Bilingualism: Language and Cognition

cambridge.org/bil

Domain-dependent mechanisms of L2 proficiency in interference suppression: A drift-diffusion analysis of verbal and spatial Stroop in Chinese–English bilinguals

Liqun Wang[1], Ruili Chu[1], Tongyi Zhang[2], Hongli Li[3], Haoyang Shi[2] and Zihao Wang[2]

[1]School of Foreign Languages, Weinan Normal University, Weinan, China; [2]School of Psychology, Northwest Normal University, Lanzhou, China and [3]School of Education, Anyang Normal University, Anyang, China

Research Article

Cite this article: Wang, L., Chu, R., Zhang, T., Li, H., Shi, H. and Wang, Z. (2026). Domain-dependent mechanisms of L2 proficiency in interference suppression: A drift-diffusion analysis of verbal and spatial Stroop in Chinese–English bilinguals. *Bilingualism: Language and Cognition* 1–16. https://doi.org/10.1017/S1366728926101965

Received: 24 October 2025 UTC
Revised: 16 September 2026 UTC
Accepted: 17 September 2026 UTC

Keywords:
bilingual advantage; Chinese–English bilinguals; drift-diffusion modeling; interference suppression; second language proficiency

Corresponding author:
Zihao Wang;
Email: zhqaz814123@163.com

This research article was awarded Open Data badge for transparent practices. See the Data Availability Statement for details.

Abstract

Whether and how second-language (L2) proficiency shapes cognitive control remain debated. We examined its influence on interference suppression across verbal and spatial domains in 154 Chinese–English bilinguals who completed parallel *verbal and spatial Stroop tasks*. L2 proficiency was modeled in three groups and continuously, and latent processes were decomposed via hierarchical Bayesian drift-diffusion modeling. Higher proficiency was associated with reduced interference in both domains, with a greater reduction in the verbal domain. Group-based model comparison favored verbal drift-rate variation and spatial decision-threshold variation. Joint regressions included proficiency, daily L2 use, L2 exposure, age of acquisition and years of learning. Proficiency was associated with higher drift rates and thresholds and shorter nondecision times in both tasks. Daily L2 use and L2 exposure were associated with verbal and spatial parameters, respectively. These findings provide computational evidence for a task-dependent account of bilingual cognitive control involving proficiency and other experience dimensions.

Highlights

- Higher L2 proficiency was associated with less verbal and spatial interference.
- The verbal benefit was carried mainly by a higher drift rate.
- The spatial benefit operated through a higher decision threshold.
- Findings support a task-dependent account of bilingual cognitive control.
- Daily L2 use and exposure showed conditional associations with decision parameters.

1. Introduction

Bilingual experience has been hypothesized to confer cognitive advantages, particularly in executive control (Bialystok, 2024; Bialystok & DePape, 2009); however, whether such advantages exist and how generalized they are remain debated (Paap et al., 2015), with disagreements in evidence syntheses. Meta-analytic estimates range from near zero once publication bias is corrected (Lehtonen et al., 2018) to small, task- and age-dependent advantages (Ware et al., 2020). A systematic review of 46 studies found that 54.3% favored bilinguals, 28.3% were mixed and 17.4% reported evidence against an advantage, attributing much of this divergence to methodological heterogeneity and the neglect of individual differences (van den Noort et al., 2019). A Bayesian analysis of 167 independent studies revealed that when group differences occurred, they favored bilinguals more often than monolinguals (Grundy, 2020). Analytic choices further influence whether an effect is observed (Privitera et al., 2023b). Taken together, these syntheses do not indicate a uniform presence or absence of an advantage but suggest a moderated advantage; its detection depends on the sample, task and analysis. This pattern motivates a finer-grained, mechanism-focused approach. Consistent with the view that bilingual experience produces adaptations linked to specific demands rather than uniform transfer (Bialystok, 2024), the central question shifts from whether bilingualism affects cognition to how any such effect is realized in underlying decision processes.

Several factors contribute to this inconsistency. First, the bilingual experience is itself heterogeneous, spanning individual differences in L2 proficiency, age of acquisition, frequency of use and interactional context (Gullifer et al., 2021; Semenova et al., 2026; Valian, 2015; Xia et al., 2026). Second, most prior work has contrasted "bilinguals" with "monolinguals," a comparison that introduces systematic confounds (Valian, 2015; Yow & Li, 2015). Inconsistencies persist even within relatively homogeneous samples (Privitera et al., 2023a); therefore, sampling heterogeneity alone cannot explain the mixed findings. Third, differences in tasks,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101965

populations and analytic approaches further inflate between-study variance (Privitera et al., 2023b; Privitera & Weekes, 2023; van den Noort et al., 2019). These observations have prompted complementary refinements, from richer multidimensional assessment of bilingual experience to sampling beyond the bilingual–monolingual binary and trial-level computational methods that recover mechanisms obscured by aggregate measures (Dash et al., 2022; Gullifer et al., 2021; Privitera et al., 2023a; Rothman et al., 2023; Titone & Tiv, 2023; Tiv et al., 2022), alongside theoretical reframing of the field's core question (Bialystok, 2024). This study draws on both. Following Rothman et al. (2023), we treat bilingualism as a spectrum rather than a category (cf. DeLuca et al., 2019; Luk & Bialystok, 2013) and jointly examine proficiency, daily L2 use, L2 exposure, age of acquisition and years of learning within a bilingual sample. We pair this multidimensional approach with hierarchical Bayesian drift-diffusion modeling to examine how these dimensions relate to the components underlying performance in verbal and spatial conflict tasks.

Second language (L2) proficiency has been associated with performance differences; however, the cognitive basis remains unclear. Among Chinese–English bilinguals, Wang et al. (2016) found smaller Chinese *Stroop* interference but comparable English interference in more proficient participants. They attributed this pattern to contributions from word-recognition speed and cognitive control. Moreover, Xie (2018) found faster responses across *Flanker* conditions in the highest proficiency group than in the lowest, without a corresponding difference in interference. Tse and Altarriba (2012) linked proficiency to different aspects of *Stroop* reaction time (RT) distributions using ex-Gaussian distributional analyses. These findings motivated us to focus on proficiency with the following questions: which decision components accompany these performance differences? Do their associations vary across verbal and spatial conflict?

Proficiency and daily language use are correlated but distinguishable dimensions (Luk & Bialystok, 2013). Other dimensions also show meaningful cognitive associations. Yow and Li (2015), for example, found that earlier L2 acquisition and more balanced language use predicted smaller *Stroop* interference. In contrast, proficiency balance did not significantly predict interference in a separate regression model. In Chinese learners of English, Xia et al. (2026) reported cognitive-control associations involving proficiency and exposure. Daily use and exposure describe current experience, whereas age of acquisition and years of learning describe when learning began and how long it has continued. We therefore examine all five dimensions as predictors of interest and interpret each association with decision processes as conditional on the others. Proficiency is the focal dimension because of this specific process-level question, not because it is assumed to be more influential than the other dimensions or to represent bilingual experience as a whole.

Accounts of bilingual cognitive control diverge most sharply on one underexamined question: the domain specificity of proficiency effects. Three positions can be distinguished. The domain-general enhancement view draws on evidence that bilinguals recruit language-control regions during nonlinguistic switching (Rodríguez-Pujadas et al., 2013). When applied to the present study, a strong version of this account predicts comparable proficiency effects across linguistic and nonlinguistic tasks. Language-control accounts describe how bilinguals manage competition between their languages (Abutalebi & Green, 2007; Green, 1998). A domain-specific restriction hypothesis predicts that any proficiency-related benefit remains confined to the linguistic task. The task-dependent modulation view holds that proficiency effects are distributed differentially across domains – neither symmetric across tasks nor confined to verbal ones. This is consistent with the unity-and-diversity architecture of executive function, in which control draws on shared and domain-specific components (Friedman & Miyake, 2017; Miyake et al., 2000). The well-documented inconsistency of bilingual advantages on nonlinguistic interference tasks (Hilchey & Klein, 2011) sits more comfortably with this task-dependent view than with strong, uniform domain-general enhancement – these positions concern where an effect appears, not which latent process carries it. Decomposing performance using the diffusion framework lets us ask the second question – whether proficiency operates through evidence-accumulation efficiency (drift rate), response caution (decision threshold) or nondecision time – and determine whether the answer differs by domain.

The proficiency findings also point to a limitation of measurement. The literature has primarily analyzed aggregate behavioral indices (mean RT and accuracy) on classical conflict tasks such as the *Stroop*, *Simon* and *Flanker*. However, a faster mean RT can reflect more efficient information processing, reduced response caution or quicker motor execution – possibilities with different theoretical implications that mean RT and accuracy alone cannot separate (Ambrosi et al., 2019; Forstmann et al., 2016). The challenge is therefore to separate these sources of variance within the same behavioral data.

Testing these accounts raises two further challenges. First, cognitive tasks differ in difficulty, load and processing demands, complicating cross-task comparison. Second, even an ostensibly nonlinguistic task may in principle be modulated by verbal labeling of stimulus features (Lupyan, 2012). If such labels are internally generated – for example, "left" and "right" during a spatial task – this endogenous pathway does not depend on L2 stimuli and differs from the stimulus-driven alternative, in which L2 stimuli trigger lexical access. A credible spatial-task proficiency effect in the absence of L2 stimuli would rule out the stimulus-driven account but not an internal-recoding contribution. To meet the first challenge, we use structurally parallel *verbal and spatial Stroop tasks*. Both require responding in the presence of conflicting information and therefore are matched on fundamental control demands; they differ in the nature of the conflict, which is semantic in the verbal task and visuospatial in the spatial task. Comparing the pattern of proficiency effects across the two tasks directly relates to the domain specificity of the effects of bilingual experience. Conflict-task measures can show weak intercorrelations and may involve partly distinct processes (Friedman & Miyake, 2004); however, this comparison is only informative if the analysis can localize each effect within the decision process, which aggregate RT and accuracy cannot.

The drift-diffusion model (DDM) provides this localization, decomposing aggregate behavior into latent cognitive processes (Ratcliff et al., 2016). It treats a decision as the accumulation of noisy evidence toward one of two response boundaries. By fitting the full RT distribution with accuracy, it estimates several interpretable parameters. Drift rate captures the efficiency of evidence accumulation and, therefore, the speed and quality of information processing. Because conflict slows and diverts accumulation, the drift rate is lower in incongruent trials than in congruent trials. Therefore, a higher incongruent trial drift rate reflects more efficient accumulation of task-relevant evidence under conflict – greater resistance to interference – rather than conflict resolution (Fennell & Ratcliff, 2019). Decision threshold (boundary separation) reflects the evidence required before responding and governs

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

the speed–accuracy tradeoff. The starting point captures a priori response bias, and nondecision time includes the time required for perceptual encoding and motor execution. A hierarchical Bayesian implementation estimates these parameters jointly at the group and individual levels, improving their reliability (Wiecki et al., 2013). Few studies have applied the DDM to bilingual cognition (Hartanto & Yang, 2016; Ong et al., 2017, 2019; Soares et al., 2019), and nondecision time is the most consistently reported locus. These studies show that nondecision time is shorter on incongruent *Flanker* trials in older bilinguals than in older monolinguals, with no group difference in drift rate or threshold (Ong et al., 2017), and that a dual-language switch-cost advantage is expressed in nondecision time rather than drift rate (Hartanto & Yang, 2016). The effects of L2 proficiency on drift rate and threshold based on the DDM have not been thoroughly studied. In adjacent individual-differences work, the common factor across executive-function tasks reflects the speed of information uptake, linking performance on these tasks to evidence-accumulation efficiency (Löffler et al., 2024). Against this background, we treat a shorter nondecision time as a domain-general index of processing speed and focus on the drift rate and decision threshold, the two parameters corresponding to the candidate mechanisms, namely accumulation efficiency and response caution, through which proficiency might act.

