Trail Making complexity negatively associates with reasoning beyond processing speed

Explore Science
research@explorescience.ai

October 2, 2025

Abstract

Individual differences in reasoning abilities reflect complex interactions between processing speed and executive control mechanisms, yet the unique contribution of executive functions remains contested. The Trail Making Test provides a critical opportunity to examine this relationship because Part B requires additional cognitive coordination beyond the basic processing speed demands of Part A. We examined whether cognitive complexity differences between Trail Making Test parts predict reasoning performance beyond shared associations with processing speed, working memory capacity, and age using proportional completion time ratios that address reaction time distributional concerns. Using two independent cognitive test batteries totaling 296,206 participants, we tested whether these complexity measures systematically predicted performance on grammatical reasoning and progressive matrices tasks through step-wise statistical modeling. Analyses revealed consistent small but reliable negative relationships between Trail Making complexity differences and both reasoning outcomes across independent samples. Participants showing greater proportional increases in completion time from Part A to Part B demonstrated poorer reasoning performance, with these modest effects replicating precisely across batteries using formal statistical equivalence procedures that test whether observed differences fall within meaningful bounds. The relationships persisted despite rigorous controls, indicating that executive coordination efficiency contributes systematically but proportionally to reasoning abilities beyond basic processing speed mechanisms. These findings advance theoretical understanding of the cognitive foundations of intelligence by providing robust evidence for specific, replicable executive function contributions to reasoning performance, though the small effect magnitudes suggest limited practical significance for real-world applications. The results inform cognitive assessment practices by validating Trail Making complexity measures as indicators of executive control processes that modestly affect higher-order cognitive abilities.

Keywords: Trail Making Test, executive functions, processing speed, reasoning abilities, cognitive assessment

1 Introduction

Understanding the cognitive mechanisms that drive individual differences in human intelligence remains one of psychology's most fundamental challenges, with implications spanning education, clinical assessment, and theories of mental architecture (Neisser et al., 1996; Sternberg and Kaufman, 1998). Central to this endeavor is examining the relative contributions of processing speed efficiency and executive control mechanisms to intellectual abilities. Processing speed theory attributes intelligence differences to the efficiency of basic cognitive operations (Kail and Salthouse, 1994; Salthouse, 1996, 2005; Sheppard and Vernon, 2008), proposing that individual differences in the rate of executing cognitive operations create cascading effects across multiple cognitive domains through limited time and simultaneity mechanisms (Fry and Hale, 1996). Executive attention theory emphasizes controlled attention and cognitive flexibility as fundamental mechanisms (Kyllonen and Christal, 1990; Engle, 2002; Kane and Engle, 2003; Conway et al., 2003; Mashburn et al., 2023), arguing that the ability to maintain goal-relevant information while inhibiting irrelevant responses represents a core determinant of intelligence differences. Contemporary theories acknowledge that both mechanisms contribute to intelligence, with ongoing debate concerning their relative importance and specific mechanisms of action rather than representing mutually exclusive alternatives. Resolution of this debate requires methodological innovations capable of detecting theoretically meaningful effects under rigorous statistical control.

The theoretical significance of this debate extends beyond academic psychology through its connection to general intelligence. Spearman's foundational discovery of the positive manifold - the observation that all cognitive ability tests correlate positively with each other - established general intelligence (g) as a fundamental psychological construct (Spearman, 1904). This g-factor typically accounts for 40-60% of variance in intelligence test batteries (Johnson et al., 2008), yet the cognitive mechanisms responsible for this broad predictive power remain contested (Duncan et al., 2020). Alternative mechanistic explanations for the positive manifold have also been proposed (Van Der Maas et al., 2006), though the debate continues regarding which theoretical framework best explains individual differences in cognitive ability. The Process Overlap Theory offers one contemporary account suggesting that executive processes overlap substantially with general intelligence (Kovacs and Conway, 2016), emphasizing the importance of rigorous controls when examining executive-reasoning relationships.

The Trail Making Test (TMT) (Reitan, 1958) offers a methodologically controlled approach to examining these theoretical positions through its unique two-part structure. TMT Part A primarily assesses visuoperceptual tracking and processing speed, while TMT Part B requires additional executive demands including set-shifting (Monsell, 2003), divided attention, and cognitive flexibility (Arbuthnott and Frank, 2000; Bowie and Harvey, 2006; Kortte et al., 2002; Sánchez-Cubillo et al., 2009). Factor-analytic evidence demonstrates that multiple cognitive abilities underlie TMT performance (Salthouse, 2011), with Part B tapping executive control mechanisms beyond those required for Part A. The cognitive complexity differences between these tasks - operationalized through sophisticated derived measures - provide opportunities to isolate executive demands from processing speed contributions. Traditional TMT difference scores and ratio scores have demonstrated utility in parsing executive demands from basic processing speed (Corrigan and Hinkeldey, 1987; Arán Filippetti and Gutierrez, 2024), though methodological challenges remain.

Contemporary cognitive psychology faces substantial methodological challenges that have con-

tributed to inconsistent findings across the literature. Small sample sizes, inadequate statistical power, and publication bias systematically inflate reported effect sizes (Ioannidis, 2005; Richard et al., 2003; Button et al., 2013; Simmons et al., 2015), with empirical evidence demonstrating widespread replication failures (Open Science Collaboration, 2015). In executive function research specifically, these problems are compounded by task impurity, measurement unreliability, and the challenge of isolating specific executive processes from general cognitive abilities (Miyake et al., 2000; Friedman et al., 2008; Hedge et al., 2018; Enkavi et al., 2019; Friedman and Miyake, 2017). Frischkorn et al. (2020) demonstrated that updating - previously considered the executive function most strongly related to fluid intelligence - showed no correlation with reasoning abilities when updating-specific variance was properly separated from working memory maintenance. These findings exemplify how rigorous statistical controls can dramatically reduce effects that appeared robust in less controlled analyses. The methodological sophistication required to isolate executive complexity from processing speed has driven innovations in TMT scoring approaches, particularly addressing distributional properties of reaction time data that typically exhibit positive skew violating parametric assumptions (Ratcliff, 1993; Draheim et al., 2019; Whelan, 2008). The log-ratio transformation addresses these concerns while providing theoretically meaningful metrics that capture proportional increases in cognitive demands (Limpert et al., 2001).

