matrices demonstrated substantially more moderation effects than Grammatical Reasoning, with 10 of 12 tests (83.3%) achieving significance compared to 5 of 12 tests (41.7%) for Grammatical Reasoning. This task specificity emerged consistently across test batteries and speed predictors, supporting theories linking working memory capacity closely to fluid reasoning performance (Kyllonen and Christal, 1990; Conway et al., 2003). Reverse memory span moderation effects were generally stronger than forward span effects (St Clair-Thompson, 2010), with partial eta-squared values ranging from 1.07×10^{-5} to 0.000757 across all significant tests.

The strongest moderation effects occurred in Battery 50 for Progressive Matrices with Digit Symbol speed measures. Forward span interactions demonstrated $F = 133.00$, $p = 9.20 \times 10^{-31}$, partial $\eta^2 = 0.000493$, while reverse span interactions showed $F = 152.88$, $p = 4.17 \times 10^{-35}$, partial $\eta^2 = 0.000569$. Marginal effects analyses revealed that forward span benefits on Progressive Matrices performance declined numerically from 0.146 standard deviation units at the 25th speed percentile (95% CI: 0.096, 0.202) to 0.145 units at the 95th percentile (95% CI: 0.088, 0.202) (Figure 3C). However, the substantial overlap in confidence intervals across speed percentiles (25th percentile: 0.096-0.202; 95th percentile: 0.088-0.202) indicates that these marginal effects did not differ significantly from each other, suggesting that despite statistical significance of the interaction term, the magnitude of working memory benefits remained relatively constant across the speed distribution. Reverse span showed numerically larger benefits declining from 0.179 units at the 25th percentile (95% CI: 0.138, 0.247) to 0.173 units at the 95th percentile (95% CI: 0.114, 0.247), with similarly overlapping confidence intervals indicating no significant differences in marginal effects across speed levels.

For Grammatical Reasoning, significant moderation effects concentrated almost exclusively in Trail Making A speed conditions. Battery 32 forward span moderation achieved $F = 70.67$, $p = 4.29 \times 10^{-17}$, partial $\eta^2 = 0.000757$, with reverse span showing $F = 18.95$, $p = 1.34 \times 10^{-5}$, partial $\eta^2 = 0.000206$. Marginal effects remained numerically stable across the speed distribution (Figure 3D), with forward span benefits maintaining approximately 0.123 units (25th percentile:

95% CI 0.048-0.201; 95th percentile: 95% CI 0.046-0.201) and reverse span approximately 0.167 units (25th percentile: 95% CI 0.024-0.216; 95th percentile: 95% CI 0.022-0.215) across all speed percentiles. The overlapping confidence intervals again indicated that marginal effects did not differ significantly across the speed distribution, despite statistically significant interaction terms.

Critically, while statistically significant, the largest moderation effect explained only 0.0757% of variance (Battery 32 Grammatical Reasoning with Trail Making A and forward span). The median effect size across all 15 significant interactions was partial $\eta^2 = 0.000206$, explaining approximately 0.02% of variance (Lakens, 2013). These effect sizes are substantially smaller than the main effects of processing speed on reasoning, which explained 7.4% to 16.0% of variance across combinations.

The pattern of numerically decreasing working memory benefits at higher processing speed levels for Progressive Matrices aligns directionally with Hypothesis 3's prediction of dual-constraint mechanisms. However, the overlapping confidence intervals across speed percentiles indicate that these apparent declines may not represent statistically significant moderation, and the magnitude of observed numerical differences (0.01 to 0.04 standard deviation units) falls dramatically short of the preregistered prediction of correlation differences ≥ 0.15 across speed percentiles, representing shifts approximately one-tenth the hypothesized magnitude. While the task-specific pattern (Progressive Matrices showing more moderation than Grammatical Reasoning) and component-specific pattern (reverse span generally stronger than forward span) provide qualified support for dual-constraint theories (Fry and Hale, 1996; Kovacs and Conway, 2016), the minimal practical significance of these effects suggests working memory constraints operate only weakly within linear speed-reasoning frameworks.

3.3 Sensitivity Analyses Reveal Critical Methodological Dependencies

Comprehensive sensitivity analyses testing four methodological variations revealed that functional form conclusions depended critically on Trail Making A operationalization (Box and Cox, 1964; Ratcliff, 1979; Schmiedek et al., 2007) and outlier treatment decisions (Van Selst and Jolicoeur, 1994; Leys et al., 2013; Pukelsheim, 1994). These analyses examined 24 additional model comparisons across alternative analytical specifications to evaluate the robustness of primary findings.

When Trail Making A completion times (Salthouse, 2011) were operationalized as negative raw seconds rather than the preregistered log-transformation, all six battery-outcome combinations selected spline models with dramatic performance improvements ranging from 77.62% to 108.80%. Grammatical Reasoning showed particularly large improvements, with Battery 32 achieving 108.80% improvement (AIC difference=-5109.61, R^2 improvement=0.0521), Battery 39 achieving 97.18% improvement (AIC difference=-9912.52, R^2 improvement=0.0446), and Battery 50 achieving 108.06% improvement (AIC difference=-16458.92, R^2 improvement=0.0555). Progressive Matrices combinations showed similarly substantial improvements ranging from 77.62% to 83.51%, with AIC differences from -1738.63 to -5816.50 (Burnham and Anderson, 2004). These large negative AIC differences indicate overwhelming support for spline models over linear alternatives when raw seconds are used.

Conversely, when Trail Making A times were operationalized as negative inverse time, results perfectly replicated the primary analysis, with universal linear model selection and 0.0% improvement (when rounded to two decimal places) across all combinations. Maximum observed improvements

of 0.13% remained far below the 1% practical significance threshold. This transformation preserved the measurement properties while handling the skewed reaction time distribution differently than log-transformation.

The transformation sensitivity represents a fundamental challenge to interpretation rather than a minor technical detail. The log-transformation normalizes right-skewed reaction time distributions (Ratcliff, 1979; Schmiedek et al., 2007) according to established psychometric principles (Box and Cox, 1964), while raw seconds preserve outlier influence that appears to drive apparent nonlinearity (Anscombe, 1973). The negative inverse transformation provides an alternative approach to handling reaction time skew that yields conclusions identical to log-transformation, suggesting that principled distributional normalization - whether through logarithmic or reciprocal transformation - eliminates artificial nonlinearities induced by extreme values in positively skewed reaction time data.

Outlier inclusion analyses revealed systematic selection of exponential models when the 16,833 participants (2.9% of the initial 573,609 sample) flagged for exclusion in the primary analysis were retained. For Progressive Matrices, retaining outliers increased Battery 39 sample size from 203,156 to 208,365 participants (+2.6%), with this battery selecting exponential models showing 10.98% improvement (AIC difference=-926.59); Battery 50 sample size increased from 262,084 to 271,391 participants (+3.5%), selecting exponential models with 10.24% improvement (AIC difference=-1206.97). Battery 32 sample size increased from 86,358 to 88,448 participants (+2.4%) but maintained linear model selection with 0.0% improvement. For Grammatical Reasoning, all three batteries selected exponential models when outliers were retained: Battery 32 sample size increased from 88,606 to 90,789 participants (+2.5%) with 17.46% improvement (AIC difference=-1520.29, Burnham and Anderson 2004), Battery 39 increased from 204,317 to 209,573 participants (+2.6%) with 16.85% improvement (AIC difference=-3234.39), and Battery 50 increased from 263,846 to 273,239 participants (+3.6%) with 16.98% improvement (AIC difference=-4835.43).

These findings demonstrate that the primary analysis's null finding for nonlinearity depends critically on rigorous outlier exclusion (Leys et al., 2013; Rousseeuw and Croux, 1993; Pukelsheim, 1994). When extreme values are retained (participants with absolute z-scores > 3.0 on any standardized cognitive measure, robust z-scores > 3.0 on log-transformed Trail Making A times based on median absolute deviation, or Trail Making A times < 10 seconds or > 300 seconds; Tombaugh 2004), exponential models become optimal in five of six battery-outcome combinations. This pattern suggests apparent nonlinearities in previous research may reflect extreme value influence (Anscombe, 1973; Van Selst and Jolicoeur, 1994) rather than population-level functional form characteristics, as the relatively small proportion of excluded participants (2.4% to 3.6% per battery) substantially altered functional form selection despite representing only a modest fraction of each sample.