Decomposed this way, the competing accounts become empirically separable because they do not differ in overall speed but in which parameter carries the proficiency effect and whether it is the same across domains. Two candidate mechanisms make uniform predictions. If acquiring proficiency works as sustained control training, its benefit should appear as a higher drift rate – more efficient evidence accumulation – in both tasks. Practice can make L2 processing more automatic (DeKeyser, 2007). If this reduces response caution during conflict, we would predict a lower decision threshold in both tasks. The domain-specificity accounts then specify where any such effect should appear, so that a contrast between two structurally matched tasks can distinguish them. An effect on the same parameter in both tasks would support domain-general enhancement; an effect confined to the verbal task would support domain-specific restriction and effects on different parameters across the two tasks would support task-dependent modulation, with the locus of the effect reflecting each task's control demands. Two parameters serve as anchors for interpretation. The starting point should be unaffected, as no account predicts a proficiency-driven response bias. Nondecision time includes perceptual encoding and response execution (Ratcliff et al., 2016). Drawing on a general processing-speed account (Salthouse, 1996), we predict shorter nondecision times in both tasks if proficiency is associated with faster processing outside evidence accumulation; an additional English lexical-access contribution is only expected in the verbal task. Comparing these component-level predictions across tasks identifies whether L2 proficiency effects are domain-general or domain-specific and the specific computational locus through which they operate.

Therefore, the present study investigates how L2 proficiency influences interference suppression across verbal and spatial domains and identifies the computational locus of any such effect. Chinese–English bilinguals completed structurally parallel *verbal and spatial Stroop tasks*, and we decomposed performance using hierarchical Bayesian DDM into drift rate (processing efficiency), decision threshold (strategic caution), starting point (response bias) and nondecision time (perceptual–motor processes). To balance comparability with the predominantly group-based literature against current best practice, we model L2 proficiency both in tertile groups and continuously, treating the convergence of the two as the central test of robustness. More broadly, the findings bear on theories of bilingual cognition and on the domain specificity of experience-dependent cognitive plasticity.

2. Methods

2.1. Data collection

2.1.1. Participants

No prior study provided an effect-size estimate applicable to a within-bilingual, proficiency-graded contrast. We therefore set no a priori target effect size; instead, we recruited a sample large enough to support stable individual-level estimates in the hierarchical Bayesian DDM and reported a sensitivity analysis for the group comparisons. We recruited 154 first-year English majors from Weinan Normal University (116 females, 38 males; mean age, 19.50 years, standard deviation [*SD*] = 1.20), a distribution reflecting the typical composition of English-major cohorts in China. A sensitivity analysis (G*Power 3.1; Faul et al., 2007) indicated that, for the omnibus comparison of the three proficiency groups (one-way fixed-effects ANOVA, α = .05), this sample afforded 80% power to detect f = .25 (η^2 = .06), the conventional benchmark for a medium effect (Cohen, 1988); the power to detect smaller effects was lower, and we interpret null group differences accordingly. Preliminary analyses revealed no significant main effects or interactions involving sex on any primary dependent variable (all $p > .05$).

Participants were assigned to three groups by L2 proficiency, indexed by the *English Category Verbal Fluency Test* (*CVFT*; Xie, 2018; Xie & Chen, 2026), using tertile splits of the total fluency score: a Low group (<33rd percentile; M = 30.82, SD = 4.75, n = 49), a Mid group (33rd–66th percentile; M = 40.87, SD = 5.69, n = 52) and a High group (>66th percentile; M = 50.36, SD = 5.85, n = 53). To balance comparability with prior group-based work against current best practice (Supplementary Table S1), we modeled L2 proficiency categorically and continuously. We reported the convergence of the two as our central robustness check. The continuous analyses used the raw *CVFT* score and, as a subjective index, the Language Experience and Proficiency Questionnaire (LEAP-Q) total score, each entered as a continuous predictor.

All participants were native Mandarin speakers from nonimmigrant families who began learning English as an L2 at around age 9 and had received formal English instruction for an average of 9.34 years (SD = .85) at the time of testing. Inclusion criteria were as follows: (1) right-handedness; (2) normal or corrected-to-normal vision; (3) no hearing impairment; (4) no history of neurological disorders; (5) no history of attention-deficit/hyperactivity disorder or other cognitive conditions and (6) no history of psychiatric disorders or use of psychotropic medication. We required right-handedness to minimize handedness-related variability in manual response; this was a stand-alone behavioral experiment, and we collected no neural, physiological or other additional measures from this sample. Exclusion criteria were atypical bilingual acquisition (e.g., aphasia recovery) and physical discomfort or abnormal mental state on the testing day. All participants gave written informed consent. The protocol was approved by the ethics review board of Weinan Normal University.

2.1.2. Demographic and linguistic background questionnaires

We collected demographic information from all participants, including age, sex, years of education and years of L2 learning.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

We assessed family socioeconomic status (SES) with a multidimensional composite measure (Zhao et al., 2022) spanning four dimensions: material wealth (Family Affluence Scale; Boyce et al., 2006), parental education, parental occupation and family monthly income. Total SES scores ranged from 5 to 46, with higher scores indicating higher status (see Supplementary Materials).

We used an adapted version of the LEAP-Q (Marian et al., 2007) to assess participants' language background, including learning history, current use patterns, self-rated proficiency and exposure frequency. In the self-rating section, participants rated their listening, speaking, reading and writing skills in both their first language (L1; Chinese) and L2 (English) on a 1–10 Likert scale, where 1 denoted the lowest and 10 denoted near-native proficiency. The four skill ratings were summed to yield a total proficiency score for each language (range 4–40).

We also collected participants' raw scores on the National College Entrance Examination (Gaokao; Cheng & Hamid, 2025) English test (maximum 150 points) as an external proficiency criterion. We used these scores as one of three convergent proficiency indicators rather than as a primary validated measure (see Supplementary Materials).

2.1.3. Raven's standard progressive matrices

Fluid intelligence was assessed with *a study-specific 30-item subset of Raven's Standard Progressive Matrices (SPM;* Raven, *2000), administered under a 20-minute limit* (see Supplementary Materials).

2.1.4. English category verbal fluency

In the bilingual literature, *proficiency* is a broad construct encompassing competence across domains, whereas *fluency* refers more narrowly to the speed and ease of production or retrieval (Segalowitz, 2010); the *CVFT* indexes one performance-based facet of proficiency. We therefore reserve "proficiency" for the broad construct and use "verbal fluency" or "*CVFT* performance" for the task itself.

As an objective measure of proficiency, we administered an English category verbal fluency task in which participants produced as many category-appropriate English words as possible within a time limit (Ardila, 2020) – an established index of lexical access and language fluency (Shao et al., 2014). Three semantic categories were used (occupations, furniture and fruits), each for 60 seconds, and the experimenter recorded all words produced. Scoring excluded repetitions, alternative forms of the same word (e.g., singular and plural) and out-of-category items. The total count of valid words across the three categories served as the *CVFT* score, reflecting English vocabulary size and lexical-retrieval efficiency.

2.1.5. Interference suppression tasks

Verbal Stroop task. We used the classic *color-word Stroop task* to measure interference suppression in the verbal domain (Figure 1; Jensen & Rohwer, 1966). Stimuli were two English color words ("RED" and "GREEN") and a neutral string ("###") as the neutral stimulus, each presented in red or green at the screen center against a gray background. Three conditions were defined: congruent (word meaning matched ink color, e.g., "RED" in red), incongruent (word meaning conflicted with ink color, e.g., "GREEN" in red) and neutral (the string in red or green). Participants ignored word meaning and responded only to ink color, pressing "F" for red and "J" for green. Each trial began with a fixation cross ("+") for 500 ms, followed by target presentation for 1,000 ms and then a blank screen for 1,500 ms before the next trial. The task comprised one practice block and three experimental blocks. The practice block contained 18 trials (six per condition, randomized) with trial-by-trial accuracy feedback; participants had to reach ≥85% accuracy to proceed. Each experimental block contained 36 trials (12 per condition) in pseudorandom order, with no more than three consecutive trials of the same condition, for 108 experimental trials in total (approximately 10–15 minutes). Primary behavioral measures were condition-wise mean RT and accuracy and the *verbal Stroop* interference effect (incongruent − neutral RT).

Spatial Stroop task. We used a *spatial Stroop task* to measure interference suppression in the nonverbal domain (Figure 1; Han et al., 2025). Stimuli were two directional arrows ("←" and "→") and a nondirectional circle ("●"), each appearing in the left or right screen position. Three conditions paralleled the verbal task: congruent (arrow direction matched position, e.g., a left-pointing arrow on the left), incongruent (direction conflicted with position, e.g., a left-pointing arrow on the right) and neutral (the circle in either position). Participants responded to arrow direction or to circle position on neutral trials, pressing "F" for left and "J" for right. Trial sequence, timing, practice criteria and block structure were identical to the verbal task, so that the two tasks differed primarily in the domain of conflict rather than in the procedure. The task again comprised 108 experimental trials (approximately

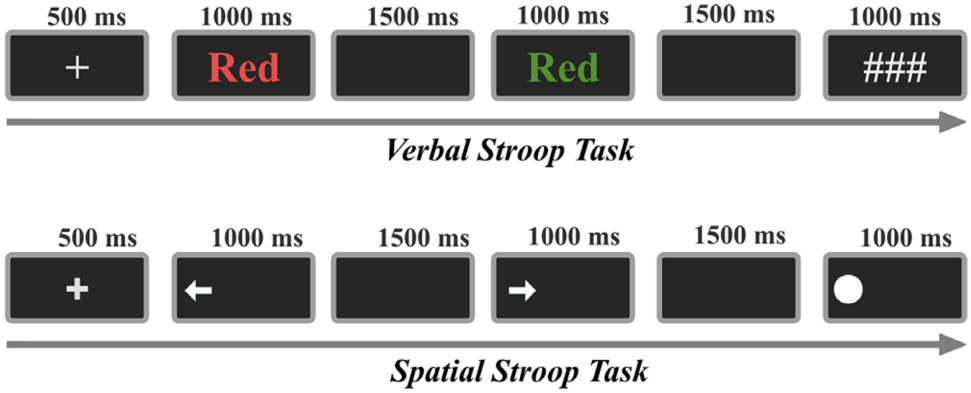


Figure 1. Trial design and procedure for the verbal and spatial *Stroop* tasks.
Note: L2 = second language. The upper panel shows the *verbal Stroop task*, in which participants identified the ink color of English (L2) color words while ignoring word meaning. Stimuli were two English color words ("RED" and "GREEN") in red or green ink, plus a neutral string ("###"). The lower panel shows the *spatial Stroop task*, in which participants reported arrow direction while ignoring screen location, or reported circle position on neutral trials.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

10–15 minutes). Primary measures were condition-wise mean RT and accuracy and the *spatial Stroop* interference effect (incongruent − neutral RT).

2.1.6. Procedure

Participants were tested individually in a quiet laboratory in a single 65- to 75-minute session, in the following order: (1) demographic and language background questionnaires and the *category verbal fluency test* (15 minutes); (2) *Raven's SPM* (20 minutes); (3) a brief break (5 minutes) and (4) the *verbal and spatial Stroop tasks* in counterbalanced order, with a 2-minute break between them (20–30 minutes). All cognitive tasks ran in *E*-Prime 2.0 on a 17-inch LCD monitor (1024 × 768 px, 60 Hz) at a viewing distance of approximately 60 cm.