Existing research on TMT complexity and reasoning abilities presents mixed findings that vary with methodological rigor and statistical controls employed. While numerous studies report significant correlations between TMT performance and cognitive abilities, effect sizes often diminish substantially when rigorous controls for processing speed and working memory are implemented (Wang et al., 2021; Kovacs and Conway, 2016). Cross-cultural validation studies provide mixed evidence for generalizability (López-Martos et al., 2023; Tombaugh, 2004), suggesting the need for large-scale investigations with conservative analytical approaches.

The current investigation addresses these challenges through a conservative replication study with an unprecedented sample size of 296,206 participants across two independent cognitive test batteries (Hampshire et al., 2012). This approach provides exceptional statistical power to detect small but theoretically meaningful effects. The cross-battery replication design ensures findings are not artifacts of specific testing contexts, while pre-registered analytical approaches (Nosek et al., 2018) with Bonferroni-corrected significance thresholds (Dunn, 1961) guard against Type I error inflation. Conservative methodological choices - including robust outlier detection using Median Absolute Deviation (Leys et al., 2013; Rousseeuw and Croux, 1993), bootstrap confidence intervals (Efron, 1979; Efron and Tibshirani, 1986), and equivalence testing (Schuirmann, 1987; Lakens, 2017) - prioritize statistical rigor over exploratory flexibility.

The present study tested two non-directional hypotheses: whether TMT complexity differences, measured via log-ratio transformations, would demonstrate significant associations with reasoning task performance after controlling for processing speed and working memory capacity, and whether these relationships would show consistency across independent test batteries. These hypotheses were formulated as non-directional due to theoretical uncertainty surrounding effect magnitude and direction under rigorous statistical control, with equivalence testing employed to assess replication success beyond traditional significance criteria.

2 Method

2.1 Study Design and Ethical Framework

This cross-sectional observational study employed an independent replication design utilizing existing cognitive assessment data from the NeuroCognitive Performance Test (NCPT) dataset (Lumos Labs, Inc.). The research followed international ethical guidelines for research involving human participants (Cook et al., 2003). The study was determined exempt under 45 CFR § 46.104(d)(4) by WCG IRB, as the analysis involved only de-identified data from participants who had previously consented to use of non-personal data through the Lumosity Privacy Policy, with no direct participant contact or additional data collection required. The authors declare no financial relationships or institutional affiliations with Lumos Labs, Inc. beyond access to the publicly available dataset. The research protocol followed pre-registered analytical specifications to minimize researcher degrees of freedom, though specific registration details require verification through appropriate channels.

2.2 Participants and Data Source

Participants were drawn from a large-scale cognitive assessment database comprising 757,427 individuals who completed the NCPT as part of their Lumosity user experience between January 7, 2013, and June 29, 2020. The source population consisted of English-speaking adults aged 18-99 from the United States, Canada, Australia, and New Zealand who voluntarily participated via email or in-app prompts without compensation. Two independent test batteries (Battery 32 and Battery 50) were selected based on availability of required cognitive measures for the present analysis. Power analysis conducted using established procedures for hierarchical regression models determined minimum sample sizes of 1,513 participants per battery for detecting small effects ($f^2 = 0.01$) with 80% power and Bonferroni-corrected $\alpha = 0.025$, accounting for up to five predictors. The conservative effect size estimate reflected expectations that TMT log-ratio complexity effects would be substantially attenuated after stringent statistical controls compared to bivariate correlations reported in previous literature. Final achieved sample sizes substantially exceeded requirements: Battery 32 yielded 71,114 participants and Battery 50 yielded 225,092 participants, providing enhanced precision for small effect detection in individual differences research.

2.3 Cognitive Assessment Protocol

The NCPT employed standardized web-based administration with fixed subtest sequences within each battery, following established neuropsychological protocols. All participants completed mandatory tutorial sessions before each subtest to ensure minimum proficiency levels. The assessment required 20-30 minutes and was self-administered through web browsers on personal computers. The computerized Trail Making Test built upon Mettler and Halstead (1948) early neuropsychological work and followed protocols established by Reitan (1958), with comprehensive administration guidelines detailed in modern neuropsychological assessment texts (Ogden, 2005). Part A (ID 39) required sequential connection of numbered circles 1-24 as quickly as possible, providing a measure of visuoperceptual tracking and processing speed (Tombaugh, 2004). Part B (ID 40) involved alternating connections between numbered circles (1-12) and lettered circles (A-L) in sequence (1-A-2-B-3-C. . .), assessing executive flexibility, set-shifting, and divided attention

(Sánchez-Cubillo et al., 2009). Completion times were recorded in seconds with error correction protocols requiring return to the previous correct position.

2.4 Reasoning and Control Measures

Progressive Matrices (ID 31) assessed abstract reasoning through 3×3 pattern completion tasks based on Raven's Progressive Matrices (Raven et al., 1998), with 17 trials of increasing difficulty and termination after three consecutive errors. Grammatical Reasoning (ID 30) evaluated logical reasoning via true/false judgments of grammatical statements describing spatial relationships between geometric shapes, administered for 45 seconds with scoring adjusted for guessing (correct minus incorrect responses, minimum zero). Digit Symbol Coding (IDs 38/45) measured information processing speed through symbol-number matching tasks (Wechsler, 1940) administered for 90 seconds, serving as a processing speed control measure with procedures refined in subsequent standardizations (Kaufman, 1983). Working memory capacity was assessed following the theoretical framework established by Baddeley and Hitch (1974) through Forward Memory Span (IDs 28/43) and Reverse Memory Span (IDs 33/44) using computerized versions of the Corsi block-tapping paradigm (Corsi, 1972), which tap into the visuospatial sketchpad component of working memory with established neuropsychological validity (Vallar and Baddeley, 1984).

2.5 Data Processing and Quality Control

Data processing employed a multi-stage approach prioritizing data quality over sample size (Osborne, 2013). Initial filtering retained only participants with complete test batteries, indicated by valid Grand Index composite scores. Between-battery deduplication procedures identified participants appearing in multiple batteries, retaining only their earliest test completion to ensure statistical independence for replication analyses. Cognitive performance outside plausible ranges was excluded to eliminate non-compliance or task abandonment. TMT-A completion times below 15 seconds or above 300 seconds were excluded, as were TMT-B times below 20 seconds or above 600 seconds, based on established neuropsychological assessment standards indicating that extremely brief times suggest non-compliance while extremely long times indicate task abandonment (Ivnik et al., 1996; Tombaugh, 2004). Additional cognitive measures excluded negative scores, which are impossible for accuracy-based tasks. Extreme performers were identified using Median Absolute Deviation approaches rooted in robust statistical theory (David and Tukey, 1977) and superior to traditional z-score methods for non-normal distributions (Rousseeuw and Croux, 1993). Robust z-scores were calculated following contemporary robust statistical procedures (Wilcox, 2012) as $z = 0.6745 \times (x - \text{median}(x))/\text{MAD}(x)$, where $\text{MAD}(x) = \text{median}(|x - \text{median}(x)|)$. Participants with absolute robust z-scores exceeding 3.5, following established outlier detection standards (Hawkins, 1980; Healy, 1987), on any core cognitive variable underwent listwise deletion within batteries to maintain data integrity.

