Bilingualism: Language and Cognition

cambridge.org/bil

A meta-analysis of Stroop task studies on bilinguals' emotional word processing

Boubaker Gar Elmetred and Sumin Zhang

School of Foreign Studies, Zhejiang Gongshang University, China

Research Article

Cite this article: Gar Elmetred, B. and Zhang, S. (2026). A meta-analysis of Stroop task studies on bilinguals' emotional word processing. *Bilingualism: Language and Cognition* 1–13. https://doi.org/10.1017/S1366728926101795

Received: 20 January 2026 UTC
Revised: 4 July 2026 UTC
Accepted: 27 July 2026 UTC

Keywords:
emotional Stroop; bilingualism; meta-analysis; age of acquisition; emotional word processing

Corresponding author:
Sumin Zhang;
Email: suminzhang@zjgsu.edu.cn

Abstract

This meta-analysis examines emotional Stroop interference in bilinguals, a literature characterised with inconsistent findings on age of acquisition (AoA), emotional valence, and L2 proficiency. Following PRISMA 2020 guidelines, a systematic search across six electronic databases identified 10 studies (12 independent effect sizes, N = 450). Random-effects meta-analysis yielded a non-significant pooled effect characterized by extreme heterogeneity. Mixed-effects moderator models identified age of L2 acquisition as the critical boundary condition: early-acquired bilinguals exhibited reliable interference, whereas late-acquired bilinguals did not. Proficiency assessment method and valence eligibility were non-significant, suggesting that superficial methodological differences do not account for the literature's variability. Formal stimulus-type analysis was precluded by a near-absence of emotion-label studies. These findings indicate that emotional Stroop interference in bilinguals is conditional on neurodevelopmental timing, supporting neurodevelopmental models of integrated language–emotion networks. The results underscore the need for standardised paradigms and more diverse bilingual samples in future primary research.

1. Introduction

Processing emotional meaning while managing competing linguistic systems places unique demands on the bilingual cognitive architecture. Bilinguals must negotiate two languages that may differ not only in proficiency and usage but also in the depth and automaticity of their emotional associations (Dewaele, 2010; Pavlenko, 2012). Behavioural and neuroimaging studies indicate that the first language typically carries stronger affective resonance than the second language, with emotional words in L2 engaging less automatic limbic activation (Caldwell-Harris, 2015; Pavlenko, 2012). The emotional Stroop paradigm provides a standardised platform for investigating this intersection, isolating attentional capture by affectively salient words during colour identification (Phaf & Kan, 2007; Williams et al., 1996). Yet its application to bilingual populations has yielded highly variable findings, with some studies reporting robust interference and others finding null or even reversed effects (Eilola et al., 2007; Sutton et al., 2007; Winskel, 2013).

This heterogeneity raises a critical question: under what conditions does bilingual emotional Stroop interference emerge reliably? Prior reviews have identified theoretical and methodological sources of variability, including inconsistent operationalisation of moderators and narrow population sampling (Kazanas et al., 2019; Sharif & Mahmood, 2023), but no quantitative synthesis has tested whether boundary conditions such as neurodevelopmental timing, valence structure or proficiency operationalisation partition the observed variance. The present meta-analysis addresses this gap by synthesizing the available literature and examining whether age of L2 acquisition, design-level valence inclusion and assessment method account for the disparate findings. In doing so, it aims to move the field beyond the binary question of whether bilinguals experience emotional interference towards a conditional understanding of when and for whom such effects are robust.

1.1. Bilingualism and cognitive control

A bilingual is an individual who uses two or more languages in daily life, regardless of proficiency level or age of acquisition (Grosjean, 2010). This definition encompasses a heterogeneous population ranging from simultaneous childhood bilinguals to late adult learners, yet all share a common cognitive characteristic: the necessity of managing multiple language systems that remain simultaneously activated during communication (Bialystok, 2011; Kroll & Bialystok, 2013). This continuous management places persistent demands on executive functions – particularly inhibitory control, conflict monitoring and attentional disengagement – because the bilingual must select the target language while suppressing the non-target alternative (Bialystok et al., 2004). Early theoretical work proposed that this routine language control confers a general cognitive advantage, with bilinguals outperforming monolinguals on nonverbal

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

executive tasks such as the Simon and Flanker paradigms (Bialystok et al., 2004). However, subsequent meta-analytic evidence has been equivocal: some syntheses report small but reliable advantages (Grundy, 2020; Yurtsever et al., 2024), while others find negligible or null effects after controlling for publication bias, demographic confounds and task demands (Donnelly et al., 2019; Lehtonen et al., 2018; Nichols et al., 2020).