2.2. Overview of the analytical strategy

Our analyses follow two main lines, including a joint analysis of bilingual experience dimensions, and are supported by robustness checks. The first establishes, behaviorally, whether and where L2 proficiency affects interference suppression; the second examines decision components using proficiency-focused models and joint models of all five experience dimensions. Throughout, L2 proficiency is treated both as tertile groups – retained for comparability with the predominantly group-based literature – and continuously (Rothman et al., 2023), and we report the convergence of the two as the central test of robustness.

Behavioral analyses combine by-participant analyses of covariance (ANCOVAs) on condition-mean RTs, which provide effect sizes comparable to prior work, with trial-level linear mixed-effects models (LMMs) that exploit the full single-trial distribution and accommodate continuous proficiency. Hierarchical Bayesian DDMs (HDDMs) partition performance into drift rate, decision threshold, nondecision time and starting point, with group-based model comparison evaluating candidate patterns of parameter variation across proficiency groups and continuous *CVFT* regression estimating proficiency associations with each parameter. The joint HDDM regression is a main analysis that examines all five measured dimensions of bilingual experience together, estimating each dimension's association conditional on the other four. Two robustness checks then probe the proficiency findings – rebinning proficiency by self-rated LEAP-Q and controlling for English reading – to ensure they are not artifacts of the proficiency index or of shared English demands (details in Supplementary Materials).

2.3. Behavioral and questionnaire analyses

Preprocessing began by removing invalid and outlying trials. For RT analyses, we excluded error trials and trials with RTs beyond 3 SD of each participant's condition mean (Ratcliff, 1993; see Supplementary Materials). Group differences in continuous demographic, language background and proficiency variables were tested with one-way ANOVAs and Tukey HSD post hoc comparisons or Kruskal–Wallis tests when residual normality was not met. Sex distributions were compared using Pearson's χ^2 test (Table 1).

Group differences on the *Stroop tasks* were first examined with by-participant ANCOVAs of condition-mean RTs – covarying age, sex, years of education, SES and fluid intelligence – with Tukey HSD post hoc comparisons. To exploit the full trial-level information, these effects were then confirmed with LMMs on trial-level data, with Group, Condition and their interaction as fixed effects; by-participant random intercepts and random slopes for Condition; the same covariates and estimation by restricted maximum likelihood (see Supplementary Materials). We operationalized proficiency continuously by replacing the categorical group factor with the standardized CVFT score (and, as a robustness check, the LEAP-Q total) in parallel LMMs. Pearson correlations characterized bivariate associations among demographic variables, proficiency indicators and *Stroop* performance. Unless otherwise noted (e.g., Tukey-adjusted pairwise contrasts), *p*-values were corrected for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995).

2.4. Hierarchical Bayesian DDM

We used HDDM (Wiecki et al., 2013) to jointly model trial-level RT and accuracy from both *Stroop tasks*. Each trial supplied participant ID, trial number, group label, condition (congruent, neutral or incongruent), response accuracy (1 = correct, 0 = error) and RT (in seconds). We excluded trials with RTs below 150 ms (likely anticipations) or beyond 3 SD above each participant's condition mean (likely lapses); as is standard for DDM fitting, fast responses above 150 ms were retained, as they carry information about the leading edge of the RT distribution (Ratcliff et al., 2016). Analyses were run in Python 3.8 using HDDM 1.0.1 (Pan et al., 2025; documentation at https://hddm.readthedocs.io/en/latest/index.html).

2.4.1. Model construction

We followed a progressive model-building strategy to localize the proficiency effect across parameters. The baseline model (M_0) included all core parameters – decision threshold (a), drift rate (v), nondecision time (t) and starting point (z) – and estimated trial-to-trial variability in drift rate (s_v), starting point (s_z) and nondecision time (s_t). All parameters were estimated freely, with no a priori response bias imposed. The hierarchical model (M_1) allowed all four parameters to vary across the three proficiency groups and served as an exploratory model to flag parameters that differed between groups. Guided by M_1 and by theory, the selective M2 series then constrained variation to a single parameter per model: $_{M2a}$ (threshold only), $_{M2v}$ (drift rate only), $_{M2t}$ (nondecision time only) and $_{M2z}$ (starting point only). To complement this categorical comparison, we additionally modeled L2 proficiency as a continuous variable. Via HDDMRegressor, the standardized *CVFT* score ($_{CVFTz}$; $M = 0$, $SD = 1$) entered as a participant-level predictor in the linear model for each diffusion parameter, $\theta_i = \beta_0 + \beta_1\, CVFTz_i$, $\theta \in \{a, v, t, z\}$, where β_1 is the regression slope of the corresponding parameter on *CVFT*. For each slope, we reported the posterior mean, posterior *SD*, 95% highest-density interval (HDI) and probability of direction (*pd*; Makowski et al., 2019), applying the same inferential criteria as for the group comparisons. The joint HDDMRegressor model additionally includes standardized daily L2 use, L2 exposure, age of acquisition and years of learning. All five predictors enter the same model for each parameter. Therefore, each coefficient estimates a conditional association rather than a dimension's overall influence or relative importance. The Supplementary Materials provide model details. We report the posterior mean, 95% HDI and the probability of direction for every coefficient.

2.4.2. Estimation, diagnostics and model selection

Parameters were estimated by the Markov chain Monte Carlo method, using four independent chains of 15,000 samples each, with the first 5,000 discarded as burn-in. Convergence was assessed by visual inspection of trace plots for mixing and stationarity and by the Gelman–Rubin statistic ($\hat{R}$; Vats & Knudson, 2021), requiring

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

Table 1. Demographic characteristics and language background of participants

Variable	Low L2 (n = 49)	Mid L2 (n = 52)	High L2 (n = 53)	$F/\chi^2/H$	p_{FDR}
Demographic characteristics					
Age (years)	19.31 ± 1.14	19.55 ± 1.21	19.74 ± 1.20	1.68	.277
Sex (male/female)	16/33	12/40	10/43	2.71[a]	.258
Years of education	12.98 ± .28	13.06 ± .28	13.01 ± .36	.70	.571
SES (range 5–46)	24.97 ± 6.06	25.54 ± 5.78	26.51 ± 6.66	.81	.551
Fluid intelligence (range 0–30)[b]	22.96 ± 1.93	22.95 ± 1.70	23.43 ± 1.77	1.21	.402
Language background					
L1 proficiency (range 4–40)[c]	37.24 ± 2.41	37.48 ± 2.39	37.51 ± 2.37	.41[d]	.815
Age of acquisition (years)	9.17 ± 1.53	9.01 ± 1.50	8.91 ± 1.24	.44	.648
Years of L2 learning	9.32 ± .82	9.43 ± .78	9.27 ± .95	.47	.668
L2 daily use (%)	22.29 ± 10.15	27.41 ± 12.45	32.95 ± 12.05	10.73	<.001***
L2 exposure (%)	27.50 ± 13.63	33.13 ± 13.19	38.95 ± 12.37	9.79	<.001***
L2 proficiency measures					
Category Verbal Fluency Test (CVFT)					
Total score[e]	30.82 ± 4.75	40.87 ± 5.69	50.36 ± 5.85	162.67	<.001***
Occupations	9.23 ± 2.56	12.73 ± 2.55	18.24 ± 3.37	129.02	<.001***
Furniture	12.46 ± 2.42	14.11 ± 2.64	13.76 ± 3.04	14.18[d]	.013*
Fruits	9.12 ± 3.45	14.03 ± 3.31	18.36 ± 3.62	90.41	<.001***
LEAP-Q self-ratings					
Total score (range 4–40)[f]	21.47 ± 2.53	23.54 ± 2.72	26.70 ± 2.57	52.18	<.001***
Listening (range 1–10)	5.20 ± 1.44	5.80 ± 1.38	7.23 ± 1.22	30.78	<.001***
Speaking (range 1–10)	5.51 ± 1.38	5.67 ± 1.50	6.39 ± 1.11	6.37	.005*
Reading (range 1–10)	5.65 ± .91	6.48 ± 1.05	7.20 ± 1.15	28.05	<.001***
Writing (range 1–10)	5.10 ± 1.31	5.59 ± 1.29	5.88 ± 1.35	4.46	.021*
Gaokao English (range 0–150)[g]	117.45 ± 19.10	116.99 ± 13.55	125.64 ± 12.54	5.33	.012*

Note: Values are $M \pm SD$ unless otherwise specified. L1 = first language; L2 = second language; SES = socioeconomic status; *CVFT = Category Verbal Fluency Test*; LEAP-Q = Language Experience and Proficiency Questionnaire. Unless otherwise noted, group differences were tested with one-way ANOVAs (F reported); p-values were adjusted across all comparisons using the Benjamini–Hochberg false discovery rate (FDR) procedure.
[a]Pearson χ^2 test of independence.
[b]Study-specific 30-item subset of Raven's Standard Progressive Matrices (1-point per item).
[c]Sum of self-rated L1 proficiency across four LEAP-Q skill domains (listening, speaking, reading, writing), each rated 1–10.
[d]Kruskal–Wallis test (H reported); see Supplementary Table S2 and Supplementary Figures S1 and S2 for distributional diagnostics.
[e]Total number of unique English words produced across three semantic categories (occupations, furniture, fruits), with 60 s per category; primary criterion for group assignment.
[f]Sum of self-rated L2 proficiency across the four LEAP-Q skill domains, each rated 1–10.
[g]Raw score on the English subtest of the National College Entrance Examination (Gaokao).
* $p_{FDR} < .05$. ** $p_{FDR} < .01$. *** $p_{FDR} < .001$.

$\hat{R} < 1.01$ for all parameters. Competing models were compared using the Deviance Information Criterion (DIC; Wiecki et al., 2013).

2.4.3. Statistical inference

Group comparisons were based on the posterior of M_1, in which all four core parameters could vary by group. For each parameter, we report (1) the posterior mean and 95% HDI of the between-group difference, computed as the first named group minus the second (Low−Mid, Mid−High or Low−High); (2) the Bayes factor (BF_{10}), indexing the strength of evidence for a group difference and (3) pd, treating a difference as credible when its 95% HDI excludes zero.

As a descriptive consistency check on the model-comparison results, we additionally decomposed the proficiency effect through the HDDM parameters using mediation analysis. Because the mediators and the behavioral outcome derive from the same trial-level data, we report this as a descriptive decomposition rather than evidence of a causal mechanism (see Supplementary Materials).

3. Results

3.1. Background characteristics

Table 1 summarizes the demographic and language background characteristics of the three groups. Omnibus tests revealed no significant group differences in any basic demographic variable – age ($F = 1.68$, $p_{FDR} = .277$), sex (χ^2 (2) = 2.71, $p_{FDR} = .258$), years of education ($F = .70$, $p_{FDR} = .571$), SES ($F = .81$, $p_{FDR} = .551$) or fluid intelligence ($F = 1.21$, $p_{FDR} = .402$) – establishing baseline comparability for the between-group analyses that follow. Distributional assumptions were checked, with nonparametric tests used for

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

variables that did not meet residual normality (Supplementary Table S2; Supplementary Figures S1 and S2). The groups also did not differ in L1 proficiency self-ratings (H (2) = .41, p_{FDR} = .815) or years of L2 learning (F = .47, p_{FDR} = .668), indicating comparable native-language ability and length of L2 experience.

The L2 proficiency measures, by contrast, showed clear graded differences. *Category verbal fluency*, the grouping criterion, differed sharply across groups (F = 162.67, p_{FDR} < .001): the High group (50.36 ± 5.85) outscored the Mid group (40.87 ± 5.69, p < .001), which in turn outscored the Low group (30.82 ± 4.75, p < .001). This incremental pattern held across two of the three category subtests – occupations (F = 129.02, p_{FDR} < .001) and fruits (F = 90.41, p_{FDR} < .001), but not for furniture. Although the omnibus test for furniture was significant (H (2) = 14.18, p_{FDR} = .013), the High (13.76 ± 3.04) and Mid (14.11 ± 2.64) groups did not differ (p = .780).