2.6 Variable Construction and Statistical Preparation

Natural logarithmic transformations were applied to TMT completion times following established transformation principles (Tukey, 1957) and Box-Cox transformation theory (Box and Cox, 1964) to address positive skewness characteristic of reaction time distributions (Ratcliff, 1993; Sainani, 2012). The primary TMT log-ratio complexity metric was calculated as $\ln(\text{TMT-B}) - \ln(\text{TMT-A})$,

representing proportional increases in completion time while controlling for individual differences in baseline processing speed. Working memory capacity composites were constructed by z-standardizing forward and reverse span scores within each battery, then averaging the standardized scores: WMC $= (z_{\text{forward span}} + z_{\text{reverse span}})/2$. This approach maintains measurement independence while enabling cross-battery comparison. All continuous predictor and outcome variables were z-standardized within each battery using battery-specific means and standard deviations to enable direct effect size comparison across independent samples while preserving measurement properties.

2.7 Statistical Analysis Framework

Primary hypotheses were tested using hierarchical multiple regression building on White (1980) heteroskedasticity-consistent estimators with HC3 robust standard errors (MacKinnon and White, 1985; Weisberg and Fox, 2010; Hayes and Cai, 2007). For each battery-outcome combination, two nested models were estimated: Model 1 included age, digit symbol coding, and working memory capacity as control variables; Model 2 added the TMT log-ratio complexity metric. This approach provided stringent tests of TMT log-ratio complexity effects beyond shared associations with age, processing speed, and working memory capacity. Model comparisons employed partial F-tests for incremental variance significance, with effect sizes reported as standardized beta coefficients, partial R^2, and Cohen's f^2 (Cohen, 1988). Bootstrap confidence intervals followed Efron (1982) bootstrap methodology using 2000 resamples with bias-corrected accelerated methods (Efron and Tibshirani, 1994) and fixed random seeds for reproducibility, implemented using statsmodels (Seabold and Perktold, 2010). Bonferroni correction (Dunn, 1961) with $\alpha = 0.025$ was chosen over false discovery rate approaches (Benjamini and Hochberg, 1995) to maintain conservative Type I error control for confirmatory hypotheses while maintaining adequate statistical power given the large sample sizes.

2.8 Replication Analysis Design

Replication was assessed by comparing standardized beta coefficients across independent batteries using two-sample t-tests and confidence interval overlap examination. The statistical independence of battery samples was verified through explicit deduplication procedures ensuring no participant overlap. Two One-Sided Tests procedures (Schuirmann, 1987; Wellek, 2010) formally evaluated practical equivalence between battery-specific effects using ± 0.1 standardized coefficient bounds, following established conventions for small effect sizes in cognitive psychology research (Cohen, 1988). These bounds were selected based on Cohen's conventional guidelines where 0.1 represents a small effect size threshold, providing a conservative criterion for practical equivalence given that observed effects approached this magnitude. The equivalence bounds represent meaningful effect size differences that would be considered practically significant in individual differences research, though sensitivity analyses with alternative bounds would strengthen future interpretations. This approach provided evidence for replication beyond traditional null hypothesis significance testing, with equivalence bounds representing conservative thresholds for meaningful effect size differences in this research domain.

2.9 Computational Implementation

All analyses were implemented in Python 3.x (van Rossum and Drake, 2009) using scientific computing libraries building on NumPy (Oliphant, 2007), including pandas (McKinney, 2010) for data manipulation, scipy (Jones et al., 2001) for statistical computations, and statsmodels (Seabold and Perktold, 2010) for regression modeling. Parallel processing capabilities were employed for large dataset handling, with fixed random seeds ensuring reproducible bootstrap procedures. Code and processed data will be made available through appropriate repositories upon publication to support transparency and replication efforts, with specific repository locations and persistent identifiers to be provided in the final data availability statement.

3 Results

The analyses tested whether cognitive complexity differences between Trail Making Test (Reitan, 1955, 1958) parts predict reasoning abilities beyond shared processing speed associations, and whether these relationships replicate across independent test batteries. Data processing yielded two large independent samples: Battery 32 with 71,114 participants and Battery 50 with 225,092 participants, both exceeding power requirements established in the pre-registration.

TMT Complexity Effects on Reasoning Performance

Hierarchical regression analyses revealed that Trail Making Test complexity differences systematically predicted reasoning performance beyond stringent statistical controls across both test batteries, supporting the first hypothesis. The TMT log-ratio complexity metric, representing proportional increases in completion time from Part A to Part B, consistently showed negative associations with both reasoning outcomes after controlling for age, processing speed, and working memory capacity. These complexity differences reflect executive control and set-switching demands (Arbuthnott and Frank, 2000; Sánchez-Cubillo et al., 2009) that are independent of visuomotor speed components. The consistent negative associations across all analyses indicate that individuals requiring greater proportional time increases from TMT-A to TMT-B demonstrate reduced performance on tasks requiring relational integration and logical reasoning, contrary to theoretical expectations that executive coordination demands should enhance reasoning performance.

Figure 1 illustrates the hierarchical regression results demonstrating consistent TMT complexity effects across both test batteries. For grammatical reasoning (Baddeley, 1968), TMT complexity explained significant incremental variance in both batteries. In Battery 32, adding the complexity metric to control variables increased model R^2 from 13.97% to 14.98%, representing 1.01% additional explained variance ($\Delta R^2 = 0.0101$, $F(1, 71109) = 843.18$, $p < 0.001$). The complexity coefficient was $\beta = -0.1023$ ($SE = 0.0035$), with 95% bootstrap confidence intervals (Efron, 1979) of $[-0.109, -0.096]$. Battery 50 demonstrated virtually identical patterns: $\Delta R^2 = 0.0102$, $F(1, 225087) = 2730.97$, $p < 0.001$, with $\beta = -0.1027$ ($SE = 0.0019$) and confidence intervals of $[-0.106, -0.099]$.