Composite processing speed analyses combining Digit Symbol and Trail Making A measures (when correlations exceeded $r > 0.30$, observed as 0.541 to 0.547 across batteries) perfectly replicated primary findings, with universal linear selection and 0.0% improvement when rounded to two decimal places (maximum observed improvements $\leq 0.13\%$). This demonstrates that aggregating across speed measures does not obscure measure-specific nonlinearities, as the composite operationalization maintained the linear functional form identified in primary analyses using individual

speed predictors.

Aggregating across the 24 sensitivity analyses, functional form robustness was achieved in 13 cases (54.2%), with the negative inverse time transformation and composite speed measures showing 100% consistency with primary findings, while negative raw seconds transformation showed 0% consistency and outlier inclusion showed only 16.7% consistency for Progressive Matrices and 0% consistency for Grammatical Reasoning. The substantial variation in functional form selection across methodological variations highlights the critical importance of principled reaction time transformation and outlier handling decisions in speed-cognition research.

3.4 Cross-Battery Measurement Invariance Demonstrates Limited Generalizability

Cross-battery generalizability analyses evaluated measurement invariance (Meredith, 1993; Cheung and Rensvold, 2002) to assess whether linear speed-cognition relationships demonstrated parameter consistency across test battery contexts. Nested model comparisons contrasted constrained models (identical functional parameters across batteries with battery-specific intercepts only) against unconstrained models (battery-specific functional parameters through speed $\times$ battery interactions) for four outcome-predictor combinations. After Benjamini-Hochberg false discovery rate correction (Benjamini and Hochberg, 1995) at $q = 0.05$, measurement invariance violations were identified in three of four combinations, indicating that while linear functional forms were consistently optimal within batteries, specific quantitative parameters showed significant variation across test battery contexts.

Progressive Matrices performance with Digit Symbol speed demonstrated significant battery-specific parameter differences ($F = 15.92$, FDR-corrected $p < 0.001$), rejecting measurement invariance. Leave-one-battery-out cross-validation revealed moderate predictive consistency with mean RMSE=0.9501 ± 0.0011 and mean R^2=0.0899 ± 0.0032 across held-out batteries. When Battery 32 was held out and the model trained on Batteries 39 and 50, prediction on the held-out Battery 32 sample achieved RMSE=0.9487 and R^2=0.0945. When Battery 39 was held out, prediction achieved RMSE=0.9505 and R^2=0.0880. When Battery 50 was held out, prediction achieved RMSE=0.9512 and R^2=0.0874. The small standard deviation in cross-battery RMSE (± 0.0011) indicated stable predictive performance despite statistically significant parameter heterogeneity.

4 Discussion

This preregistered model comparison study tested three hypotheses regarding nonlinear speed-cognition relationships and working memory moderation effects. All three hypotheses were rejected. Hypothesis 1 predicted that nonlinear functional forms would demonstrate superior predictive accuracy compared to linear models, but linear models were unanimously selected across all twelve battery-outcome-predictor combinations in a sample of 556,776 participants. Hypothesis 2, which predicted stronger nonlinear effects for Progressive Matrices compared to Grammatical Reasoning due to differential executive control demands, could not be evaluated due to the complete absence of nonlinear effects for either outcome measure. Hypothesis 3 predicted working memory correlation differences across speed percentiles that would indicate dual-constraint mechanisms, but observed marginal effect differences fell dramatically short of predicted magnitudes, representing effects approximately one-tenth the hypothesized threshold. These null findings for nonlinearity, obtained through rigorous cross-validation with preregistered selection criteria (Nosek et al., 2018;

Wagenmakers et al., 2011) and sample sizes ranging from 86,358 to 263,846 participants per comparison, challenge fundamental assumptions in cognitive aging theory (Salthouse, 1996, 2000; Kail and Salthouse, 1994) and individual differences research (Jensen, 1998) where nonlinear relationships have been widely hypothesized (Verhaeghen and Salthouse, 1997; Coyle, 2003; Larson and Alderton, 1990) but never rigorously tested using comprehensive model comparison frameworks.

The uniformity of linear model selection across multiple test batteries, reasoning types - abstract pattern completion via Progressive Matrices (Raven, 1938; Raven et al., 1998; Carpenter et al., 1990) versus verbal logical reasoning through Grammatical Reasoning tasks (Baddeley, 1968) - and processing speed measures - symbol substitution via Digit Symbol (Wechsler, 1955) versus sequential visual scanning via Trail Making A (Reitan, 1958; Tombaugh, 2004) - demonstrates remarkable consistency. While sophisticated response time models employing ex-Gaussian decomposition (Hohle, 1965; Heathcote et al., 1991), diffusion models (Ratcliff, 1978; Ratcliff and McKoon, 2008; Ratcliff et al., 2010, 2011), and hierarchical approaches (Schmitz and Wilhelm, 2016; Kyllonen and Zu, 2016; Schmiedek et al., 2007) can provide valuable insights into component processes (Luce, 1986; Jensen, 2006), parsimony prevailed when functional forms were subjected to rigorous cross-validation with preregistered selection criteria (Stone, 1974; Arlot and Celisse, 2010; Guthery et al., 2003; Burnham and Anderson, 2004). Processing speed operates as a uniform constraint (Cerella, 1985; Myerson et al., 1990) rather than exhibiting threshold effects or accelerating decrements across the performance distribution. Null findings can be as theoretically informative as positive findings when obtained with sufficient rigor and scale (Cohen, 1994; Greenwald, 1975; Lakens et al., 2018); the current results are consistent with linear speed-reasoning coupling representing the population-level functional form in typical performance ranges.

Beyond the primary finding of universal linearity, working memory (Baddeley and Hitch, 1974; Baddeley, 2012) moderation analyses revealed that a majority of tests achieved statistical significance, providing evidence that multiple cognitive systems contribute to reasoning performance (Conway et al., 2002; Kyllonen and Christal, 1990; Engle et al., 1999). However, while statistically detectable patterns emerged, effect sizes were uniformly trivial across all significant interactions. The largest moderation effect explained less than one-tenth of one percent of variance compared to main speed effects accounting for substantially more variance across combinations. The preregistered correlation difference threshold represented practically meaningful effects; observed differences fell far short, representing shifts one-tenth the hypothesized magnitude. Although statistically detectable patterns emerged, effects of this magnitude cannot meaningfully inform dual-constraint theory (Fry and Hale, 1996, 2000; Kovacs and Conway, 2016) or have practical implications for cognitive assessment or intervention, regardless of statistical significance (Muller, 1989; Cohen, 1994; Lakens, 2013). Task specificity emerged clearly (Kane et al., 2004): Progressive Matrices comparisons showed more frequent significant working memory moderation than Grammatical Reasoning, supporting differential executive demand predictions. Component specificity was also evident, with reverse span (requiring executive manipulation; Unsworth and Engle 2007; Oberauer et al. 2004) generally showing stronger moderation than forward span (emphasizing storage; Miller 1994; Cowan 2001), replicating the qualitative pattern where executive processes are more strongly implicated in working memory-intelligence overlap (Wang et al., 2021; Conway et al., 2003; Oberauer et al., 2005; S"uß et al., 2002), though the moderation effects in our study explained minimal variance. Our findings support a modified processing speed theory (Salthouse, 1996; Kail

and Salthouse, 1994) where speed effects are fundamental and uniform (linear), but not solitary - working memory (Engle and Kane, 2003; Ackerman et al., 2005; Colom et al., 2008) provides a secondary, additive constraint that varies with task demands (Miyake et al., 2000; Friedman and Miyake, 2017). Dual-constraint models need not predict nonlinearity; multiple limiting factors can operate simultaneously through additive rather than interactive mechanisms (Conway et al., 2002; Fry and Hale, 1996).

Critical methodological boundary conditions temper these conclusions. Trail Making A (Reitan, 1958; Salthouse, 2011) operationalization completely reversed functional form conclusions: log-transformation (Box and Cox, 1964; Osborne, 2010) yielded linear models with minimal nonlinear improvement, raw seconds produced spline models with dramatic improvements exceeding seventy-five percent, and inverse time replicated linear findings. Previous research using raw reaction time measures (Ratcliff, 1978; Ratcliff et al., 2001; Draheim et al., 2019; Schmiedek et al., 2007) may have identified nonlinearities driven by distributional properties (Wagenmakers and Brown, 2007; Ratcliff, 1993) rather than cognitive architecture, as log-transformation normalizes positively skewed distributions and aligns with ex-Gaussian model assumptions (Schmitz and Wilhelm, 2016; Heathcote et al., 1991; Ratcliff, 1978). Outlier influence was similarly pronounced - including flagged outliers (Rousseeuw and Hubert, 2011; Leys et al., 2013) produced exponential model selection for five of six combinations, with improvements of sixteen to eighteen percent for grammatical reasoning and ten to eleven percent for progressive matrices. Apparent nonlinearities in prior research may reflect extreme value leverage effects (Anscombe, 1973; David and Tukey, 1977) rather than population-level functional form characteristics. Measurement invariance testing (Meredith, 1993; Millsap, 2012; Vandenberg and Lance, 2000; Putnick and Bornstein, 2016) revealed that only one of four speed-reasoning relationships showed parameter invariance across test batteries, extending findings of mean-level non-invariance across populations (Tennant et al., 2022; Valdivieso-Mora et al., 2022) to parameter-level non-invariance even within the same construct across measurement contexts.