The LEAP-Q total score corroborated the grouping (F = 52.18, p_{FDR} < .001), with the High group (26.70 ± 2.57) above the Mid group (23.54 ± 2.72, p < .001) and the Mid group above the Low group (21.47 ± 2.53, p < .001). The individual skill dimensions varied in how cleanly they separated the groups. Listening (F = 30.78, p_{FDR} < .001) and reading (F = 28.05, p_{FDR} < .001) showed clear three-level gradients. Overall, speaking differed (F = 6.37, p_{FDR} = .005), but only the High–Low contrast was significant (6.39 ± 1.11 versus 5.51 ± 1.38, p = .003); writing differed overall (F = 4.46, p_{FDR} = .021) but yielded no significant pairwise contrast. Gaokao English scores, used as an external criterion, also supported the grouping (F = 5.33, p_{FDR} = .012): the High group (125.64 ± 12.54) outscored the Low (117.45 ± 19.10, p = .020) and Mid (116.99 ± 13.55, p = .012) groups, which did not differ from each other (p = .988). Together, these measures confirm that the *category-verbal-fluency* split yielded three groups with distinct L2 proficiency. The groups also differed in daily L2 use (F = 10.73, p_{FDR} < .001) and L2 exposure (F = 9.79, p_{FDR} < .001; Table 1). These differences further motivate examining the experience dimensions jointly rather than attributing group differences to proficiency alone.

3.2. Interference suppression performance and associations with L2 proficiency

All participants reached ≥85% accuracy in the practice blocks of both tasks, with no group differences (all p > .24), confirming comparable task comprehension and adequate familiarity with the English color words across proficiency levels. Table 2 reports behavioral performance on the *verbal and spatial Stroop tasks*. Covarying age, sex, years of education, SES and fluid intelligence, by-participant ANCOVAs revealed distinct patterns of proficiency effects across the two domains.

In the *verbal Stroop task*, groups differed on incongruent-condition RT (F = 6.11, p_{FDR} = .011, η_p^2 = .077): the High group responded faster than the Low (p = .003) and Mid (p = .042) groups (Figure 2A), whereas the Mid–Low contrast was not significant (p = .108). Congruent (F = 2.76, p_{FDR} = .106) and neutral (F = 2.08, p_{FDR} = .172) RTs did not differ across groups. The verbal interference effect also differed (F = 7.55, p_{FDR} = .006, η_p^2 = .094), with the High group showing smaller interference (40.59 ± 50.56 ms) than the Low group (p < .001); the verbal interference effect in the Mid group fell between the High and Low groups, approaching but not reaching significance against the Low (p = .058) and High (p = .054) groups.

The *spatial Stroop task* showed a different profile. Incongruent-condition RT again differed across groups (F = 4.48, p_{FDR} = .026, η_p^2 = .058), with the High group being faster than the Low (p = .021) and Mid (p = .023) groups but no Low–Mid difference (p = .998) (Figure 2A); congruent (F = .97, p_{FDR} = .437) and neutral (F = .64, p_{FDR} = .527) RTs were comparable. Spatial interference likewise differed (F = 5.75, p_{FDR} = .011, η_p^2 = .073): the High group showed smaller interference than the Mid (p = .005) and Low (73.71 ± 80.85 ms, p = .022) groups, with no Low–Mid difference (p = .891).

To confirm that these effects were robust to the analytic approach and to model the full single-trial RT distribution, we refitted the data with LMMs on log-transformed trial-level RTs (fixed effects: Group × Condition plus the same covariates; by-participant random intercepts and condition slopes; Satterthwaite degrees of freedom; reference Group = Low, Condition = Neutral; Supplementary Table S3 and Supplementary Figure S3). The trial-level models converged with the by-participant ANCOVAs. Interference (incongruent − neutral, log-RT scale) was significantly smaller in the High than in the Low group in both tasks – verbal: Low−High = .060, 95% *CI* (.021, .099), t (151) = 3.03, p = .008; spatial: Low−High = .067, 95% *CI* (.020, .114), t (151) = 2.82, p = .015 – whereas the Low and Mid groups did not

Table 2. Behavioral performance on verbal and spatial Stroop tasks

Variable	Low L2 (n = 49)	Mid L2 (n = 52)	High L2 (n = 53)	F	p_{FDR}	η_p^2
Verbal Stroop task						
Incongruent (ms)	722.36 ± 165.63	674.79 ± 136.15	615.36 ± 114.06	6.11	.011*	.077
Congruent (ms)	590.66 ± 131.64	629.28 ± 125.46	572.11 ± 112.01	2.76	.106	.036
Neutral (ms)	630.40 ± 142.05	608.61 ± 126.72	574.77 ± 108.07	2.08	.172	.028
Interference effect (ms)	91.96 ± 65.46	66.19 ± 51.89	40.59 ± 50.56	7.55	.006**	.094
Spatial Stroop task						
Incongruent (ms)	702.08 ± 130.94	701.31 ± 111.25	630.81 ± 119.82	4.48	.026*	.058
Congruent (ms)	588.27 ± 121.49	615.47 ± 115.48	611.92 ± 107.74	.97	.437	.013
Neutral (ms)	628.38 ± 127.19	621.42 ± 105.65	593.14 ± 113.99	.64	.527	.009
Interference effect (ms)	73.71 ± 80.85	79.88 ± 60.03	37.67 ± 61.20	5.75	.011*	.073

Note: L2 = second language. All values are presented as mean ± standard deviation. The models controlled for the following covariates: age, sex, years of education, socioeconomic status and fluid intelligence. All p-values were adjusted for multiple comparisons using the false discovery rate (FDR) correction with the Benjamini–Hochberg procedure.
*p_{FDR} < .05. **p_{FDR} < .01. *** p_{FDR} < .001.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

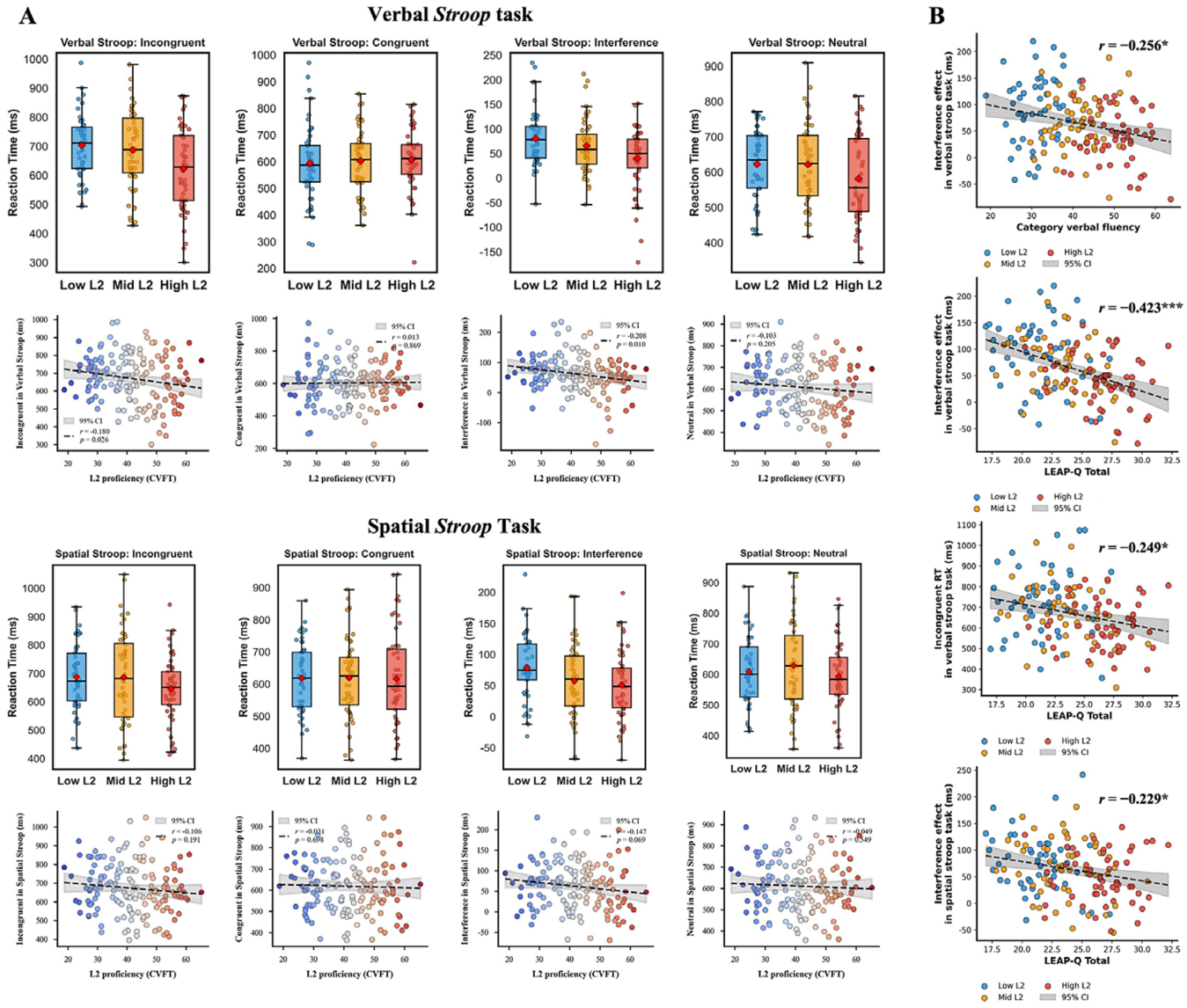


Figure 2. Group differences in *Stroop* task performance and their associations with second language proficiency.
Note: L2 = second language; FDR = false discovery rate. (A) *Stroop* performance and its relation to continuous L2 proficiency, for the verbal (top) and spatial (bottom) tasks. Within each task, the upper portion shows boxplots of condition-wise RTs and interference scores (ms) for the three proficiency groups (Low, Mid, High). The columns show incongruent RT, congruent RT, interference and neutral RT; the lower portion shows scatterplots of continuous proficiency (*CVFT*) against *Stroop* RTs, with regression lines and 95% CIs, by condition (columns) and task (rows). (B) Correlations between the L2 proficiency measures and *Stroop* performance indices. *CVFT = Category Verbal Fluency Test*; LEAP-Q = Language Experience and Proficiency Questionnaire; RT = reaction time. Significance is reported after FDR correction. * $p_{FDR} < .05$. ** $p_{FDR} < .01$. *** $p_{FDR} < .001$.

differ (verbal $p = .616$; spatial $p = .580$; Supplementary Table S4). The interference gradient was monotonic in the verbal task and driven by the High group in the spatial task.

The correlational pattern (Figure 2B) was consistent with this domain asymmetry. *Category verbal fluency* correlated negatively with verbal interference ($r = -.256$, $p_{FDR} = .011$), and the LEAP-Q total score more strongly so ($r = -.423$, $p_{FDR} < .001$), with the latter also predicting incongruent-condition RT ($r = -.249$, $p_{FDR} = .014$). For the spatial task, the LEAP-Q total score again correlated with interference ($r = -.229$, $p_{FDR} = .027$), but more weakly than in the verbal domain. Modeling proficiency continuously reproduced this pattern: within the same mixed-effects framework, continuous *category verbal fluency* predicted smaller verbal interference ($\beta = -.27$, 95% CI [−.42, −.11], t [146] = −3.31, $p_{FDR} = .003$) and smaller, though weaker, spatial interference ($\beta = -.16$, 95% CI [−.31, −.01], t [146] = −2.08, $p_{FDR} = .047$), with the association descriptively stronger in the verbal domain (*CVFT* × Domain interaction: $\beta = .13$, $p_{FDR} = .049$). The domain-dependent behavioral pattern is therefore not an artifact of the categorical split. We report the zero-order correlations as a descriptive complement for comparability with prior work; covariate-adjusted inference for proficiency as a continuous variable is determined using continuous HDDM regression.