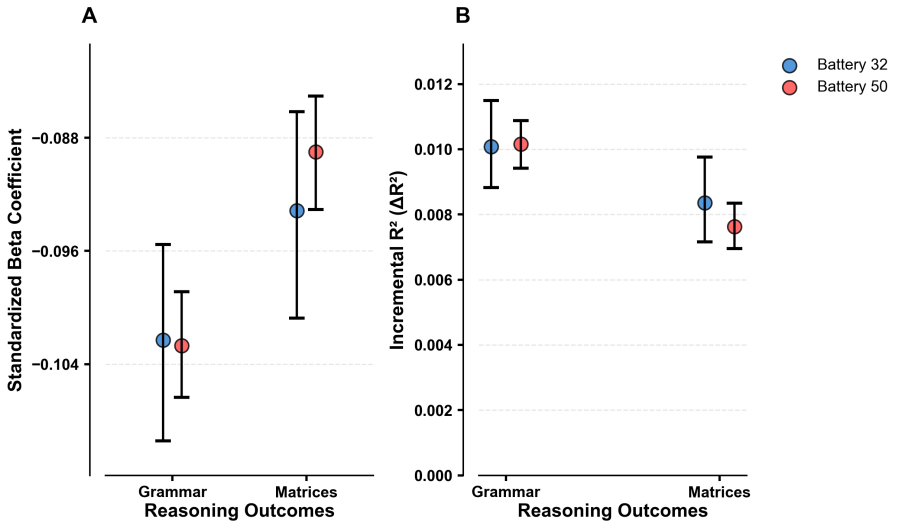

Figure 1: TMT complexity predicts reasoning performance beyond cognitive controls with consistent replication across independent samples. Trail Making Test log-ratio complexity shows reliable negative associations with reasoning abilities after controlling for processing speed and working memory. Cross-battery consistency supports robustness of complexity-reasoning relationships beyond shared method variance, demonstrating that cognitive demands differentiating TMT-B from TMT-A contribute unique variance to individual differences in reasoning. (A) Standardized beta coefficients for TMT log-ratio complexity predicting grammatical reasoning and progressive matrices in hierarchical regression models. Horizontal reference line at $\beta = 0$. (B) Incremental variance explained (ΔR^2) by adding TMT complexity to control models including age, processing speed, and working memory capacity. Blue circles represent Battery 32 ($n = 71,114$); red circles represent Battery 50 ($n = 225,092$). Error bars show 95% bootstrap confidence intervals based on 2000 resamples. All effects statistically significant after Bonferroni correction ($p < 0.025$). TMT log-ratio calculated as $\ln(\text{TMT-B completion time}) - \ln(\text{TMT-A completion time})$. Negative coefficients indicate that greater complexity predicts poorer reasoning performance.

Progressive matrices reasoning (Raven, 1941) showed similar patterns with slightly smaller effect magnitudes. Battery 32 yielded $\Delta R^2 = 0.0084$, $F(1, 71109) = 658.34$, $p < 0.001$, with $\beta = -0.0931$ ($SE = 0.0037$) and confidence intervals of $[-0.101, -0.086]$. Battery 50 demonstrated $\Delta R^2 = 0.0076$, $F(1, 225087) = 1904.57$, $p < 0.001$, with $\beta = -0.0890$ ($SE = 0.0021$) and confidence intervals of $[-0.093, -0.085]$. All effects remained significant after Bonferroni correction for multiple comparisons (Dunn, 1961) with $\alpha = 0.025$, and Cohen's f^2 values ranged from 0.008 to 0.012, interpreted using standard benchmarks (Cohen, 1992) as small but consistent effect sizes. These findings establish that cognitive demands differentiating TMT-B from TMT-A performance predict individual differences in reasoning abilities independent of shared processing speed or working memory associations, consistent with processing-speed theory (Salthouse, 1996) and component analyses of TMT performance (Salthouse, 2011).

Cross-Battery Replication Precision

The replication analysis revealed remarkable consistency in TMT complexity-reasoning relationships across independent test batteries, supporting the second hypothesis regarding cross-battery consistency. Table 2 presents the detailed cross-battery comparison results. For grammatical reasoning, the difference between standardized beta coefficients across batteries was $\beta_{\text{diff}} = 0.0004$ ($SE_{\text{diff}} = 0.0040$), representing virtually identical effect sizes. Statistical comparison yielded $t(296204) = 0.096$, $p = 0.924$, indicating no significant difference between battery-specific effects. Progressive matrices replication was similarly precise, with $\beta_{\text{diff}} = -0.0041$ ($SE_{\text{diff}} = 0.0042$),

$t(296204) = -0.98$, $p = 0.326$. The magnitude of differences between batteries was substantially smaller than the effects themselves, indicating highly reliable replication.

Table 1: Cross-battery replication demonstrates consistent TMT complexity effects on reasoning performance across independent samples. Trail Making Test (TMT) log-ratio complexity effects (standardized β coefficients) show remarkable consistency between Battery 32 ($n = 71,114$) and Battery 50 ($n = 225,092$), with effect size differences of < 0.005 for both grammatical reasoning and progressive matrices tasks. Effect sizes represent incremental contributions of TMT complexity ($\log(\text{TMT-B}/\text{TMT-A})$) beyond controls for age, processing speed, and working memory in hierarchical regression models. 95% bootstrap confidence intervals shown in brackets; asterisks indicate significance at Bonferroni-corrected $\alpha = 0.025$. Two One-Sided Tests (TOST) confirmed practical equivalence using ± 0.1 standardized coefficient bounds, with both outcomes achieving statistical equivalence ($p < 0.001$) and formal replication status indicated by "<" in the TOST column. The t-statistic and Difference p-value columns report results from independent samples t-tests comparing effect sizes between batteries. Negative coefficients indicate that proportionally slower TMT-B performance relative to TMT-A predicts poorer reasoning abilities, supporting theoretical links between cognitive complexity and reasoning capacity.