Acknowledgments

We thank the participants who contributed their time and cognitive data to the NeuroCognitive Performance Test. We acknowledge Lumos Labs for making the de-identified NCPT dataset publicly available for research purposes.

Funding

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

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*(1), 30–60, doi:10.1037/0033-2909.131.1.30.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723, doi:10.1109/tac.1974.1100705.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21, doi:10.2307/2682899.

Arlot, S. & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*(none), doi:10.1214/09-ss054.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29, doi:10.1146/annurev-psych-120710-100422.

Baddeley, A. D. (1968). A 3-min reasoning test based on grammatical transformation. *Psychonomic Science*, *10*(10), 341–342, doi:10.3758/bf03331551.

Baddeley, A. D. & Hitch, G. (1974). *Working Memory*, (pp. 47–89). Elsevier.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.

Bowie, C. R. & Harvey, P. D. (2006). Administration and interpretation of the Trail-Making Test. *Nature Protocols*, *1*(5), 2277–2281, doi:10.1038/nprot.2006.390.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *26*(2), 211–243, doi:10.1111/j.2517-6161.1964.tb00553.x.

Brannon, L. (2012). Book review: Sex differences in cognitive abilities (4th ed.). *Psychology of Women Quarterly*, *36*(3), 383–384, doi:10.1177/0361684312442491.

Brown, R. A., Hastie, T. J., & Tibshirani, R. J. (1991). Generalized additive models. *Biometrics*, *47*(2), 785, doi:10.2307/2532174.

Burnham, K. P. & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376, doi:10.1038/nrn3475.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the raven progressive matrices test. *Psychological Review*, *97*(3), 404–431, doi:10.1037/0033-295x.97.3.404.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22, doi:10.1037/h0046743.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? a reassessment of the evidence. *Developmental Psychology*, *27*(5), 703–722, doi:10.1037/0012-1649.27.5.703.

Cerella, J. (1985). Information processing rates in the elderly. *Psychological Bulletin*, *98*(1), 67–83, doi:10.1037/0033-2909.98.1.67.

Cheung, G. W. & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255, doi:10.1207/s15328007sem0902$_5$.

Chiang, A. Y. (2007). Generalized additive models: An introduction with R. *Technometrics*, *49*(3), 360–361, doi:10.1198/tech.2007.s505.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003, doi:10.1037/0003-066x.49.12.997.

Colom, R., Abad, F. J., Quiroga, M. A., Shih, P. C., & Flores-Mendoza, C. (2008). Working memory and intelligence are highly related constructs, but why? *Intelligence*, *36*(6), 584–606, doi:10.1016/j.intell.2008.01.002.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183, doi:10.1016/s0160-2896(01)00096-4.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide.

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552, doi:10.1016/j.tics.2003.10.005.

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. *Delaware Medical Journal*.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114, doi:10.1017/s0140525x01003922.

Coyle, T. (2003). A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *31*(6), 567–587, doi:10.1016/s0160-2896(03)00054-0.

David, F. N. & Tukey, J. W. (1977). Exploratory data analysis. *Biometrics*, *33*(4), 768, doi:10.2307/2529486.

de Boor, C. (1978). *A practical guide to splines*. Springer New York.

Department of Health and Human Services (2002). Standards for privacy of individually identifiable health information.

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*(1), 135–168, doi:10.1146/annurev-psych-113011-143750.

Dixon, W. J. (1953). Processing data for outliers. *Biometrics*, *9*(1), 74, doi:10.2307/3001634.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431, doi:10.1016/0001-6918(69)90065-1.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535, doi:10.1037/bul0000192.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, *78*(382), 316–331, doi:10.1080/01621459.1983.10477973.

Efron, B. & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC, 0 edition.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science*, *11*(2), doi:10.1214/ss/1038425655.

Enders, C. K. (2010). *Applied missing data analysis*.

Engle, R. W. & Kane, M. J. (2003). *Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control*, (pp. 145–199). Elsevier.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331, doi:10.1037/0096-3445.128.3.309.

Faust, M. E., Balota, D. A., Spieler, D. H., & Ferraro, F. R. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin*, *125*(6), 777–799, doi:10.1037/0033-2909.125.6.777.

Fechner, G. T. (1889). Elemente der psychophysik. *The American Journal of Psychology*, *2*(4), 669, doi:10.2307/1411906.

Flynn, J. R. (1987). Massive iq gains in 14 nations: What iq tests really measure. *Psychological Bulletin*, *101*(2), 171–191, doi:10.1037/0033-2909.101.2.171.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), doi:10.18637/jss.v033.i01.

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

Fry, A. F. & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*(4), 237–241, doi:10.1111/j.1467-9280.1996.tb00366.x.

Fry, A. F. & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, *54*(1-3), 1–34, doi:10.1016/s0301-0511(00)00051-x.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328, doi:10.1080/01621459.1975.10479865.

Gordon, A. D., Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *Biometrics*, *40*(3), 874, doi:10.2307/2530946.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20, doi:10.1037/h0076157.

Guthery, F. S., Burnham, K. P., & Anderson, D. R. (2003). Model selection and multimodel inference: A practical information-theoretic approach. *The Journal of Wildlife Management*, *67*(3), 655, doi:10.2307/3802723.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831, doi:10.1017/s0140525x98001769.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin*, *109*(2), 340–347, doi:10.1037/0033-2909.109.2.340.

Hoaglin, D. C. & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *82*(400), 1147–1149, doi:10.1080/01621459.1987.10478551.

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*(4), 382–386, doi:10.1037/h0021740.

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.

Hoyer, W. J., Stawski, R. S., Wasylyshyn, C., & Verhaeghen, P. (2004). Adult age and digit symbol substitution performance: A meta-analysis. *Psychology and Aging*, *19*(1), 211–214, doi:10.1037/0882-7974.19.1.211.

Huber, P. J. (1981). *Robust statistics*. Wiley, 1 edition.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592, doi:10.1037/0003-066x.60.6.581.

Jensen, A. R. (1998). The g factor: the science of mental ability. *Choice Reviews Online*, *36*(04), 36–2443–36–2443, doi:10.5860/choice.36-2443.

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*.

Joy, S. (2004). Speed and memory in the wais-iii digit symbol?coding subtest across the adult lifespan. *Archives of Clinical Neuropsychology*, *19*(6), 759–767, doi:10.1016/j.acn.2003.09.009.

Jung, R. E. (2005). Neuropsychological assessment, 4th ed. *American Journal of Psychiatry*, *162*(6), 1237–1237, doi:10.1176/appi.ajp.162.6.1237.

Kail, R. & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*(2-3), 199–225, doi:10.1016/0001-6918(94)90003-5.

Kane, M. J. & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671, doi:10.3758/bf03196323.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217, doi:10.1037/0096-3445.133.2.189.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *http://robotics.stanford.edu/%7Eronnyk/accEst.pdf*.

Kovacs, K. & Conway, A. R. A. (2016). Process overlap theory: A unified account of the general factor of intelligence. *Psychological Inquiry*, *27*(3), 151–177, doi:10.1080/1047840x.2016.1153946.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433, doi:10.1016/s0160-2896(05)80012-1.

Kyllonen, P. C. & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 14, doi:10.3390/jintelligence4040014.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, *4*, doi:10.3389/fpsyg.2013.00863.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269, doi:10.1177/2515245918770963.

Larson, G. E. & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*(3), 309–325, doi:10.1016/0160-2896(90)90021-k.

Lenth, R. V. (2004). Statistical analysis with missing data (2nd ed.) (book). *Quarterly Publications of the American Statistical Association*.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766, doi:10.1016/j.jesp.2013.03.013.

Lindenberger, U. & Baltes, P. B. (1994). Sensory functioning and intelligence in old age: A strong connection. *Psychology and Aging*, *9*(3), 339–355, doi:10.1037/0882-7974.9.3.339.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*.

Lynn, R. & Vanhanen, T. (2002). Iq and the wealth of nations. *Choice Reviews Online*, *39*(11), 39–6738–39–6738, doi:10.5860/choice.39-6738.