3.3. HDDM results

HDDMs were fitted to the trial-level data from both tasks. All parameters had $\hat{R}$ values below 1.01 (Supplementary Table S5), indicating good convergence. Model comparison favored a drift-rate model for the verbal task ($_{M2v}$; *DIC* = 7,659.59) and a decision-

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

threshold model for the spatial task ($_{M2a}$; *DIC* = 7,523.18), each outperforming the alternative single-parameter models (Supplementary Table S6).

Posterior between-group comparisons (Figure 3A, Supplementary Table S7) showed that, in the verbal task, the drift rate (*v*) followed a graded pattern across groups. The High group had higher drift rate than the Low group (Low−High = −.270, 95% HDI [−.391, −.149], *pd* > .999, BF_{10} > 100), with the drift rate of the Mid group falling between and credibly differing from those of the Low (Low−Mid = −.130, *pd* = .995, BF_{10} = 15.20) and High (Mid − High = −.140, *pd* = .999, BF_{10} = 28.50) groups. Decision threshold (*a*) also differed, with the High group adopting a more conservative criterion than the Low group (Low−High = −.270, 95% HDI [−.382, −.166], *pd* > .999, BF_{10} > 100). Nondecision time (*t*) was shorter in the High group than in the Low group (Low − High = .140, 95% HDI [.071, .206], *pd* > .999, BF_{10} > 100), indicating faster perceptual encoding and motor execution. The starting point (*z*) did not differ across groups (all *pd* < .61, BF_{10} < .20), indicating no systematic response bias.

The spatial task showed a contrasting parameter profile (Figure 3B, Supplementary Table S7). Group differences in decision threshold (*a*) and nondecision time (*t*) paralleled those in the verbal task. Drift rate (*v*) also differed between the High and Low groups (Low − High = −.110, 95% HDI [−.203, −.014], *pd* = .982, BF_{10} = 9.20), although the evidence was weaker than that for the decision-threshold difference; no credible drift-rate difference was observed between the Mid group and either of the other groups (all *pd* < .91). Thus, in the spatial domain, the proficiency effect appears to operate mainly through the decision criterion rather than through evidence-accumulation efficiency.

To confirm that this parameter-level dissociation did not depend on the tertile split, we re-estimated the HDDM with *CVFT* entered as a continuous, standardized predictor ($_{CVFTz}$) of each diffusion parameter (Supplementary Table S8). The overall pattern matched the group-based analysis, although effects were milder and more uncertain under this continuous, single-stage modeling. In the verbal task, higher *CVFT* credibly predicted a higher drift rate (β = .24, 95% HDI [.07, .41], *pd* = .997), a shorter nondecision time

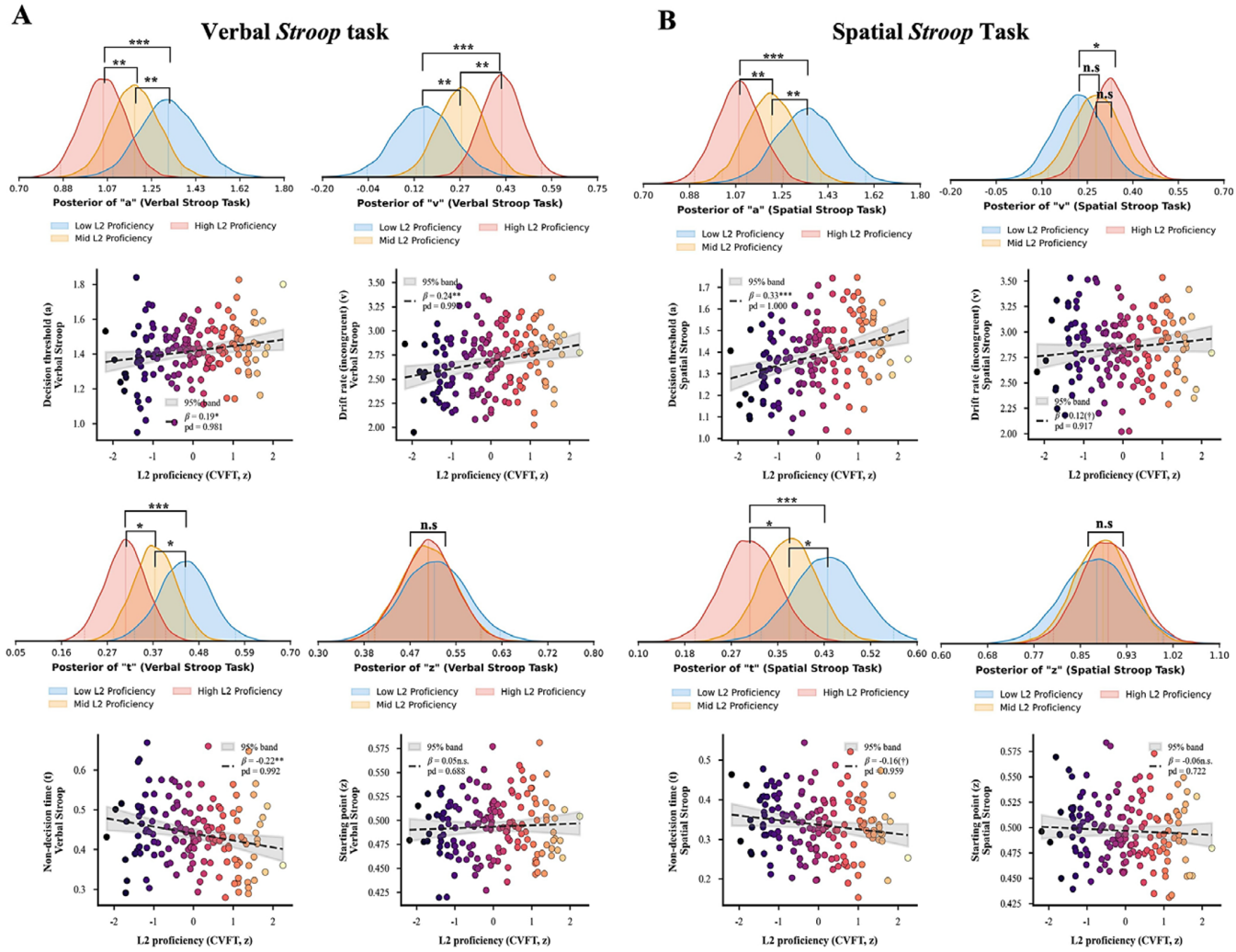


Figure 3. Hierarchical Bayesian drift-diffusion model parameter estimates.

Note: CVFT = English Category Verbal Fluency Test; HDI = highest-density interval. (A) Verbal Stroop task. Posterior distributions show the estimated parameters – drift rate (v), decision threshold (a), nondecision time (t) and starting point (z) – for the three proficiency groups (Low, Mid, High). Accompanying scatterplots display each parameter against standardized CVFT: points are per-participant posterior means, the dashed line is the model-implied regression in the plotted parameter units, and the shaded band is the 95% HDI. Each panel is annotated with the standardized coefficient (β) and the probability of direction (pd). (B) Spatial Stroop task, in the same format as Panel A. The starting point (z) showed no credible group or proficiency effect in either task. * pd ≥ .975; ** pd ≥ .99; *** pd ≥ .999. † .90 ≤ pd < .975 (marginal). n.s. = pd < .90.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

(β = −.22, 95% HDI [−.40, −.04], *pd* = .992) and a higher decision threshold (β = .19, 95% HDI [.01, .37], *pd* = .981); drift rate remained the leading parameter, and the threshold association was credible, although its lower HDI bound was close to zero. In the spatial task, only the decision threshold was robustly credible (β = .33, 95% HDI [.16, .50], *pd* > .999) and dominated the profile. In contrast, the drift rate (β = .12, 95% HDI [−.05, .29], *pd* = .917) and nondecision time (β = −.16, 95% HDI [−.34, .02], *pd* = .959) did not meet the credibility criterion because their HDIs included zero. Therefore, the spatial nondecision-time association was less certain here than in the group comparison. The starting point was unrelated to *CVFT* in both tasks (verbal *pd* = .688; spatial *pd* = .722). Despite the smaller and more uncertain estimates, the core direction was preserved: the verbal advantage was driven primarily by drift rate, the spatial advantage primarily by decision threshold and the starting point was null in both domains.

3.4. Associations with multiple dimensions of bilingual experience

The joint regressions included proficiency, daily L2 use, L2 exposure, age of acquisition and years of learning as standardized predictors of each diffusion parameter. The largest absolute pairwise correlation among the five predictors was .384, between proficiency and daily L2 use. Proficiency was also positively correlated with L2 exposure (*r* = .307). Table 3 reports every coefficient, its 95% HDI and its probability of direction; Supplementary Figure S4 summarizes the estimates. Each coefficient describes an association conditional on the other four dimensions.

In both tasks, proficiency was positively associated with drift rate and decision threshold and negatively associated with nondecision time. All six associations had 95% HDIs excluding zero. Descriptively, the threshold coefficient was larger in the *spatial task* (β = .406) than in the *verbal task* (β = .224), whereas the drift rate coefficients were similar (spatial β = .309; verbal β = .334). The nondecision-time coefficient was more negative in the *spatial task* (β = −.443) than in the *verbal task* (β = −.346).

Daily L2 use was positively associated with verbal drift rate (β = .299) and verbal decision threshold (β = .176). L2 exposure was positively associated with spatial drift rate (β = .196) and spatial decision threshold (β = .188). The remaining use and exposure coefficients, and all coefficients for age of acquisition and years of learning, had 95% HDIs including zero. Starting-point intervals included zero for all five predictors.

The proficiency associations with spatial drift rate and spatial nondecision time had intervals including zero in the *CVFT*-only model but excluding zero in the joint model. The former estimates describe associations without adjustment for the other four language-experience dimensions, whereas the joint estimates describe the corresponding conditional associations.

3.5. Sensitivity analyses

3.5.1. A noncognitive proficiency index

Because *category verbal fluency* itself recruits cognitive control, the proficiency–control association could in principle be partly circular. Rebinning participants by LEAP-Q tertiles – a self-rated index requiring no cognitive-control performance – retained the principal group differences: the High group again showed the least interference, a higher verbal drift rate, higher decision thresholds in both tasks and shorter nondecision times throughout, with no credible spatial drift effect and no starting-point differences (Supplementary Figure S5). Because the LEAP-Q does not engage cognitive control, this convergence indicates that the dissociation is not an artifact of shared demands between the proficiency measure and the *Stroop tasks*.

3.5.2. English reading proficiency

After adjustment for LEAP-Q self-rated English reading, *CVFT* remained associated with verbal drift rate (β = .291, 95% HDI [.124, .458]) and spatial decision threshold (β = .371, 95% HDI [.199, .544]). Nondecision-time associations were also credible in both tasks (verbal β = −.343, 95% HDI [−.507, −.179]; spatial β = −.385, 95% HDI [−.551, −.219]). The intervals for verbal decision threshold (β = .154, 95% HDI [−.032, .340]) and spatial drift rate (β = .124, 95% HDI [−.059, .307]) included zero, as did the starting-point intervals in both tasks (Supplementary Table S9; Supplementary Figure S6).

4. Discussion

The present study contributes to a growing body of DDM research in bilingual cognition (e.g., Hartanto & Yang, 2016; Ong et al., 2017). Its specific contribution is a within-participant DDM decomposition applied to structurally parallel verbal and spatial conflict tasks in a single bilingual sample, which reveals a task-dependent pattern in the proficiency-related decision components. The group-based model comparison favored verbal drift-rate variation and spatial decision-threshold variation. Furthermore, the joint regressions showed broader proficiency associations and additional associations involving daily L2 use and exposure.