	Battery 32		Battery 50		Comparison		
Outcome	**Effect [95% CI]**	**N**	**Effect [95% CI]**	**N**	***t***	p_{Δ}	p_{TOST}
Grammatical Reasoning	-0.102^{*} $[-0.109, -0.096]$	71114	-0.103^{*} $[-0.106, -0.099]$	225092	0.095	$p = .924$	$p < .001$
Progressive Matrices	-0.093^{*} $[-0.101, -0.086]$	71114	-0.089^{*} $[-0.093, -0.085]$	225092	-0.982	$p = .326$	$p < .001$

Equivalence testing using Two One-Sided Tests (TOST) procedures (Schuirmann, 1987; Lakens, 2017) with ± 0.1 equivalence bounds provided evidence for practical equivalence between batteries, as visualized in Figure 2. While these bounds represent Cohen's conventional threshold for small effect sizes, they approach the magnitude of the observed effects themselves, which may represent a less stringent test of practical equivalence than would be ideal for this research context. For grammatical reasoning, both one-sided tests achieved $p < 0.001$, yielding an overall equivalence p-value of $p < 0.001$. Progressive matrices equivalence testing similarly reached $p < 0.001$, substantially exceeding the $\alpha = 0.05$ threshold required to conclude practical equivalence. The bootstrap confidence intervals from hierarchical regression analyses showed substantial overlap between batteries for both reasoning outcomes, providing additional convergent evidence for replication consistency.

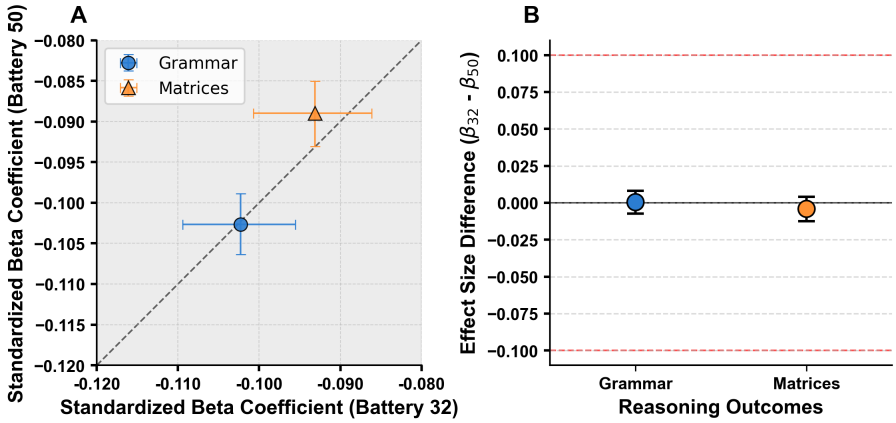

Figure 2: TMT complexity effects on reasoning show robust cross-battery replication with statistical equivalence. Standardized beta coefficients demonstrate remarkable consistency across independent samples, with both reasoning measures clustering tightly around perfect replication. The negligible effect size differences (≤ 0.004 standardized units) fall well within pre-specified equivalence bounds, supporting reliable measurement of cognitive complexity-reasoning relationships. Panel A displays direct comparison of effect sizes between Battery 32 (x-axis) and Battery 50 (y-axis), with grammatical reasoning (blue circles) and progressive matrices (orange triangles) positioned near the diagonal reference line (dashed black) indicating perfect replication. Gray shaded region marks ± 0.1 equivalence bounds. Panel B shows effect size differences between batteries (Battery 32 - Battery 50) with horizontal reference lines at perfect replication (solid black, 0) and equivalence bounds (red dashed, ± 0.1). Error bars represent 95% bootstrap confidence intervals throughout. TMT complexity calculated as $\log(\text{TMT-B}/\text{TMT-A})$ standardized beta coefficients from hierarchical regression controlling for age, processing speed, and working memory. Two One-Sided Tests confirmed statistical equivalence for both outcomes ($p < 0.001$) using ± 0.1 bounds. Sample sizes: Battery 32 $n = 71,114$, Battery 50 $n = 225,092$.

Effect Size Interpretation and Theoretical Implications

Small effects are common in individual-differences applications of robust cognitive tasks (Hedge et al., 2018), and the observed TMT complexity effects, while statistically robust and highly replicable, represent small effect sizes in absolute terms. Incremental R^2 values ranging from 0.76% to 1.02% indicate that complexity differences explain modest proportions of reasoning variance beyond established predictors. However, the precision of replication across independent samples totaling nearly 300,000 participants suggests these represent reliable population parameters rather than sampling artifacts. The complexity differences are often interpreted as cognitive flexibility and set maintenance (Kortte et al., 2002) and are consistent with general task-switching costs (Monsell, 2003) in line with the unity/diversity framework of executive functions (Miyake et al., 2000). This pattern may reflect individual differences in cognitive efficiency rather than executive control capacity per se, particularly on tasks requiring relational integration (Halford et al., 1998). Multicollinearity assessments revealed no problematic intercorrelations among predictors with maximum VIFs less than 1.77 and were well below levels of concern (O'brien, 2007), confirming that complexity effects were independent of processing speed and working memory controls. The manipulation checks demonstrated expected theoretical relationships: TMT-A correlated negatively with processing speed measures ($r \approx -0.49$), working memory capacity showed positive associations with reasoning as expected from developmental cascade accounts (Fry and Hale, 1996; Kyllonen and Christal, 1990), and span tasks demonstrated appropriate intercorrelations in line with span task intercorrelations (Conway et al., 2005) ranging from $r = 0.39$-0.59, validating the theoretical constructs underlying the analyses.

4 Discussion

These findings illuminate a longstanding theoretical tension in cognitive science between processing speed and executive attention accounts of intelligence, dating back to Spearman (1904) foundational work on general intelligence. The current study's results suggest a more nuanced understanding than either pure processing speed models (Kail and Salthouse, 1994) or strong executive function theories (Engle et al., 1999; Kane and Engle, 2002; Mashburn et al., 2023) would predict. The modest but reliable effects observed indicate that neither framework fully captures the complexity of cognitive coordination mechanisms underlying reasoning abilities, particularly when assessed through computerized versions of traditional neuropsychological measures.

The observed Trail Making Test complexity effects reveal a pattern that defies simple theoretical categorization. TMT complexity differences consistently predicted reasoning performance beyond shared associations with processing speed and working memory capacity, indicating that coordination demands capture unique cognitive variance. However, the magnitude of these effects was modest, with standardized coefficients showing small but reliable negative associations across both test batteries. This pattern suggests neither the fundamental importance proposed by executive attention theories nor the negligible role implied by pure processing speed accounts. Importantly, these small effect sizes may have limited practical significance for clinical or educational applications, as the additional variance explained represents a minimal contribution to real-world cognitive performance.

Rather than proposing entirely new theoretical constructs, these findings may be best understood within existing frameworks that emphasize the fractionation of executive functions (Miyake et al., 2000; Diamond, 2013). The coordination demands captured by TMT complexity differences likely reflect the efficiency with which individuals coordinate multiple cognitive operations - maintaining attention to relevant stimulus dimensions while inhibiting prepotent responses and updating working memory representations (Baddeley and Hitch, 1974; Baddeley, 2012). This coordination efficiency represents a genuine individual difference that is associated with reasoning performance, albeit within constrained effect size boundaries. However, this interpretation remains speculative, as the negative relationships could equally reflect measurement artifacts specific to computerized assessment contexts or speed-accuracy tradeoffs inherent in online cognitive testing.