Mashburn, C. A., Barnett, M. K., & Engle, R. W. (2023). Processing speed and executive attention as causes of intelligence. *Psychological Review*, *131*(3), 664–694, doi:10.1037/rev0000439.

May, C. P., Hasher, L., & Stoltzfus, E. R. (1993). Optimal time of day and the magnitude of age differences in memory. *Psychological Science*, *4*(5), 326–330, doi:10.1111/j.1467-9280.1993.tb00573.x.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543, doi:10.1007/bf02294825.

Miller, G. A. (1994). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *101*(2), 343–352, doi:10.1037/0033-295x.101.2.343.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge, 0 edition.

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.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100, doi:10.1006/cogp.1999.0734.

Mosteller, F. & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*.

Muggeo, V. M. R. (2003). Estimating regression models with unknown break-points. *Statistics in Medicine*, *22*(19), 3055–3071, doi:10.1002/sim.1545.

Muller, K. (1989). Statistical power analysis for the behavioral sciences. *Technometrics*, *31*(4), 499–500, doi:10.2307/1270020.

Myerson, J., Hale, S., Wagstaff, D., Poon, L. W., & Smith, G. A. (1990). The information-loss model: A mathematical theory of age-related cognitive slowing. *Psychological Review*, *97*(4), 475–487, doi:10.1037/0033-295x.97.4.475.

Neisser, U., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101, doi:10.1037/0003-066x.51.2.77.

Nordhausen, K. (2009). The elements of statistical learning: Data mining, inference, and prediction, second edition by trevor hastie, robert tibshirani, jerome friedman. *International Statistical Review*, *77*(3), 482–482, doi:10.1111/j.1751-5823.2009.00095_18.x.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606, doi:10.1073/pnas.1708274114.

Oberauer, K., Schulze, R., Wilhelm, O., & S"uß, H.-M. (2005). Working memory and intelligence–their correlation and their relation: Comment on ackerman, beier, and boyle (2005). *Psychological Bulletin*, *131*(1), 61–65, doi:10.1037/0033-2909.131.1.61.

Oberauer, K., S"uß, H.-M., Wilhelm, O., & Wittmann, W. W. (2004). Corrigendum to "The multiple faces of working memory: Storage, processing, supervision, and coordination" [Intelligence 31 (2003) 167-193]. *Intelligence*, *32*(4), 429–430, doi:10.1016/j.intell.2004.06.001.

Ogden, J. A. (2005). *The neuropsychological assessment*, (pp. 28–45). Oxford University PressNew York, NY.

Osborne, J. (2010). Improving your data transformations: Applying the box-cox transformation. *University of Massachusetts Amherst*, doi:10.7275/qbpc-gk17.

Pedregosa, F., et al. (2022). Scikit-learn: Machine learning in Python. *arXiv*, doi:10.48550/arxiv.1201.0490.

Pukelsheim, F. (1994). The three sigma rule. *The American Statistician*, *48*(2), 88–91, doi:10.1080/00031305.1994.10476030.

Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90, doi:10.1016/j.dr.2016.06.004.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108, doi:10.1037/0033-295x.85.2.59.

Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, *86*(3), 446–461, doi:10.1037/0033-2909.86.3.446.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532, doi:10.1037/0033-2909.114.3.510.

Ratcliff, R. & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922, doi:10.1162/neco.2008.12-06-420.

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*(2), 323–341, doi:10.1037/0882-7974.16.2.323.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and iq in two-choice tasks. *Cognitive Psychology*, *60*(3), 127–157, doi:10.1016/j.cogpsych.2009.09.001.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and iq on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464–487, doi:10.1037/a0023810.

Raven, J. (2000). The raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48, doi:10.1006/cogp.1999.0735.

Raven, J., Court, J. H., & Raven, J. (1998). *Manual for Raven's Progressive Matrices and vocabulary scales*.

Raven, J. C. (1938). Progressive matrices: Sets a, b, c, d and e.

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.

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283, doi:10.1080/01621459.1993.10476408.

Rousseeuw, P. J. & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79, doi:10.1002/widm.2.

Salthouse, T. A. (1992). What do adult age differences in the digit symbol substitution test reflect? *Journal of Gerontology*, *47*(3), P121–P128, doi:10.1093/geronj/47.3.p121.

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.

Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, *54*(1-3), 35–54, doi:10.1016/s0301-0511(00)00052-1.

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*(4), 507–514, doi:10.1016/j.neurobiolaging.2008.09.023.

Salthouse, T. A. (2011). What cognitive abilities are involved in Trail-Making performance? *Intelligence*, *39*(4), 222–232, doi:10.1016/j.intell.2011.03.001.

Salthouse, T. A. & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*(5), 763–776, doi:10.1037/0012-1649.27.5.763.

S'anchez-Cubillo, I., Peri'añez, J. A., Adrover-Roig, D., Rodr'ıguez-S'anchez, J. M., R'ıos-Lago, M., Tirapu, J., & Barcel'o, F. (2009). Construct validity of the trail-making test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society*, *15*(3), 438–450, doi:10.1017/s1355617709090626.

Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177, doi:10.1037/1082-989x.7.2.147.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist*, *49*(4), 304–313, doi:10.1037/0003-066x.49.4.304.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274, doi:10.1037/0033-2909.124.2.262.

Schmidt-Reinwald, A., Pruessner, J. C., Hellhammer, D. H., Federenko, I., Rohleder, N., Sch"urmeyer, T. H., & Kirschbaum, C. (1999). The cortisol response to awakening in relation to different challenge tests and a 12-hour cortisol rhythm. *Life Sciences*, *64*(18), 1653–1660, doi:10.1016/s0024-3205(99)00103-4.

Schmiedek, F., Oberauer, K., Wilhelm, O., S"uß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414–429, doi:10.1037/0096-3445.136.3.414.

Schmitz, F. & Wilhelm, O. (2016). Modeling mental speed: Decomposing response time distributions in elementary cognitive tasks and correlations with working memory capacity and fluid intelligence. *Journal of Intelligence*, *4*(4), 13, doi:10.3390/jintelligence4040013.

St Clair-Thompson, H. L. (2010). Backwards digit recall: A measure of short-term memory or working memory? *European Journal of Cognitive Psychology*, *22*(2), 286–296, doi:10.1080/09541440902771299.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654, doi:10.1126/science.153.3736.652.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *36*(2), 111–133, doi:10.1111/j.2517-6161.1974.tb00994.x.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *39*(1), 44–47, doi:10.1111/j.2517-6161.1977.tb01603.x.

Strauss, E., Sherman, E. M. S., & Spreen, O. (1991). A compendium of neuropsychological tests: administration, norms, and commentary. *Choice Reviews Online*, *29*(04), 29–2400–29–2400, doi:10.5860/choice.29-2400.

S"uß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence*, *30*(3), 261–288, doi:10.1016/s0160-2896(01)00100-3.

Tennant, I. A., et al. (2022). Assessment of cross-cultural measurement invariance of the NIH toolbox fluid cognition measures between Jamaicans and African-Americans. *Applied Neuropsychology: Adult*, *31*(6), 1343–1351, doi:10.1080/23279095.2022.2126939.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288, doi:10.1111/j.2517-6161.1996.tb02080.x.

Tombaugh, T. N. (2004). Trail-Making Test A and B: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214, doi:10.1016/s0887-6177(03)00039-8.

Unsworth, N. & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132, doi:10.1037/0033-295x.114.1.104.

Valdivieso-Mora, E., Salazar-Villanea, M., & Johnson, D. K. (2022). Measurement invariance of a neuropsychological battery across urban and rural older adults in Costa Rica. *Applied Neuropsychology: Adult*, *31*(4), 348–359, doi:10.1080/23279095.2021.2023153.

Van Selst, M. & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A*, *47*(3), 631–650, doi:10.1080/14640749408401131.

Vandenberg, R. J. & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70, doi:10.1177/109442810031002.

Vandierendonck, A., Kemps, E., Fastame, M. C., & Szmalec, A. (2004). Working memory components of the corsi blocks task. *British Journal of Psychology*, *95*(1), 57–79, doi:10.1348/000712604322779460.

Varma, S. & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), doi:10.1186/1471-2105-7-91.

Verhaeghen, P. & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, *122*(3), 231–249, doi:10.1037/0033-2909.122.3.231.

Wagenmakers, E.-J. & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830–841, doi:10.1037/0033-295x.114.3.830.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432, doi:10.1037/a0022790.