4.1. Differential effects of L2 proficiency on verbal and spatial interference suppression

The results reveal the domain-dependent influences of L2 proficiency on interference suppression. In the *verbal Stroop task*, high-proficiency participants showed shorter incongruent-condition RTs than did participants in the Mid and Low groups, but smaller interference effects only relative to the Low group. Correlational analyses converged: *category verbal fluency* correlated negatively with verbal interference, and self-rated LEAP-Q proficiency showed a strong association with verbal performance. These within-task effects align with prior evidence that L1 and L2 proficiency modulate conflict resolution in a *color-word Stroop task* (Tse & Altarriba, 2012); their study, like the present one, used distributional rather than mean-based analyses to determine proficiency effects. L2 proficiency has also been linked to conflict monitoring in a nonverbal *saccadic arrow-Stroop task* under high monitoring demand (Singh & Mishra, 2013), indicating that proficiency effects are not confined to verbal materials; however, such paradigms do not specifically isolate verbal conflict. The spatial task showed a weaker version of the same advantage. High-proficiency participants retained a conflict-processing benefit; however, the between-group effect size was smaller than in the verbal task (η_p^2 = .073 vs. .094), and the LEAP-Q correlations were correspondingly weaker. The spatial difference emerged only between the High group and the Mid and Low groups, with no separation between the latter two, suggesting that the spatial advantage may require a higher proficiency threshold to appear.

The proficiency-only pattern is consistent with the task-dependent modulation view, under which the locus of the proficiency effect tracks each task's control demands rather than reflecting a single uniform mechanism. The proficiency-related gains apply most

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

Table 3. Joint regression of the HDDM parameters on five dimensions of bilingual experience

Outcome (HDDM parameter)	Predictor	β	95% HDI	*pd*
Verbal decision threshold (a_{verbal})	Proficiency (*CVFT*)	**.224**	[.051, .398]	.994
	L2 use	**.176**	[.009, .344]	.980
	L2 exposure	.057	[−.106, .218]	.757
	AoA	.089	[−.066, .244]	.871
	Years of learning	−.071	[−.227, .084]	.815
Verbal drift rate (v_{verbal})	Proficiency (*CVFT*)	**.334**	[.175, .492]	.999
	L2 use	**.299**	[.148, .452]	.999
	L2 exposure	−.006	[−.154, .144]	.527
	AoA	.032	[−.109, .174]	.671
	Years of learning	−.078	[−.220, .063]	.860
Verbal nondecision time (t_{verbal})	Proficiency (*CVFT*)	**−.346**	[−.522, −.172]	.999
	L2 use	.012	[−.154, .180]	.558
	L2 exposure	.060	[−.103, .222]	.765
	AoA	.091	[−.065, .246]	.874
	Years of learning	−.002	[−.158, .156]	.509
Verbal starting point (z_{verbal})	Proficiency (*CVFT*)	.105	[−.080, .289]	.870
	L2 use	.060	[−.115, .238]	.748
	L2 exposure	−.110	[−.282, .060]	.897
	AoA	.107	[−.056, .269]	.901
	Years of learning	−.056	[−.220, .108]	.752
Spatial decision threshold ($a_{spatial}$)	Proficiency (*CVFT*)	**.406**	[.248, .567]	.999
	L2 use	.037	[−.118, .190]	.679
	L2 exposure	**.188**	[.039, .336]	.993
	AoA	.027	[−.116, .170]	.645
	Years of learning	−.069	[−.212, .076]	.827
Spatial drift rate ($v_{spatial}$)	Proficiency (*CVFT*)	**.309**	[.137, .477]	.999
	L2 use	.004	[−.158, .166]	.521
	L2 exposure	**.196**	[.039, .355]	.992
	AoA	.103	[−.047, .254]	.913
	Years of learning	−.071	[−.223, .080]	.822
Spatial nondecision time ($t_{spatial}$)	Proficiency (*CVFT*)	**−.443**	[−.604, −.281]	.999
	L2 use	−.078	[−.233, .077]	.842
	L2 exposure	−.071	[−.220, .079]	.822
	AoA	−.025	[−.169, .118]	.629
	Years of learning	.103	[−.042, .246]	.918
Spatial starting point ($z_{spatial}$)	Proficiency (*CVFT*)	−.052	[−.236, .132]	.713
	L2 use	.067	[−.108, .245]	.775
	L2 exposure	−.098	[−.269, .072]	.869
	AoA	−.116	[−.281, .047]	.919
	Years of learning	−.031	[−.196, .132]	.650

Note: HDDM = hierarchical Bayesian drift-diffusion model; L2 = second language; CVFT = English Category Verbal Fluency Test; AoA = age of acquisition; *HDI* = highest-density interval; *pd* = probability of direction. All five predictors were standardized ($M = 0$, $SD = 1$) and entered simultaneously, so each β is the association of one dimension with the parameter, with the other four held constant. An association is classified as credible when its 95% HDI excludes zero; such coefficients are shown in bold. Standardized coefficients and HDI bounds retain leading zeros (they may exceed 1); *pd* is reported without a leading zero.

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

directly to the verbal task and transfer only partially to the spatial task, given its different stimulus–response mappings. Within the unity-and-diversity framework of executive function (Friedman & Miyake, 2017), this implies that L2 proficiency influences domain-general and domain-specific components: effects on the shared component account for the spatial difference, while additional effects on a verbal-specific component account for the larger verbal effect. These within-task effects are best read against two meta-analytic findings. First, overall bilingual advantages in executive function are small to null and shrink further after correcting for publication bias (Lehtonen et al., 2018), implying that any reliable effect is likely specific rather than global. Second, the magnitude of bilingual effects depends on the task type and participant age (Ware et al., 2020). The present results fit both: proficiency credibly modulated several latent decision parameters within each task, and the modulated parameter profile differed by domain rather than reflecting a global advantage. The computational decomposition clarifies why the verbal effect is the larger one – verbal performance benefits from enhanced accumulation efficiency (drift rate) and a general speed-up (nondecision time), whereas spatial performance benefits from strategic adjustment (decision threshold) and the same general speed-up. This is a graded domain specificity: the benefit is somewhat stronger in language-related tasks and extends partially into nonlinguistic ones.

The adaptive control hypothesis (Green & Abutalebi, 2013) centers on the interactional context, which we did not assess here, and its predictions have not held consistently (Paap et al., 2021). The task-dependent, unity-and-diversity framework adopted here accounts for the domain-general and domain-specific effects. Specific neural mechanisms cannot be determined from these behavioral data. Future work combining diffusion modeling with model-based neuroimaging (Forstmann et al., 2016) could test whether the parameter-level effects reported here correspond to dissociable neural pathways.

4.2. Differential computational mechanisms of L2 proficiency effects

The group-based HDDM comparison identifies different leading parameters across the two domains. Additionally, the joint model examines how proficiency and the other experience dimensions relate to each parameter. In the verbal task, the drift rate was the discriminating parameter, with high-proficiency participants accumulating evidence more efficiently than low-proficiency participants. Drift rate reflects evidence-accumulation efficiency per unit time – the ability to extract task-relevant information while forming a decision (Forstmann et al., 2016). In a *Stroop* context, a higher drift rate on incongruent trials reflects more rapid and accurate identification of the target feature (color) under interference from the distractor (word meaning) (Fennell & Ratcliff, 2019). This gain plausibly reflects several interrelated processes. Attentional selection is a possible contributor, given evidence linking attention to drift rate in perceptual decisions (Nunez et al., 2017). Daily language use may contribute separately to these associations, as indicated by the verbal drift rate association with L2 use in the joint model, and the need to keep two language systems distinct may support clearer, more stable target representations in cognitive tasks generally. We treat the drift rate as the primary verbal locus based on the model comparison, which selected the drift-rate model even though the verbal task also showed a secondary threshold difference.

In the spatial task, the decision threshold was the discriminating parameter, with high-proficiency participants adopting more conservative criteria. Decision threshold reflects the evidence required before responding and, therefore, response caution and the speed–accuracy trade-off (Cochrane et al., 2023). Notably, this effect runs opposite to an automaticity account, which predicts that greater proficiency lowers the threshold; instead, higher proficiency was associated with greater, not reduced, caution – arguing against automaticity in the spatial domain and favoring strategic recruitment of caution. Related bilingual and second-language research has examined metacognition, task switching and rationality (Öztürk & Şenaydın, 2019; Prior & MacWhinney, 2010; Purpuri et al., 2025). However, these studies did not test decision thresholds in a nonlinguistic task, and the processes producing the present threshold difference remain unknown. The weaker and ungraded drift-rate difference in the spatial task does not imply that proficiency leaves spatial accumulation efficiency untouched; rather, adjusting the threshold may simply be the more accessible or more effective route to better spatial performance. This likely reflects task-specific demands: the *verbal Stroop* engages linguistic material closely tied to bilingual experience, and thus shows pronounced efficiency gains, whereas the *spatial Stroop* is further from language experience and elicits greater reliance on strategic optimization.

The group-based model comparison therefore favors verbal drift-rate variation and spatial decision-threshold variation, whereas the joint regressions reveal a broader pattern: proficiency is associated with drift rate, decision threshold and nondecision time in both tasks. The model comparison evaluates candidate patterns of group differences; the joint coefficients estimate proficiency associations conditional on the other dimensions of experience (Supplementary Figure S4).

Unlike the task-specific drift-rate and threshold patterns, nondecision time differed between groups in both tasks. This cross-task consistency points to the domain-general facilitation of basic processing. Nondecision time captures perceptual encoding and motor execution (Forstmann et al., 2016), stages that, although outside the decision proper, shape overall performance; the consistent speed-up may reflect more efficient attentional engagement, consistent with bilingual experience producing adaptive change in higher-level control and lower-level attention systems (Bialystok, 2024).

Because modeling proficiency as a continuous rather than a categorical variable reproduces the same core direction – verbal effects led by drift rate, spatial effects led by the decision threshold and no starting-point effect – these conclusions do not depend on the tertile split. Under continuous modeling, the estimates were smaller and more uncertain: the spatial drift and nondecision-time intervals included zero, and the verbal decision-threshold effect was credible only by a narrow margin. The direction of the dissociation was reproduced, but with greater uncertainty. We retained the grouping for comparability with the substantial group-based literature (Supplementary Table S1) while encouraging future work to model bilingual experience continuously and multidimensionally (de Bruin, 2019; Gullifer et al., 2021). The two models estimate different associations and should be interpreted together: the spatial drift-rate and nondecision-time intervals included zero in the proficiency-only model but not in the joint model.

4.3. What does the DDM reveal beyond aggregate analyses?

Aggregate RT cannot reveal whether an interference effect arises from more efficient evidence accumulation under conflict, a change in response caution, faster perceptual–motor processing or some

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

mixture of these. The DDM separates them into distinct channels, and doing so here clarifies where the two tasks differ and where they behave alike. Behaviorally, higher proficiency credibly reduced interference in both tasks but more strongly in the verbal task (Figure 2). The decomposition shows that these reductions arise through different channels. In the verbal task, higher proficiency raised the drift rate – the evidence-accumulation channel – with a smaller secondary effect on the threshold. In the proficiency-only regression, the spatial task showed a credible decision-threshold association and an uncertain drift-rate association; the joint model showed credible proficiency associations with both parameters. The distinction matters because proficiency-only and joint regressions estimate different associations. Nondecision time differed between groups in both tasks, although the proficiency-only spatial estimate was uncertain.