The persistence of TMT complexity effects after controlling for digit symbol coding performance (Wechsler, 1955; Joy, 2004) provides evidence against purely processing speed-based explanations. Both measures involve speeded performance, yet TMT complexity captured additional variance beyond basic processing efficiency. However, several alternative explanations warrant consideration. Speed-accuracy tradeoffs may lead participants to sacrifice accuracy for speed on complex tasks, potentially confounding complexity measures with strategic differences. Task-specific artifacts related to computerized administration may differ substantially from traditional paper-and-pencil neuropsychological assessment, limiting generalizability to clinical contexts where traditional TMT administration predominates. The log-ratio transformation, while addressing distributional concerns, may introduce systematic biases that contribute to the observed negative relationships through measurement confounds rather than genuine cognitive mechanisms.

These findings connect to broader theoretical developments in cognitive control research, though their interpretation requires caution given the specific constraints of computerized TMT assessment. The modest but reliable TMT complexity effects are consistent with theories emphasizing distributed

cognitive mechanisms rather than a single executive system (Duncan et al., 2020; Salthouse, 2011). Individual differences in coordination efficiency may reflect optimization of neural networks supporting attention switching (Posner and Petersen, 1990; Petersen and Posner, 2012) and working memory updating (Conway et al., 2003), consistent with neurocognitive models that emphasize network efficiency (Corbetta and Shulman, 2002). However, the cross-sectional observational design precludes causal inferences about these neural mechanisms, and the negative direction of effects suggests that alternative explanations related to task demands or measurement artifacts may be equally plausible.

The theoretical implications extend to dual-process theories of cognition (Evans, 2008; Kahneman, 2003), where controlled processes supplement automatic processing. TMT complexity differences may index individual variation in the efficiency of controlled processing mechanisms (Stanovich and West, 2000), particularly those involved in maintaining and manipulating multiple cognitive representations simultaneously. This interpretation aligns with evidence for genetic influences on individual differences in executive functions (Friedman et al., 2008), suggesting that coordination efficiency represents a measurable individual difference. However, this mechanistic account remains speculative given the correlational nature of the data and the possibility that observed relationships reflect methodological factors rather than genuine cognitive processes.

The current findings provide limited evidence regarding domain-general executive function models, as the study examined TMT complexity in relation to only two reasoning tasks within a specific online assessment context. While the small effect sizes observed align with recent findings questioning the broad importance of executive functions in cognitive abilities (Frischkorn et al., 2020), the scope of evidence is insufficient to challenge comprehensive models of executive control. The specificity of TMT complexity effects may support models emphasizing fractionated executive functions, where different coordination demands show distinct patterns of association with cognitive outcomes, but this conclusion extends beyond what the current measures and analyses can definitively support.

The study advances methodological standards in cognitive psychology through several approaches that address replicability concerns, though some innovations require careful interpretation. The implementation of Two One-Sided Tests (TOST) equivalence testing (Schuirmann, 1987; Lakens, 2017) represents a shift from traditional null hypothesis testing toward precision estimation in replication research, addressing fundamental concerns about the reproducibility crisis in psychological science (Open Science Collaboration, 2015; Ioannidis, 2005). The massive sample sizes employed enable detection of theoretically meaningful but empirically small effects that typical studies cannot reliably identify, directly addressing chronic statistical power limitations that have plagued psychological research (Cohen, 1988; Button et al., 2013). Cross-battery independence ensures that replication tests are based on statistically independent observations, while analytical innovations including log-ratio transformations and Median Absolute Deviation approaches (David and Tukey, 1977; Leys et al., 2013) for robust outlier detection provide methodological improvements. The implementation of HC3 robust standard errors (MacKinnon and White, 1985; Long and Ervin, 2000) addresses heteroscedasticity concerns in hierarchical regression models, collectively establishing rigorous analytical standards for large-scale cognitive research.

Acknowledgments

The authors thank the participants who contributed their data to the NeuroCognitive Performance Test dataset, making this large-scale investigation possible. We acknowledge Lumos Labs, Inc. for providing access to the de-identified cognitive assessment data used in this study.

Funding

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

References

Arbuthnott, K. & Frank, J. (2000). Trail making test, part b as a measure of executive control: Validation using a set-switching paradigm. *Applied Neuropsychology*, *22*(4), 518–528, doi:10.1076/1380-3395(200008)22:4;1-0;FT518.

Arán Filippetti, V. & Gutierrez, M. (2024). Interpreting the direct- and derived-trail making test scores in argentinian children: Regression-based norms, convergent validity, test-retest reliability, and practice effects. *The Clinical Neuropsychologist*, *39*(6), 1696–1721, doi:10.1080/13854046.2024.2423414.

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.

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159, doi:10.1037/0033-2909.112.1.155.

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. *Psychonomic Bulletin & Review*, *12*(5), 769–786, doi:10.3758/bf03196772.

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.

Cook, R. J., Dickens, B. M., & Fathalla, M. F. (2003). *World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects*, (pp. 428–432). Oxford University Press.

Corbetta, M. & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215, doi:10.1038/nrn755.

Corrigan, J. D. & Hinkeldey, N. S. (1987). Relationships between parts a and b of the trail making test. *Archives of Clinical Neuropsychology*, *2*(4), 329–334, doi:10.1017/S1355617700000606.

Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. *McGill University*.

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

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

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.

Duncan, J., Assem, M., & Shashidhara, S. (2020). Integrated intelligence from distributed brain activity. *Trends in Cognitive Sciences*, *24*(10), 838–852, doi:10.1016/j.tics.2020.06.012.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64, doi:10.1080/01621459.1961.10482090.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), doi:10.1214/aos/1176344552.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics.

Efron, B. & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), doi:10.1214/ss/1177013815.

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23, doi:10.1111/1467-8721.00160.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). *Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex*, (pp. 102–134). Cambridge University Press, 1 edition.

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477, doi:10.1073/pnas.1818430116.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*(1), 255–278, doi:10.1146/annurev.psych.59.103006.093629.

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.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, *137*(2), 201–225, doi:10.1037/0096-3445.137.2.201.