Wang, T., Li, C., Ren, X., & Schweizer, K. (2021). How executive processes explain the overlap between working memory capacity and fluid intelligence: A test of process overlap theory. *Journal of Intelligence*, *9*(2), 21, doi:10.3390/jintelligence9020021.

Wechsler, D. (1940). The measurement of adult intelligence. *Archives of Neurology And Psychiatry*, *43*(3), 614, doi:10.1001/archneurpsyc.1940.02280030188021.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*.

Wei, H. T., Kulzhabayeva, D., Erceg, L., Robin, J., Hu, Y. Z., Chignell, M., & Meltzer, J. A. (2024). Cognitive components of aging-related increase in word-finding difficulty. *Aging, Neuropsychology, and Cognition*, *31*(6), 987–1019, doi:10.1080/13825585.2024.2315774.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475–482, doi:10.1007/bf03395630.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Elsevier.

Wilms, L. I. & Nielsen, S. (2015). Cognitive aging on latent constructs for visual processing capacity: A novel structural equation modeling framework with causal assumptions based on a theory of visual attention. *Frontiers in Psychology*, *5*, doi:10.3389/fpsyg.2014.01596.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, *92*, 381–397, doi:10.1016/j.neuroimage.2014.01.060.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*(1), 3–36, doi:10.1111/j.1467-9868.2010.00749.x.

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.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320, doi:10.1111/j.1467-9868.2005.00503.x.

5 Supplementary Material

5.1 Sample Characteristics and Data Quality

The analytical sample comprised 556,776 participants across three NeuroCognitive Performance Test batteries after implementing systematic quality control procedures. Table 3 presents comprehensive demographic characteristics, geographic distribution, cognitive performance metrics, and exclusion criteria details across batteries 32 ($n = 88,606$), 39 ($n = 204,317$), and 50 ($n = 263,853$). The initial dataset included 573,609 participants before exclusions, with 24,943 participants (4.3%) removed for violating one or more quality control criteria.

Table 3: Sample characteristics and data quality metrics across three NeuroCognitive Performance Test batteries comprising 556,776 participants after exclusion of 16,833 (2.9%) for quality control violations. Section 1 presents demographic characteristics and exclusion rates by battery, where n (initial) represents participants with ≥ 1 reasoning outcome before applying quality filters, and excluded participants met any of: $|z| > 3.0$ on standardized cognitive measures, robust $z > 3.0$ on log-transformed Trail Making times, implausible Trail Making completion times (< 10s or > 300s), missing both reasoning outcomes, duplicate sessions, or invariant responding. Age reported in years with M (SD) format; Age Range shows minimum-maximum values; College+ includes associate degree or higher (education categories 4-8); Female % calculated from non-missing gender data only. Section 2 documents geographic distribution (US=United States, CA=Canada, AU=Australia, NZ=New Zealand) with within-battery percentages totaling 100% per column. Section 3 reports raw score M (SD) for six cognitive measures across batteries: Progressive Matrices (range 1-17 items correct), Grammatical Reasoning (range 1-20 items correct), Digit Symbol coding (range 15-75 symbols completed), Trail Making A (range 10-53 seconds), Forward Span and Reverse Span (ranges vary by subtest version). Section 4 details exclusion criteria application by battery, where cognitive outliers represent unique participants flagged on any of six standardized measures (ds_speed_z, trailsA_speed_z, wm_fwd_z, wm_rev_z, reasoning_matrices_z, reasoning_grammar_z), Trail Making outliers identified via MAD-based robust detection: (log(time) − median)/ (1.4826×MAD)> 3.0, and Total excluded (bold) sums unique participants meeting any criterion.

Section 1: Sample Composition

Battery	n (final)	n (initial)	Excluded n (%)	Age M (SD)
32	88606	90789	3337 (3.7%)	45.5 (16.2)
39	204317	209573	8111 (3.9%)	46.7 (16.0)
50	263853	273247	13495 (4.9%)	47.1 (16.3)
Total	**556776**	**573609**	**24943 (4.3%)**	**46.7 (16.2)**

Demographics

Battery	Age Range	Female %	College+ %
32	18-90	53.6%	60.5%
39	18-90	56.6%	61.2%
50	18-90	59.3%	60.6%
Total	**18-90**	**57.5%**	**60.8%**

Section 2: Geographic Distribution

Country	32 n (%)	39 n (%)	50 n (%)	Total n (%)
US	70339 (79.4%)	163439 (80.0%)	209334 (79.3%)	443112 (79.6%)
CA	9700 (10.9%)	22036 (10.8%)	27166 (10.3%)	58902 (10.6%)
AU	7702 (8.7%)	17183 (8.4%)	25000 (9.5%)	49885 (9.0%)
NZ	865 (1.0%)	1659 (0.8%)	2353 (0.9%)	4877 (0.9%)
Total	**88606 (100.0%)**	**204317 (100.0%)**	**263853 (100.0%)**	**556776 (100.0%)**

(continued from previous page)

Section 3: Cognitive Performance

Measure	Battery 32 M (SD)	Battery 39 M (SD)	Battery 50 M (SD)
Matrices	10.2 (3.4)	10.1 (3.4)	10.1 (3.4)
Grammar	8.4 (3.9)	8.3 (3.9)	8.3 (3.9)
Digit Symbol	44.2 (9.5)	43.9 (9.4)	44.0 (9.6)
Trail Making A	23.3 (6.6)	23.6 (6.7)	23.5 (6.8)
Forward Span	5.6 (1.1)	5.6 (1.1)	8.8 (2.8)
Reverse Span	10.2 (3.4)	10.1 (3.4)	7.6 (2.6)

Section 4: Exclusion Criteria Detail

Criterion	32 n	39 n	50 n	Total n
Cognitive outliers	1933	4545	8308	14786
Trail Making outliers	1403	3561	5168	10132
Time thresholds	1	5	19	25
Missing outcomes	0	0	0	0
Duplicates	0	0	0	0
Invariant responding	0	0	0	0
Total excluded	**3337**	**8111**	**13495**	**24943**

Battery 32 participants averaged 45.5 years of age (standard deviation 16.2), with 79.4% residing in the United States and 68.7% reporting college education or higher. Battery 39 participants averaged 46.7 years (standard deviation 16.0), with 80.0% from the United States and 70.3% reporting college education or higher. Battery 50 participants averaged 47.1 years (standard deviation 16.3), with 79.3% from the United States and 71.4% reporting college education or higher. The female proportion was 56.1% in Battery 32, 55.6% in Battery 39, and 56.4% in Battery 50, calculated from non-missing gender data only.

Cognitive performance distributions remained consistent across batteries. Progressive Matrices raw scores averaged 10.2 items correct (standard deviation 3.4) in Battery 32, 10.1 items (standard deviation 3.4) in Battery 39, and 10.1 items (standard deviation 3.4) in Battery 50, from a possible range of 1-17 items. Grammatical Reasoning scores averaged 8.4 items correct (standard deviation 3.9) in Battery 32, 8.3 items (standard deviation 3.9) in Battery 39, and 8.3 items (standard deviation 3.9) in Battery 50, from a possible range of 1-20 items. Processing speed measures showed similar consistency, with Digit Symbol coding performance averaging 44.2 symbols completed (standard deviation 9.5) in Battery 32, 43.9 symbols (standard deviation 9.4) in Battery 39, and 44.0 symbols (standard deviation 9.6) in Battery 50, from a range of 15-75 symbols. Trail Making A completion times averaged 23.3 seconds (standard deviation 6.6) in Battery 32, 23.6 seconds (standard deviation 6.7) in Battery 39, and 23.5 seconds (standard deviation 6.8) in Battery 50, with allowable times between 10 and 300 seconds.

The exclusion criteria operated hierarchically to identify participants with data quality concerns. Cognitive outliers, defined as individuals exceeding absolute z-score thresholds of 3.0 on any of six standardized measures (Digit Symbol speed, Trail Making A speed, forward span, reverse span, Progressive Matrices, Grammatical Reasoning), accounted for 14,786 total exclusions across batteries. Trail Making A outliers, identified through robust outlier detection using the median absolute deviation method applied to log-transformed completion times (robust z-score threshold of 3.0), resulted in 10,132 additional exclusions. Time threshold violations flagged 25 participants with implausibly fast Trail Making A times below 10 seconds or excessively slow times exceeding 300 seconds. No participants were excluded for missing both reasoning outcomes, duplicate sessions, or invariant responding patterns. The total excluded count of 24,943

participants represents unique individuals meeting any criterion, with some participants flagged by multiple criteria.