This decomposition reveals an interpretive consequence that aggregate RT would hide. The group-based comparison highlights evidence accumulation in the verbal task and response caution in the spatial task, whereas the joint model shows how these components relate to proficiency, use and exposure. A smaller behavioral interference effect can therefore accompany multiple decision components and multiple dimensions of bilingual experience. That the parameter profile varies with the demands a task engages accords with a demand-specific account of bilingual adaptation (Bialystok, 2024): proficiency need not confer a uniform control resource expressed identically across domains but may instead shape processing in ways tied to each task's demands. The value of the DDM is that it makes the shape of this demand-specificity visible where aggregate RT leaves it hidden. This adds to recent methodological work examining whether analytic and measurement choices affect the conclusions drawn from the same data – on analysis method (Privitera et al., 2023b), measurement modality (Zhou & Privitera, 2025) and the value of time-sensitive performance measures (Privitera, 2026) – by supplying an instance at the level of within-task latent processes, where two behaviorally distinct effects resolve into partly shared, partly dissociable parameter profiles. The corresponding descriptive mediation models are shown in Supplementary Figure S7 of the Supplementary Materials.

4.4. Implications

By identifying differential mechanisms across domains, the present findings advance the understanding of how L2 proficiency influences interference suppression through multiple pathways whose relative weight varies by domain (Bialystok, 2024). This multipathway, multilevel pattern offers a complementary lens on the inconsistencies in the bilingual-advantage literature, because a single behavioral measure can conflate mechanistically distinct effects. Beyond proficiency, daily L2 use was associated with verbal drift rate and decision threshold, whereas L2 exposure was associated with the corresponding spatial parameters. These associations were estimated with the other four dimensions held constant. Together, the findings link decision processes to proficiency and current language experience. These exploratory associations warrant replication. Age of acquisition and years of learning had intervals including zero for every parameter. Their restricted ranges in this sample leave limited scope for conclusions about learning history across bilingual populations. These results support assessing current experience and acquisition history together, as advocated in multidimensional accounts of bilingualism (de Bruin, 2019; Gullifer et al., 2021).

4.5. Limitations and future considerations

Several limitations qualify these conclusions. First, the sample is homogeneous, which is a double-edged sword. Participants were English-major undergraduates aged 18–22 with similar educational backgrounds and largely classroom-based L2 experience, which limits external validity. The same homogeneity, however, is how this individual-differences approach controls the cultural and educational confounds that complicate bilingual–monolingual comparisons (Valian, 2015); the trade-off is that restricted ranges in age of acquisition and years of L2 learning constrain our ability to detect associations with learning history. Intervals including zero for these dimensions do not establish that acquisition history is less important than proficiency or current experience. Two further dimensions that recent frameworks emphasize – code-switching frequency and interactional context (Green & Abutalebi, 2013; Semenova et al., 2026) – were not assessed; the multidimensional analysis reported above therefore extends beyond proficiency but does not exhaust bilingual experience. Future work should test more diverse bilingual populations and examine how sample characteristics, language pairings and typological distance moderate the proficiency–control relationship.

Second, the cross-sectional design precludes causal inference. Concurrent measurement cannot determine whether stronger L2 proficiency sharpens interference suppression, or whether more efficient interference suppression instead aids L2 acquisition; longitudinal and training designs with pre- and post-DDM assessments would be needed to establish directionality.

Third, the proficiency index itself raises a circularity concern because *category verbal fluency* recruits cognitive control (Shao et al., 2014). A sensitivity analysis that regrouped participants by the self-rated LEAP-Q total score – which requires no cognitive-control performance – retained the verbal drift rate and spatial decision-threshold differences, indicating that the effects are not merely an artifact of the index. However, the present measures cannot fully separate proficiency from executive control. Future research should index L2 proficiency with measures that minimally depend on executive control (e.g., receptive vocabulary or grammaticality judgment) and add an assessment of speaking proficiency that fluency tasks and standardized written tests such as the Gaokao do not capture.

Fourth, although controlling for self-rated English reading proficiency left the key effects credible (a higher drift rate in the verbal task and a higher decision threshold in the spatial task), our proficiency index and the *verbal Stroop* rely on English; therefore, a contribution of English lexical access to the verbal-task effect cannot be fully excluded. Reassuringly, the parallel effect of proficiency in the nonlinguistic spatial task and the domain-general nondecision-time effect indicate that the findings are not reducible to English lexical access. Future work could pair a proficiency measure independent of English lexical production (e.g., a standardized comprehension or grammar test) with an objective index of lexical access (e.g., vocabulary size or picture-naming latency) to separate linguistic proficiency from cognitive control more cleanly.

Fifth, the *spatial Stroop* contained no linguistic stimuli, yet proficiency effects emerged on its DDM parameters. We cannot exclude a contribution from endogenous verbal recoding (Lupyan, 2012) – a pathway independent of L2 stimulus presence, which therefore does not undercut the spatial-task evidence against a stimulus-driven lexical-access account. A complementary design – an L1 (Chinese-character) *verbal Stroop* together with articulatory suppression during the spatial task in an independent sample –

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S1366728926101965

would dissociate stimulus-driven lexical access, internal verbal recoding and interference suppression; replicating this should be prioritized.

Finally, the findings are specific to interference suppression, indexed by *a Stroop-type* conflict between task-relevant and task-irrelevant stimulus dimensions. Cognitive control comprises, at minimum, inhibition (including interference suppression), working-memory updating and task-set switching (Friedman & Miyake, 2017), each potentially related to bilingual experience in its own way. Both tasks here index stimulus–stimulus conflict; stimulus–response conflict tasks such as the *Simon task*, where competition arises at response selection, engage partly distinct control processes (Hommel, 2011; Kornblum et al., 1990) and have shown inconsistent bilingual effects (Hilchey & Klein, 2011). Whether the parameter profiles observed here generalize to other control components and to stimulus–response conflict is an open question that applying the same DDM decomposition across a battery of tasks within a single sample would resolve.

5. Conclusions

Using HDDM, in this study, we examined the computational mechanisms underlying L2 proficiency effects on domain-specific and domain-general interference suppression. L2 proficiency was associated with enhanced interference suppression in the verbal and spatial domains. The association was stronger in the verbal domain and showed a task-dependent computational pattern, with group-based model comparison favoring verbal drift-rate variation and spatial decision-threshold variation. Joint regressions revealed broader proficiency associations and additional parameter associations with daily L2 use and exposure, linking decision processes to proficiency and current language experience. These results characterize the bilingual influence on interference suppression as domain-specific and domain-general, with its expression and underlying mechanism varying by domain, and they show how a computational decomposition can recover the mechanistic structure that aggregate measures leave hidden. Future studies should test these patterns in more diverse samples, use longitudinal and training designs to determine causal direction and model bilingual experience as the multidimensional construct it is.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926101965.

Data availability statement. The data that support the findings of this study are openly available on OSF at https://osf.io/p6udh/.

Acknowledgments. The authors thank the students from Weinan Normal University for their participation.

Competing interests. The authors declare that they have no competing interests.

References

Abutalebi, J., & **Green, D.** (2007). Bilingual language production: The neurocognition of language representation and control. *Journal of Neurolinguistics*, **20**(3), 242–275. https://doi.org/10.1016/j.jneuroling.2006.10.003

Ambrosi, S., **Servant, M.**, **Blaye, A.**, & **Burle, B.** (2019). Conflict processing in kindergarten children: New evidence from distribution analyses reveals the dynamics of incorrect response activation and suppression. *Journal of Experimental Child Psychology*, **177**, 36–52. https://doi.org/10.1016/j.jecp.2018.06.006

Ardila, A. (2020). A cross-linguistic comparison of category verbal fluency test (ANIMALS): A systematic review. *Archives of Clinical Neuropsychology*, **35**(2), 213–225. https://doi.org/10.1093/arclin/acz060

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 (Methodological)*, **57**(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bialystok, E. (2024). Bilingualism modifies cognition through adaptation, not transfer. *Trends in Cognitive Sciences*, **28**(11), 987–997. https://doi.org/10.1016/j.tics.2024.07.012

Bialystok, E., & **DePape, A.-M.** (2009). Musical expertise, bilingualism, and executive functioning. *Journal of Experimental Psychology: Human Perception and Performance*, **35**(2), 565–574. https://doi.org/10.1037/a0012735

Boyce, W., **Torsheim, T.**, **Currie, C.**, & **Zambon, A.** (2006). The family affluence scale as a measure of national wealth: Validation of an adolescent self-report measure. *Social Indicators Research*, **78**(3), 473–487. https://doi.org/10.1007/s11205-005-1607-6

Cheng, Y., & **Hamid, M. O.** (2025). Social impact of Gaokao in China: A critical review of research. *Language Testing in Asia*, **15**(1), Article 22. https://doi.org/10.1186/s40468-025-00355-y

Cochrane, A., **Sims, C. R.**, **Bejjanki, V. R.**, **Green, C. S.**, & **Bavelier, D.** (2023). Multiple timescales of learning indicated by changes in evidence-accumulation processes during perceptual decision-making. *npj Science of Learning*, **8**(1), Article 19. https://doi.org/10.1038/s41539-023-00168-9

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Dash, T., **Joanette, Y.**, & **Ansaldo, A. I.** (2022). Multifactorial approaches to study bilingualism in the aging population: Past, present, future. *Frontiers in Psychology*, **13**, Article 917959. https://doi.org/10.3389/fpsyg.2022.917959

de Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, **9**(3), Article 33. https://doi.org/10.3390/bs9030033

DeKeyser, R. M. (Ed.). (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511667275

DeLuca, V., **Rothman, J.**, **Bialystok, E.**, & **Pliatsikas, C.** (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proceedings of the National Academy of Sciences*, **116**(15), 7565–7574. https://doi.org/10.1073/pnas.1811513116

Faul, F., **Erdfelder, E.**, **Lang, A.-G.**, & **Buchner, A.** (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, **39**(2), 175–191. https://doi.org/10.3758/BF03193146

Fennell, A., & **Ratcliff, R.** (2019). Does response modality influence conflict? Modelling vocal and manual response Stroop interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **45**(11), 2098–2119. https://doi.org/10.1037/xlm0000689

Forstmann, B. U., **Ratcliff, R.**, & **Wagenmakers, E.-J.** (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, **67**(1), 641–666. https://doi.org/10.1146/annurev-psych-122414-033645

Friedman, N. P., & **Miyake, A.** (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, **133**(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

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

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, **1**(2), 67–81. https://doi.org/10.1017/S1366728998000133

Green, D. W., & **Abutalebi, J.** (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, **25**(5), 515–530. https://doi.org/10.1080/20445911.2013.796377

Grundy, J. G. (2020). The effects of bilingualism on executive functions: An updated quantitative analysis. *Journal of Cultural Cognitive Science*, **4**(2), 177–199. https://doi.org/10.1007/s41809-020-00062-5

Gullifer, J. W., **Kousaie, S.**, **Gilbert, A. C.**, **Grant, A.**, **Giroud, N.**, **Coulter, K.**, **Klein, D.**, **Baum, S.**, **Phillips, N.**, & **Titone, D.** (2021). Bilingual language experience as a multidimensional spectrum: Associations with objective and subjective language proficiency. *Applied Psycholinguistics*, **42**(2), 245–278. https://doi.org/10.1017/S0142716420000521

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101965

Han, X., **Li, W.**, & **Filippi, R.** (2025). Modulating bilingual language production and cognitive control: How bilingual language experience matters. *Bilingualism: Language and Cognition*, **28**(1), 120–134. https://doi.org/10.1017/S1366728924000191

Hartanto, A., & **Yang, H.** (2016). Disparate bilingual experiences modulate task-switching advantages: A diffusion-model analysis of the effects of interactional context on switch costs. *Cognition*, **150**, 10–19. https://doi.org/10.1016/j.cognition.2016.01.016

Hilchey, M. D., & **Klein, R. M.** (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review*, **18**(4), 625–658. https://doi.org/10.3758/s13423-011-0116-7