Frischkorn, G. T., von Bastian, C. V., Souza, A. S., & Oberauer, K. (2020). Individual differences in updating are not related to reasoning ability and working memory capacity. *Journal of Experimental Psychology: General*, doi:10.31234/osf.io/92uhb.

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.

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.

Hampshire, A., Highfield, R. R., Parkin, B. L., & Owen, A. M. (2012). Fractionating human intelligence. *Neuron*, *76*(6), 1225–1237, doi:10.1016/j.neuron.2012.06.022.

Hawkins, D. M. (1980). *Identification of outliers*. Springer Netherlands.

Hayes, A. F. & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in ols regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722, doi:10.3758/bf03192961.

Healy, J. D. (1987). A note on multivariate cusum procedures. *Technometrics*, *29*(4), 409–412, doi:10.2307/1269451.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186, doi:10.3758/s13428-017-0935-1.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124, doi:10.1371/journal.pmed.0020124.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: Cowat, bnt, mae token, wrat-r reading, amnart, stroop, tmt, and jlo. *The Clinical Neuropsychologist*, *10*(3), 262–278, doi:10.1080/13854049608406689.

Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, *36*(1), 81–95, doi:10.1016/j.intell.2007.06.001.

Jones, E., Oliphant, T., & Peterson, P. (2001). *SciPy: Open source scientific tools for Python*.

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.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697–720, doi:10.1037/0003-066x.58.9.697.

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. *Progress in Brain Research*, *133*, 309–323, doi:10.1016/S0079-6123(02)33024-9.

Kane, M. J. & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to stroop interference. *Journal of Experimental Psychology: General*, *132*(1), 47–70, doi:10.1037/0096-3445.132.1.47.

Kaufman, A. S. (1983). Test review: Wechsler, d. manual for the wechsler adult intelligence scale, revised. new york: Psychological corporation, 1981. *Journal of Psychoeducational Assessment*, *1*(3), 309–313, doi:10.1177/073428298300100310.

Kortte, K. B., Horner, M. D., & Windham, W. K. (2002). The trail making test, part b: Cognitive flexibility or ability to maintain set? *Applied Neuropsychology*, *9*(2), 106–109, doi:10.1207/s15324826an0902_5.

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.

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses.

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.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341, doi:10.1641/0006-3568(2001)051[0341:lndats]2.0.co;2.

Long, J. S. & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224, doi:10.1080/00031305.2000.10474549.

López-Martos, D., et al. (2023). Reference data for attentional, executive, linguistic, and visual processing tests obtained from cognitively healthy individuals with normal alzheimer's disease cerebrospinal fluid biomarker levels. *Journal of Alzheimer's Disease*, *95*(1), 237–249, doi:10.3233/JAD-230290.

MacKinnon, J. G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325, doi:10.1016/0304-4076(85)90158-7.

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.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the Python in Science Conference* (pp. 56–61).

Mettler, F. A. & Halstead, W. C. (1948). Brain and intelligence; a quantitative study of the frontal lobes. *The American Journal of Psychology*, *61*(3), 440, doi:10.2307/1417176.

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.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140, doi:10.1016/s1364-6613(03)00028-7.

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

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.

O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690, doi:10.1007/s11135-006-9018-6.

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

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, *9*(3), 10–20, doi:10.1109/mcse.2007.58.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), doi:10.1126/science.aac4716.

Osborne, J. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. SAGE Publications, Inc.

Petersen, S. E. & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, *35*(1), 73–89, doi:10.1146/annurev-neuro-062111-150525.

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

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

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

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology*, *19*(1), 137–150, doi:10.1111/j.2044-8341.1941.tb00316.x.

Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology*, *19*(5), 393–394, doi:10.1037/h0044509.

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.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363, doi:10.1037/1089-2680.7.4.331.

Rousseeuw, P. J. & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283, doi:10.2307/2291267.

Sainani, K. L. (2012). Dealing with non-normal data. *PM&R*, *4*(12), 1001–1005, doi:10.1016/j.pmrj.2012.10.013.

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. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology*, *19*(4), 532–545, doi:10.1037/0894-4105.19.4.532.

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.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680, doi:10.1007/bf01068419.

Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the Python in Science Conference* (pp. 92–96).

Sheppard, L. D. & Vernon, P. A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality and Individual Differences*, *44*(3), 535–551, doi:10.1016/j.paid.2007.09.015.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2015). *False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant*, (pp. 547–555). American Psychological Association.

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

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665, doi:10.1017/s0140525x00003435.

Sternberg, R. J. & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology*, *49*, 479–502, doi:10.1146/annurev.psych.49.1.479.

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., & Barceló, 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.

Tombaugh, T. (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.

Tukey, J. W. (1957). On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, *28*(3), 602–632, doi:10.1214/aoms/1177706875.

Vallar, G. & Baddeley, A. D. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 151–161, doi:10.1016/s0022-5371(84)90104-x.

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

van Rossum, G. & Drake, F. L. (2009). *Python 3 reference manual*.

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*.

Weisberg, S. & Fox, J. (2010). *An R companion to applied regression*.

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC, 0 edition.

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

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817, doi:10.2307/1912934.

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

5 Supplementary Material

Supplementary Methods and Results

The analysis employed a comprehensive five-step pipeline designed to ensure methodological rigor and statistical independence across two large-scale cognitive test batteries. This approach enabled robust evaluation of Trail Making Test complexity effects on reasoning abilities through independent replication analyses.

Data Processing and Quality Control Procedures

The analysis began with harmonization of two independent cognitive test batteries from the NeuroCognitive Performance Test dataset. Battery 32 initially contained 851,177 individual subtest records, while Battery 50 contained 2,302,948 records. Systematic filtering procedures retained only complete test administrations with valid grand index scores, removing 55,145 records from Battery 32 and 131,980 records from Battery 50. Subtest selection focused on seven core cognitive measures: Trail Making Test Parts A and B, grammatical reasoning, progressive matrices, digit symbol coding, and forward and reverse memory span tasks.

Cross-battery deduplication procedures identified 62 participants who completed assessments in both batteries. To maintain statistical independence required for replication analyses, these participants were retained only from the battery containing their earliest test completion, resulting in removal of 90 cross-battery records. Data restructuring converted long-format records to wide-format test runs, yielding 87,065 test runs in Battery 32 and 266,442 test runs in Battery 50.