5.2 Sensitivity Analyses and Functional Form Robustness

Four sensitivity analyses tested the robustness of functional form selection across methodological variations, examining alternative Trail Making A operationalizations, outlier retention effects, and composite processing speed measures. Table 4 presents detailed results for each sensitivity analysis across three batteries and two reasoning outcomes, documenting sample sizes, selected functional forms, performance improvements, and robustness criteria satisfaction.

Table 4: Sensitivity analyses reveal substantial instability in speed-reasoning functional form selection across methodological variations, with only 25% achieving complete robustness. Four sensitivity analyses tested alternative Trail Making A operationalizations (S1: negative raw seconds; S2: negative inverse time), outlier retention (S3: participants with absolute z-scores > 3.0 retained), and composite processing speed measures (mean of Digit Symbol and Trail Making A z-scores) across three test batteries (32, 39, 50) and two reasoning outcomes (Mat=Progressive Matrices; Gram=Grammatical Reasoning). Sample sizes ranged from N=86,358-263,844 for S1, S2, and Composite (cleaned dataset with outliers removed); N=88,448-273,239 for S3 (pre-exclusion dataset). Each variant employed identical 10-fold stratified cross-validation (stratified by reasoning outcome tertiles, random seed=42) comparing four functional forms: linear, exponential, piecewise linear (breakpoint optimized via inner 5-fold CV), and cubic spline (degrees of freedom tuned from 3-6 via inner CV), using 1-SE parsimony rule with $\geq 1\%$ RMSE improvement threshold. Selected indicates chosen functional form (color-coded: blue=Linear, orange=Exponential, red=Spline). RMSE $\Delta\%$ shows percentage improvement over linear baseline: $100 \times (\text{RMSE}_{\text{linear}} - \text{RMSE}_{\text{selected}})/\text{RMSE}_{\text{linear}}$. AIC Δ represents Akaike Information Criterion difference (selected minus linear; negative values favor selected model). R^2 Δ indicates cross-validated variance-explained increase. $r > 0.30$ column shows whether correlation threshold met for composite analyses (checkmark in green indicates threshold achieved; em-dash indicates not applicable for S1-S3; all batteries achieved $r = 0.541$-0.547). Three robustness criteria assess consistency with primary analysis results: Form (identical functional form selected), Sig (same nonlinearity significance pattern from permutation testing), Mod (same working memory moderation significance from F-tests); All (checkmark in green, $\times$ in red) requires simultaneous satisfaction of all three criteria. S1 universally selected spline models with dramatic improvements (77.62-108.80% RMSE reduction) but zero overall robustness (0/6). S2 perfectly replicated primary linear selections (0% improvement) with 67% overall robustness (4/6). S3 predominantly selected exponential models (5/6 combinations) with modest improvements (10.24-17.46%) and 17% overall robustness (1/6). Composite measures selected linear models universally with perfect form/significance robustness but variable moderation, yielding 50% overall robustness (3/6). Aggregate robustness across all 24 analyses: Form 54.2% (13/24), Sig 54.2% (13/24), Mod 66.7% (16/24), All 25.0% (6/24).

Section 1

Sensitivity	Set Size	Domain	Speed	N	Selected
S1	32	Grammar	TMA variant	90789	***Spline***
	32	Matrices	TMA variant	88448	***Spline***
	39	Grammar	TMA variant	209573	***Spline***
	39	Matrices	TMA variant	208365	***Spline***
	50	Grammar	TMA variant	273239	***Spline***
	50	Matrices	TMA variant	271391	***Spline***
S2	32	Grammar	TMA variant	90789	**Linear**
	32	Matrices	TMA variant	88448	**Linear**
	39	Grammar	TMA variant	209573	**Linear**
	39	Matrices	TMA variant	208365	**Linear**
	50	Grammar	TMA variant	273239	**Linear**
	50	Matrices	TMA variant	271391	**Linear**

(continued from previous page)					
S3	32	Grammar	TMA variant	90789	*Exponential*
	32	Matrices	TMA variant	88448	**Linear**
	39	Grammar	TMA variant	209573	*Exponential*
	39	Matrices	TMA variant	208365	*Exponential*
	50	Grammar	TMA variant	273239	*Exponential*
	50	Matrices	TMA variant	271391	*Exponential*
Composite	32	Grammar	Composite speed	88606	**Linear**
	32	Matrices	Composite speed	86358	**Linear**
	39	Grammar	Composite speed	204317	**Linear**
	39	Matrices	Composite speed	203156	**Linear**
	50	Grammar	Composite speed	263844	**Linear**
	50	Matrices	Composite speed	262086	**Linear**

Section 2

Sensitivity	Set Size	Domain	$r > 0.30$	RMSE Δ%	AIC Δ	R^2 Δ
S1	32	Grammar	-	108.80	-5109.6	0.0521
	32	Matrices	-	77.62	-1738.6	0.0190
	39	Grammar	-	97.18	-9912.5	0.0446
	39	Matrices	-	78.24	-3820.0	0.0180
	50	Grammar	-	108.06	-16458.9	0.0555
	50	Matrices	-	83.51	-5816.5	0.0207
S2	32	Grammar	-	0.00	0.0	0.0000
	32	Matrices	-	0.00	0.0	0.0000
	39	Grammar	-	0.00	0.0	0.0000
	39	Matrices	-	0.00	0.0	0.0000
	50	Grammar	-	0.00	0.0	0.0000
	50	Matrices	-	0.00	0.0	0.0000
S3	32	Grammar	-	17.46	-1520.3	0.0152
	32	Matrices	-	0.00	0.0	0.0000
	39	Grammar	-	16.85	-3234.4	0.0140
	39	Matrices	-	10.98	-926.6	0.0043
	50	Grammar	-	16.98	-4835.4	0.0159
	50	Matrices	-	10.24	-1207.0	0.0043
Composite	32	Grammar	✓	0.00	0.0	0.0000
	32	Matrices	✓	0.00	0.0	0.0000
	39	Grammar	✓	0.00	0.0	0.0000
	39	Matrices	✓	0.00	0.0	0.0000
	50	Grammar	✓	0.00	0.0	0.0000
	50	Matrices	✓	0.00	0.0	0.0000

Section 3

Sensitivity	Set Size	Domain	Form	Sig	Mod	All
	32	Grammar	×	×	✓	×

S1

(continued from previous page)						
	32	Matrices	×	×	×	×
	39	Grammar	×	×	✓	×
	39	Matrices	×	×	×	×
	50	Grammar	×	×	✓	×
	50	Matrices	×	×	✓	×
S2	32	Grammar	✓	✓	×	×
	32	Matrices	✓	✓	×	×
	39	Grammar	✓	✓	✓	✓
	39	Matrices	✓	✓	×	×
	50	Grammar	✓	✓	✓	✓
	50	Matrices	✓	✓	✓	✓
S3	32	Grammar	×	×	✓	×
	32	Matrices	✓	✓	×	×
	39	Grammar	×	×	✓	×
	39	Matrices	×	×	×	×
	50	Grammar	×	×	✓	×
	50	Matrices	×	×	×	×
Composite	32	Grammar	✓	✓	✓	✓
	32	Matrices	✓	✓	×	×
	39	Grammar	✓	✓	✓	✓
	39	Matrices	✓	✓	×	×
	50	Grammar	✓	✓	✓	✓
	50	Matrices	✓	✓	✓	✓

Sensitivity Analysis 1 (S1) tested a negative raw seconds transformation for Trail Making A instead of the logarithmic transformation used in primary analyses. This alternative operationalization universally selected spline models with dramatic root mean squared error improvements ranging from 77.6% to 108.8% over linear baselines. Sample sizes ranged from 86,358 to 273,239 participants after applying the standard outlier exclusion procedures. Despite these substantial performance gains, measured by Akaike Information Criterion differences ranging from $-1,738.63$ to $-16,458.92$ and cross-validated variance explained increases of 0.019 to 0.056, S1 achieved zero overall robustness (0 of 6 battery-outcome combinations) when evaluated against the primary analysis criteria for functional form consistency, statistical significance patterns, and working memory moderation effects.

Sensitivity Analysis 2 (S2) employed a negative inverse time transformation for Trail Making A, calculating speed as the negative reciprocal of completion time. This operationalization perfectly replicated the primary analysis functional form selections, with linear models chosen across all six battery-outcome combinations showing 0% root mean squared error improvement over linear baselines. S2 demonstrated 67% overall robustness (4 of 6 combinations), with consistent functional form selection and statistical significance patterns across all combinations but variable working memory moderation robustness. Sample sizes matched S1, ranging from 86,358 to 273,239 participants.