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, **136**(2), 189–202. https://doi.org/10.1016/j.actpsy.2010.04.011

Jensen, A. R., & **Rohwer, W. D.** (1966). The Stroop color-word test: A review. *Acta Psychologica*, **25**(1), 36–93. https://doi.org/10.1016/0001-6918(66)90004-7

Kornblum, S., **Hasbroucq, T.**, & **Osman, A.** (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility—A model and taxonomy. *Psychological Review*, **97**(2), 253–270. https://doi.org/10.1037/0033-295x.97.2.253

Lehtonen, M., **Soveri, A.**, **Laine, A.**, **Järvenpää, J.**, **de Bruin, A.**, & **Antfolk, J.** (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, **144**(4), 394–425. https://doi.org/10.1037/bul0000142

Löffler, C., **Frischkorn, G. T.**, **Hagemann, D.**, **Sadus, K.**, & **Schubert, A.-L.** (2024). The common factor of executive functions measures nothing but speed of information uptake. *Psychological Research*, **88**(4), 1092–1114. https://doi.org/10.1007/s00426-023-01924-7

Luk, G., & **Bialystok, E.** (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, **25**(5), 605–621. https://doi.org/10.1080/20445911.2013.795574

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, **3**, Article 54. https://doi.org/10.3389/fpsyg.2012.00054

Makowski, D., **Ben-Shachar, M. S.**, **Chen, S. H. A.**, & **Lüdecke, D.** (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, **10**, Article 2767. https://doi.org/10.3389/fpsyg.2019.02767

Marian, V., **Blumenfeld, H. K.**, & **Kaushanskaya, M.** (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, **50**(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

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. https://doi.org/10.1006/cogp.1999.0734

Nunez, M. D., **Vandekerckhove, J.**, & **Srinivasan, R.** (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology*, **76**(Part B), 117–130. https://doi.org/10.1016/j.jmp.2016.03.003

Ong, G., **McKague, M.**, **Weekes, B.**, & **Sewell, D. K.** (2019). Diffusing the bilingual lexicon: Task-based and lexical components of language switch costs. *Cognitive Psychology*, **114**, Article 101225. https://doi.org/10.1016/j.cogpsych.2019.101225

Ong, G., **Sewell, D. K.**, **Weekes, B.**, **McKague, M.**, & **Abutalebi, J.** (2017). A diffusion model approach to analysing the bilingual advantage for the Flanker task: The role of attentional control processes. *Journal of Neurolinguistics*, **43**(Part A), 28–38. https://doi.org/10.1016/j.jneuroling.2016.08.002

Öztürk, N., & **Şenaydın, F.** (2019). Dichotomy of EFL reading: Metacognition vs. proficiency. *Journal of Language and Linguistic Studies*, **15**(2), 605–617. https://doi.org/10.17263/jlls.586757

Paap, K. R., **Johnson, H. A.**, & **Sawi, O.** (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*, **69**, 265–278. https://doi.org/10.1016/j.cortex.2015.04.014

Paap, K. R., **Mason, L.**, & **Anders-Jefferson, R.** (2021). Predictions about the cognitive consequences of language switching on executive functioning inspired by the adaptive control hypothesis fail more often than not. *Brain Sciences*, **11**(9), Article 1217. https://doi.org/10.3390/brainsci11091217

Pan, W., **Geng, H.**, **Zhang, L.**, **Fengler, A.**, **Frank, M. J.**, **Zhang, R.-Y.**, & **Chuan-Peng, H.** (2025). dockerHDDM: A user-friendly environment for Bayesian hierarchical drift-diffusion modeling. *Advances in Methods and Practices in Psychological Science*, **8**(1), Article 25152459241298700. https://doi.org/10.1177/25152459241298700

Prior, A., & **MacWhinney, B.** (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, **13**(2), 253–262. https://doi.org/10.1017/s1366728909990526

Privitera, A. J. (2026). Bilingual effects on cognitive control: Are we looking in the right place? *Bilingualism: Language and Cognition*, **29**(4), 1004–1017. https://doi.org/10.1017/s1366728925100576

Privitera, A. J., **Momenian, M.**, & **Weekes, B.** (2023a). Graded bilingual effects on attentional network function in Chinese high school students. *Bilingualism: Language and Cognition*, **26**(3), 527–537. https://doi.org/10.1017/s1366728922000803

Privitera, A. J., **Momenian, M.**, & **Weekes, B. S.** (2023b). Modeling the bilingual advantage: Do results differ between analysis methods? *Ampersand*, **11**, Article 100134. https://doi.org/10.1016/j.amper.2023.100134

Privitera, A. J., & **Weekes, B.** (2023). Scoping review of research practices in the investigation of bilingual effects on inhibition and attention in young people. *International Journal of Bilingualism*, **27**(5), 749–775. https://doi.org/10.1177/13670069221121498

Purpuri, S., **Vasta, N.**, **Filippi, R.**, **Treccani, B.**, **Wei, L.**, & **Mulatti, C.** (2025). Beyond the foreign language effect: Unravelling the impact of L2 proficiency on rationality. *Bilingualism: Language and Cognition*, **28**(3), 740–747. https://doi.org/10.1017/s1366728924000750

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

Ratcliff, R., **Smith, P. L.**, **Brown, S. D.**, & **McKoon, G.** (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, **20**(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

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

Rodríguez-Pujadas, A., **Sanjuán, A.**, **Ventura-Campos, N.**, **Román, P.**, **Martin, C.**, **Barceló, F.**, **Costa, A.**, & **Ávila, C.** (2013). Bilinguals use language-control brain areas more than monolinguals to perform non-linguistic switching tasks. *PLoS One*, **8**(9), Article e73028. https://doi.org/10.1371/journal.pone.0073028

Rothman, J., **Bayram, F.**, **DeLuca, V.**, **Di Pisa, G.**, **Duñabeitia, J. A.**, **Gharibi, K.**, **Hao, J.**, **Kolb, N.**, **Kubota, M.**, **Kupisch, T.**, **Laméris, T.**, **Luque, A.**, **van Osch, B.**, **Pereira Soares, S. M.**, **Prystauka, Y.**, **Tat, D.**, **Tomić, A.**, **Voits, T.**, & **Wulff, S.** (2023). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics*, **44**(3), 316–329. https://doi.org/10.1017/s0142716422000315

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

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357

Semenova, E., **Norkina, M.**, **Logvinenko, T.**, **Ryseva, K.**, **Chinn, L. K.**, **Crabb, K.**, & **Grigorenko, E. L.** (2026). The modulating role of interactional contexts in executive functioning of bilinguals: A scoping review. *Bilingualism: Language and Cognition*, **29**(2), 527–561. https://doi.org/10.1017/s1366728925100084

Shao, Z., **Janse, E.**, **Visser, K.**, & **Meyer, A. S.** (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, **5**, Article 772. https://doi.org/10.3389/fpsyg.2014.00772

Singh, N., & **Mishra, R. K.** (2013). Second language proficiency modulates conflict-monitoring in an oculomotor Stroop task: Evidence from Hindi-English bilinguals. *Frontiers in Psychology*, **4**, Article 322. https://doi.org/10.3389/fpsyg.2013.00322

Soares, S. M. P., **Ong, G.**, **Abutalebi, J.**, **Del Maschio, N.**, **Sewell, D.**, & **Weekes, B.** (2019). A diffusion model approach to analyzing performance on the Flanker task: The role of the DLPFC. *Bilingualism: Language and Cognition*, **22**(5), 1194–1208. https://doi.org/10.1017/s1366728918000974

Titone, D. A., & **Tiv, M.** (2023). Rethinking multilingual experience through a Systems Framework of Bilingualism. *Bilingualism: Language and Cognition*, **26**(1), 1–16. https://doi.org/10.1017/s1366728921001127

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101965

Tiv, M., **Kutlu, E.**, **Gullifer, J. W.**, **Feng, R. Y.**, **Doucerain, M. M.**, & **Titone, D. A.** (2022). Bridging interpersonal and ecological dynamics of cognition through a systems framework of bilingualism. *Journal of Experimental Psychology: General*, **151**(9), 2128–2143. https://doi.org/10.1037/xge0001174

Tse, C.-S., & **Altarriba, J.** (2012). The effects of first- and second-language proficiency on conflict resolution and goal maintenance in bilinguals: Evidence from reaction time distributional analyses in a Stroop task. *Bilingualism: Language and Cognition*, **15**(3), 663–676. https://doi.org/10.1017/s1366728912000077

Valian, V. (2015). Bilingualism and cognition: A focus on mechanisms. *Bilingualism: Language and Cognition*, **18**(1), 47–50. https://doi.org/10.1017/s1366728914000698

van den Noort, M., **Struys, E.**, **Bosch, P.**, **Jaswetz, L.**, **Perriard, B.**, **Yeo, S.**, **Barisch, P.**, **Vermeire, K.**, **Lee, S.-H.**, & **Lim, S.** (2019). Does the bilingual advantage in cognitive control exist and if so, what are its modulating factors? A systematic review. *Behavioral Sciences*, **9**(3), Article 27. https://doi.org/10.3390/bs9030027

Vats, D., & **Knudson, C.** (2021). Revisiting the Gelman–Rubin diagnostic. *Statistical Science*, **36**(4), 518–529. https://doi.org/10.1214/20-sts812

Wang, R., **Fan, X.**, **Liu, C.**, & **Cai, Z. G.** (2016). Cognitive control and word recognition speed influence the Stroop effect in bilinguals. *International Journal of Psychology*, **51**(2), 93–101. https://doi.org/10.1002/ijop.12115

Ware, A. T., **Kirkovski, M.**, & **Lum, J. A. G.** (2020). Meta-analysis reveals a bilingual advantage that is dependent on task and age. *Frontiers in Psychology*, **11**, Article 1458. https://doi.org/10.3389/fpsyg.2020.01458

Wiecki, T. V., **Sofer, I.**, & **Frank, M. J.** (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, **7**, Article 14. https://doi.org/10.3389/fninf.2013.00014

Xia, L., **Sorace, A.**, **Vega-Mendoza, M.**, **Deng, X.**, & **Bak, T. H.** (2026). The effect of language proficiency, usage, and exposure on cognitive control: A study in early adulthood Chinese learners of English. *International Journal of Bilingualism*, **30**(1), 249–266. https://doi.org/10.1177/13670069241307606

Xie, Z. (2018). The influence of second language (L2) proficiency on cognitive control among young adult unbalanced Chinese-English bilinguals. *Frontiers in Psychology*, **9**, Article 412. https://doi.org/10.3389/fpsyg.2018.00412

Xie, Z., & **Chen, M.** (2026). The influence of L2 proficiency and interpreting experience on cognitive control among Chinese–English bilinguals. *International Journal of Bilingualism*, **30**(4), 1038–1062. https://doi.org/10.1177/13670069251358703

Yow, W. Q., & **Li, X.** (2015). Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: Why variations in bilingual experiences matter. *Frontiers in Psychology*, **6**, Article 164. https://doi.org/10.3389/fpsyg.2015.00164

Zhao, X., **Wang, Y.**, & **Maes, J. H. R.** (2022). The effect of working memory capacity and training on intertemporal decision making in children from low-socioeconomic-status families. *Journal of Experimental Child Psychology*, **216**, Article 105347. https://doi.org/10.1016/j.jecp.2021.105347

Zhou, Y., & **Privitera, A. J.** (2025). Subjective versus objective language proficiency measures in the investigation of bilingual effects on cognitive control. *International Journal of Bilingualism*, **29**(3), 549–565. https://doi.org/10.1177/13670069241229393

Downloaded from https://www.cambridge.org/core. IP address: 64.118.141.134, on 17 Oct 2026 at 09:08:03, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101965