Robust quality control procedures included feasibility bounds for cognitive test performance and outlier detection using Median Absolute Deviation-based z-scores. Trail Making Test completion times below 15 seconds (TMT-A) or 20 seconds (TMT-B) were excluded as implausible, as were times exceeding 300 seconds (TMT-A) or 600 seconds (TMT-B). Participants with absolute robust z-scores exceeding 3.5 on any cognitive measure were removed through listwise deletion. These procedures excluded 4,027 participants from Battery 32 and 12,792 participants from Battery 50, resulting in final analytic samples of 71,114 participants in Battery 32 and 225,092 participants in Battery 50.

Theoretical Construct Validation

Manipulation checks confirmed expected theoretical relationships between cognitive constructs across both independent batteries, demonstrating robust TMT complexity-reasoning associations across independent samples. Processing speed measures showed strong negative correlations between Trail Making Test Part A completion times and digit symbol coding performance in both Battery 32 ($r = -0.488$) and Battery 50 ($r = -0.496$), validating their shared measurement of processing speed. Working memory capacity composites, derived from standardized forward and reverse span scores, demonstrated moderate positive correlations with both reasoning outcomes across batteries, with slightly stronger relationships observed in Battery 50 compared to Battery 32.

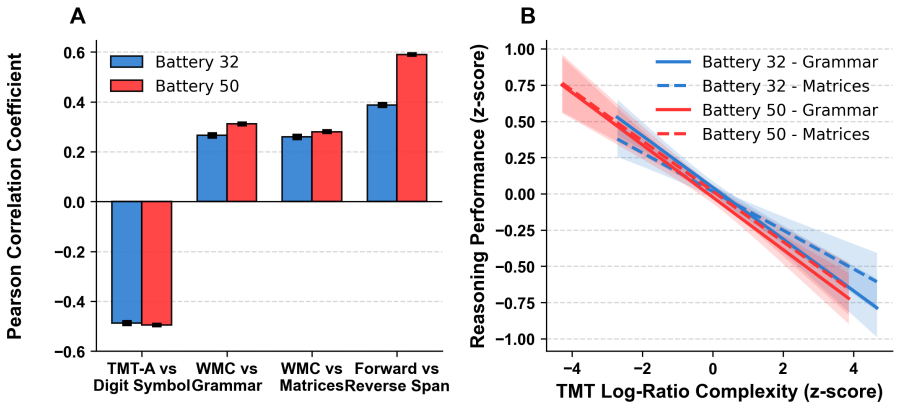

Figure 3: Theoretical construct validation confirms expected cognitive relationships and demonstrates robust TMT complexity-reasoning associations across independent samples. Panel A reveals consistent theoretical correlations validating cognitive constructs across batteries, with stronger working memory-reasoning relationships in Battery 50 than Battery 32. Strong negative TMT-A/digit symbol correlations confirm both assess processing speed; moderate positive working memory-reasoning correlations support expected cognitive architecture; substantial forward-reverse span intercorrelations validate working memory assessment. Panel B demonstrates consistent negative TMT complexity-reasoning relationships before covariate control, with nearly identical slopes across batteries supporting replication robustness. Higher complexity scores indicate proportionally slower TMT-B relative to TMT-A performance, consistently predicting poorer reasoning across conditions. Blue bars/lines represent Battery 32, red bars/lines represent Battery 50. Panel A shows 95% confidence intervals as error bars; Panel B displays regression lines with 95% confidence intervals as shaded regions. Solid lines indicate grammatical reasoning, dashed lines indicate progressive matrices. TMT log-ratio calculated as $\log(\text{TMT-B}) - \log(\text{TMT-A})$; working memory composite derived from standardized forward and reverse span scores. All Panel A correlations significant at $p < 0.001$. Sample sizes: Battery 32 $n = 71,114$, Battery 50 $n = 225,092$. Panel B visualization uses 2000 randomly sampled points per battery-outcome combination.

Forward and reverse span tasks showed substantial intercorrelations supporting their combination into working memory composites, with correlations of $r = 0.387$ in Battery 32 and $r = 0.590$ in Battery 50. Most importantly for the primary research hypotheses, TMT log-ratio complexity metrics showed consistent negative relationships with reasoning abilities across both batteries, with comparable effect sizes for grammatical reasoning (Battery 32: $r = -0.169$; Battery 50: $r = -0.172$) and progressive matrices (Battery 32: $r = -0.146$; Battery 50: $r = -0.142$).

Cross-Battery Replication Analysis

Hierarchical regression analyses revealed remarkably consistent TMT complexity effects across independent samples. For grammatical reasoning, standardized beta coefficients were nearly identical between Battery 32 ($\beta = -0.1023$, $SE = 0.0035$) and Battery 50 ($\beta = -0.1027$, $SE = 0.0019$). Progressive matrices showed similarly consistent effects (Battery 32: $\beta = -0.0931$, $SE = 0.0037$; Battery 50: $\beta = -0.0890$, $SE = 0.0021$). All complexity effects remained statistically significant after Bonferroni correction for multiple testing ($p < 0.025$), indicating that greater proportional increases in TMT-B relative to TMT-A completion times consistently predicted poorer reasoning performance across both samples.

Statistical comparison of beta coefficients between batteries revealed no significant differences for either reasoning outcome. For grammatical reasoning, the coefficient difference was $\beta_{\text{diff}} = 0.0004$ ($SE = 0.0040$, $t = 0.096$, $p = 0.924$), while progressive matrices yielded $\beta_{\text{diff}} = -0.0041$ ($SE = 0.0042$, $t = -0.98$, $p = 0.326$). Bootstrap confidence intervals for complexity effects showed substantial overlap between batteries, further supporting effect size consistency.

Statistical Equivalence Testing

Two One-Sided Tests procedures with ± 0.1 equivalence bounds provided definitive evidence for practical equivalence between battery-specific effects. For grammatical reasoning, both lower ($t = 25.09$, $p < 0.001$) and upper ($t = -24.90$, $p < 0.001$) equivalence tests reached significance, yielding an overall TOST p-value of $p < 0.001$. Progressive matrices demonstrated similar equivalence patterns (lower: $t = 22.71$, $p < 0.001$; upper: $t = -24.67$, $p < 0.001$; combined: $p < 0.001$). These results conclusively demonstrated that any true differences between battery-specific effects fall below the threshold considered practically meaningful for this research domain.

The convergent evidence from significance testing, effect size consistency, confidence interval overlap, and formal equivalence testing established robust replication of TMT complexity-reasoning relationships across independent samples totaling 296,206 participants. This methodological approach provided compelling evidence that cognitive complexity differences captured by proportional TMT performance changes represent a reliable predictor of reasoning abilities across diverse testing contexts.