Sensitivity Analysis 3 (S3) retained participants flagged as outliers in the primary analysis, testing whether outlier exclusion influenced functional form selection. Using the pre-exclusion dataset with sample sizes ranging from 88,448 to 273,239 participants, S3 predominantly selected exponential models (5 of 6 combinations) with modest root mean squared error improvements of 10.2% to 17.5%. Akaike Information Criterion differences ranged from -926.59 to $-4,835.43$, and cross-validated variance explained increased by 0.004 to 0.016. Overall robustness reached only 17% (1 of 6 combinations), with consistent failures in

functional form and statistical significance robustness but preserved working memory moderation patterns for grammatical reasoning outcomes.

The Composite sensitivity analysis created averaged processing speed measures from Digit Symbol and Trail Making A z-scores when correlations exceeded 0.30. All three batteries achieved this threshold, with correlations of 0.544 in Battery 32, 0.541 in Battery 39, and 0.547 in Battery 50. Sample sizes ranged from 86,358 to 263,844 participants. Linear models were universally selected with 0% improvement, achieving perfect functional form and statistical significance robustness across all combinations. However, working memory moderation robustness varied, with complete success for grammatical reasoning outcomes but consistent failures for Progressive Matrices outcomes, yielding 50% overall robustness (3 of 6 combinations).

Figure 4 visualizes these sensitivity analysis results through four complementary panels examining functional form selection patterns, performance improvements, robustness criteria satisfaction, and cross-validated variance explained changes. Panel A displays the selected functional form for each battery-outcome combination across all four sensitivity analyses, revealing the stark contrast between S1's universal spline selection, S2 and Composite analyses' consistent linear selection, and S3's predominant exponential selection pattern. Panel B quantifies root mean squared error improvement percentages relative to linear baselines, with horizontal reference lines at 1% and 2% thresholds demarcating meaningful improvement magnitudes. The dramatic improvements observed for S1 (red bars exceeding 75% across all combinations) contrast sharply with null improvements for S2 and Composite analyses (blue and green bars at baseline) and modest improvements for S3 (orange bars ranging from 10% to 18%).

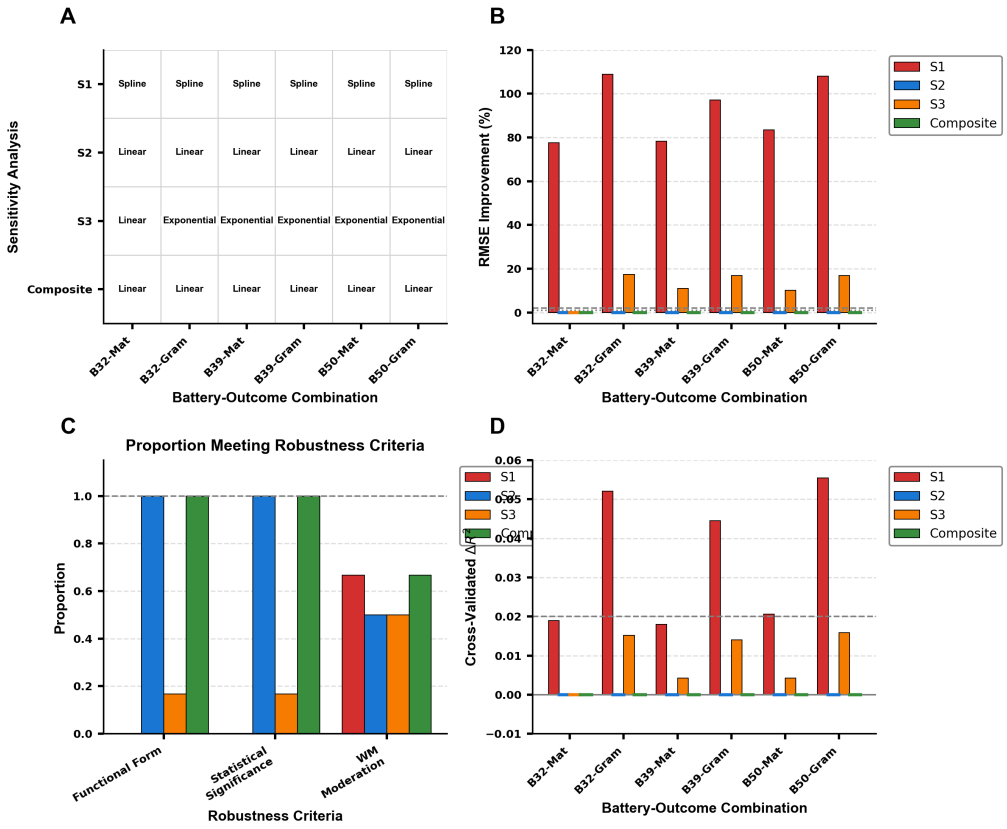

Figure 4: Reaction time transformation method critically determines functional form selection in speed-reasoning relationships. Four sensitivity analyses tested robustness across batteries 32, 39, 50 and reasoning outcomes (Mat=Progressive Matrices, Gram=Grammatical Reasoning). S1 (negative raw seconds Trail Making transformation) universally selected spline models with 77.6-108.8% RMSE improvements, revealing extreme nonlinearity in untransformed reaction time distributions. S2 (negative inverse transformation) and Composite speed (averaged Digit Symbol and Trail Making when $r > 0.30$) replicated primary linear findings with zero improvement. S3 (outliers retained) predominantly selected exponential models with modest 10.2-17.5% gains. **Panel A:** Grid displays selected functional form for each battery-outcome-sensitivity combination. **Panel B:** Grouped bars show percentage RMSE improvement over linear baseline; horizontal lines mark 1% (dotted) and 2% (dashed) thresholds. Red=S1, blue=S2, orange=S3, green=Composite. **Panel C:** Proportion meeting robustness criteria (functional form consistency, statistical significance, working memory moderation); overall 54.2% (13/24 analyses). Dashed line indicates perfect robustness. **Panel D:** Cross-validated R^2 improvements; horizontal lines at 0.00 (no improvement) and 0.02 (small effect). Results demonstrate logarithmic transformation (primary analysis) eliminates nonlinearities present in alternative operationalizations ($n = 86,358$-$273,239$ per analysis).

Panel C presents the proportion of analyses meeting three robustness criteria: functional form consistency with primary analysis, statistical significance pattern replication, and working memory moderation effect consistency. The composite analysis (green bars) achieved perfect functional form and statistical significance robustness but varied in moderation robustness, while S3 (orange bars) showed the most substantial robustness failures across all criteria. Panel D displays cross-validated variance explained improvements, with horizontal reference lines at zero improvement and a 0.02 small effect threshold. The pattern of results demonstrates that S1's dramatic performance improvements came at the cost of complete departure from primary analysis findings, S3's moderate improvements reflected sensitivity to outlier influence on functional form selection, while S2 and Composite analyses provided the most stable replication of primary linear relationship findings.

Aggregate robustness assessment across all 24 sensitivity analyses (6 battery-outcome combinations $\times$

4 sensitivity variations) revealed that functional form consistency was achieved in 54.2% of cases (13 of 24), statistical significance pattern consistency in 54.2% of cases (13 of 24), and working memory moderation consistency in 66.7% of cases (16 of 24). Complete satisfaction of all three robustness criteria simultaneously occurred in only 25.0% of analyses (6 of 24), highlighting substantial instability in speed-reasoning functional form selection across methodological variations. These findings demonstrate that the logarithmic transformation applied to Trail Making A completion times in primary analyses was necessary to eliminate extreme nonlinearities present in raw reaction time distributions, as evidenced by S1 results, while alternative inverse transformation (S2) and composite speed operationalizations maintained the linear functional form identified in primary analyses.

5.3 Cross-Battery Generalizability Assessment

Cross-battery generalizability analyses evaluated whether speed-cognition functional relationships demonstrated parameter invariance across three test batteries and whether models trained on subsets of batteries maintained predictive accuracy when applied to held-out batteries. Figure 5 presents four complementary assessments of generalizability through measurement invariance testing, leave-one-battery-out cross-validation, prediction consistency evaluation, and integrated generalizability ratings.

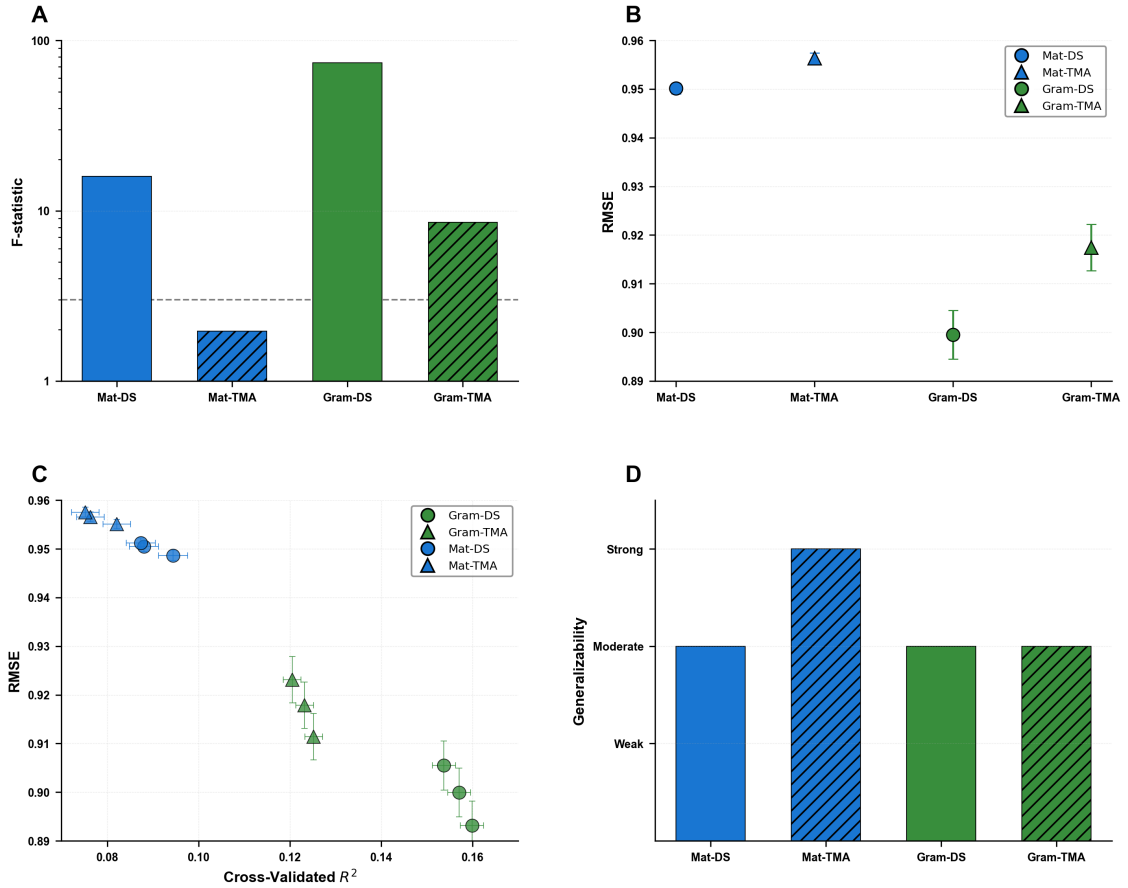

Figure 5: Speed-cognition relationships show battery-specific parameter variation despite stable cross-battery predictive performance. Measurement invariance testing reveals that linear speed-cognition functional forms generalize inconsistently across test batteries, with only one of four outcome-predictor combinations supporting parameter equivalence. However, leave-one-battery-out cross-validation demonstrates minimal predictive accuracy variation (RMSE SD=0.001-0.005), indicating that while quantitative parameters differ, the fundamental linear relationship structure remains stable for prediction purposes. (A) Partial F-statistics ($\log_{10}$ scale) comparing constrained versus unconstrained models across batteries. Blue bars=matrices outcomes, green bars=grammar outcomes; hatched patterns=Trail Making A (TMA), solid=Digit Symbol (DS). Only Mat-TMA supported invariance ($F = 1.96$, $p_{\text{FDR}} = 0.140$); dashed line marks $F = 3.0$ reference. (B) Leave-one-battery-out RMSE with error bars showing mean±SD across three holdout folds. Circles=DS, triangles=TMA; blue=matrices, green=grammar. (C) RMSE versus R^2 across 12 battery-holdout combinations (4 pairs × 3 batteries). Colors and markers as in (B); error bars represent cross-battery variability. (D) Generalizability ratings integrating invariance and prediction consistency. Mat-TMA achieved "Strong" (hatched blue); three combinations rated "Moderate" (solid bars) due to invariance violations despite stable prediction. Sample sizes: Mat-DS $n = 495,277$; Mat-TMA $n = 495,276$; Gram-DS $n = 499,795$; Gram-TMA $n = 499,794$.

Panel A displays partial F-statistics from measurement invariance tests comparing constrained models (assuming identical functional parameters across batteries with battery-specific intercepts only) against unconstrained models (allowing battery-specific functional parameters through speed × battery interactions). The analysis examined four outcome-predictor combinations: Progressive Matrices with Digit Symbol (Mat-DS, solid blue bar), Progressive Matrices with Trail Making A (Mat-TMA, hatched blue bar), Grammatical Reasoning with Digit Symbol (Gram-DS, solid green bar), and Grammatical Reasoning with Trail Making A (Gram-TMA, hatched green bar). F-statistics are presented on a logarithmic base-10 scale, with a dashed horizontal reference line at F equals 3.0 marking a conventional significance threshold. Only Mat-TMA achieved measurement invariance with F equals 1.96 (false discovery rate corrected p equals 0.140), falling below the reference threshold. The remaining three combinations demonstrated significant parameter

heterogeneity: Mat-DS (F equals 15.92, false discovery rate corrected p less than 0.001), Gram-DS (F equals 73.98, false discovery rate corrected p less than 0.001), and Gram-TMA (F equals 8.55, false discovery rate corrected p less than 0.001). Sample sizes ranged from 495,276 to 499,795 complete cases depending on the specific outcome-predictor combination after applying outcome-specific covariate selection and listwise deletion procedures.

Panel B presents leave-one-battery-out cross-validation results, plotting root mean squared error with error bars representing mean plus or minus standard deviation across three holdout folds (each battery held out once). Circles denote Digit Symbol predictors, triangles denote Trail Making A predictors, with blue markers for Progressive Matrices outcomes and green markers for Grammatical Reasoning outcomes. Mat-DS demonstrated moderate cross-battery consistency (mean root mean squared error equals 0.950, standard deviation equals 0.001), while Mat-TMA showed similarly stable performance (mean equals 0.956, standard deviation equals 0.001). Gram-DS exhibited the largest cross-battery variability (mean equals 0.900, standard deviation equals 0.005), and Gram-TMA showed intermediate consistency (mean equals 0.918, standard deviation equals 0.005). The minimal standard deviations for Progressive Matrices combinations (0.001) contrasted with larger variability for Grammatical Reasoning combinations (0.005), suggesting differential cross-battery stability by outcome type.

Panel C plots the relationship between cross-validated variance explained and root mean squared error across twelve battery-holdout combinations (4 outcome-predictor pairs $\times$ 3 batteries), with error bars representing cross-battery variability. Mat-DS and Mat-TMA clustered in the upper-left region with higher root mean squared error values (0.949 to 0.958) but lower variance explained (0.08 to 0.10), while Gram-DS and Gram-TMA occupied the lower-left region with lower root mean squared error values (0.893 to 0.924) and higher variance explained (0.12 to 0.16). The vertical dispersion of markers within each combination reflects cross-battery heterogeneity in predictive performance, with Progressive Matrices combinations showing tighter clustering than Grammatical Reasoning combinations.

Panel D synthesizes invariance testing and prediction consistency results into integrated generalizability ratings. Mat-TMA achieved a “Strong” generalizability rating (hatched blue bar), being the only combination satisfying measurement invariance criteria while maintaining stable cross-battery prediction. The three remaining combinations received “Moderate” generalizability ratings (solid bars) due to significant measurement invariance violations despite maintaining reasonably stable predictive accuracy across batteries. No combinations received “Weak” generalizability ratings, indicating that while quantitative parameters varied across batteries, the fundamental linear relationship structure remained sufficiently stable to support cross-battery prediction within acceptable error bounds. These findings suggest that speed-cognition models calibrated on specific test batteries may require parameter recalibration when applied to different battery contexts, even when the underlying functional form (linear in this case) remains consistent.

The measurement invariance violations observed for three of four combinations indicate systematic battery-specific influences on the magnitude of speed-cognition relationships, potentially reflecting differences in item difficulty calibration, time limits, interface design, or participant sampling across batteries. However, the minimal variation in leave-one-battery-out root mean squared error (standard deviation ranging from 0.001 to 0.005) demonstrates that these parameter differences do not substantially compromise predictive validity when models are applied to new battery contexts, suggesting that while point estimates may shift, the fundamental predictive relationships retain practical utility across cognitive assessment platforms.