The field has therefore moved beyond the binary question of whether bilingualism enhances cognition and towards identifying the specific conditions under which such effects emerge (Bialystok, 2024). This shift reflects a broader recognition that bilingualism is not a monolithic experience but a multidimensional construct involving age of acquisition, proficiency balance, language-switching frequency and sociolinguistic context (Yurtsever et al., 2024). Recent theoretical developments support this conditional view. Bialystok (2024) proposed an adaptation theory arguing that bilingual experience modifies cognition not by transferring language-control skills to domain-general executive function, but by adapting the underlying attention system itself. According to this framework, bilinguals develop a more efficient attentional deployment system optimized for rapid reconfiguration between competing stimulus dimensions – a capacity directly tested when emotionally salient stimuli compete with goal-directed behaviour.

Recent evidence supports this conditional view using sensitive process measures. Yurtsever et al. (2024) demonstrated that L2 age of acquisition and language-switching frequency predict distinct executive function trajectories in young adults, with earlier acquisition and higher switching associated with more adaptive conflict monitoring. These findings suggest that conventional behavioural measures may underestimate bilingual–monolingual differences because they capture only endpoint performance rather than the dynamic trajectory of cognitive control (Yurtsever et al., 2024). The implication is that bilingual cognitive consequences are emergent properties of ongoing experience, not fixed traits, and that the field requires more sensitive methodologies to detect them.

These adaptive dynamics extend naturally into emotional language processing. Emotional words engage both limbic and language networks, so bilinguals must negotiate affective salience across two systems that may differ in the depth and automaticity of their emotional associations (Dewaele, 2010; Pavlenko, 2012). Because emotional stimuli inherently capture attention, they provide an ideal test case for examining whether bilingual attentional adaptations generalise to affectively charged contexts. Systematic reviews confirm that bilingual emotional processing constitutes a distinct research domain, with neuroimaging and behavioural studies revealing divergent L1–L2 reactivity patterns (Sharif & Mahmood, 2023). However, these reviews also note that the field employs diverse paradigms, stimulus types, and population definitions that resist straightforward synthesis (Sharif & Mahmood, 2023), raising the question of how executive control mechanisms identified in non-emotional tasks extend to emotionally salient word processing.

1.2. Emotional word processing in bilinguals

Emotional words are processed differently in a second language. Bilinguals consistently report feeling more spontaneous and emotionally connected when using their L1, a phenomenon attributed to deeper encoding of emotional concepts during childhood acquisition (Dewaele, 2008; Pavlenko, 2012). Behavioural evidence complicates this straightforward picture: some studies find reduced emotional effects in L2 compared to L1 (Anooshian & Hertel, 1994; Degner et al., 2012), while others report equivalent or stronger L2 resonance among highly proficient bilinguals (Caldwell-Harris et al., 2011; Ferré et al., 2010). These inconsistencies likely reflect the diversity of paradigms employed – lexical decision, word recall, affective priming, and Stroop interference – each tapping different levels of automatic versus controlled processing (Kazanas et al., 2019).

A further source of complexity concerns stimulus type. Emotion-label words explicitly name emotional states (e.g., "fear," "joy"), whereas emotion-laden words carry emotional connotation without naming an emotion (e.g., "war," "cancer," "holiday"). Constructionist accounts propose that language provides the conceptual categories through which emotional experiences are constructed, implying that emotion-label words should engage conceptual affective systems more directly than emotion-laden words (Lindquist, 2017). Recent evidence from Chinese–English bilinguals supports this distinction: Gu and Chen (2024) found that emotion-label words produced faster reaction times and higher accuracy than emotion-laden words in both languages, with negative emotion-laden words showing a disembodiment effect in L2 (i.e., no processing advantage over neutral words). This stimulus-type imbalance – where researchers have overwhelmingly favoured emotion-laden materials – may have obscured genuine differences in conceptual affective engagement between the two stimulus classes (Gu & Chen, 2024).

Neuroimaging studies provide converging evidence for differential L1–L2 emotional processing. Del Maschio et al. (2024) conducted an fMRI emotional Stroop experiment with proficient Italian–English bilinguals and found stronger L1 than L2 activations in the posteromedial cortex, a region implicated in attention modulation and affective processing. Critically, these neural differences emerged without detectable behavioural effects, suggesting that behavioural and hemodynamic responses exhibit different sensitivity profiles (Del Maschio et al., 2024). This dissociation underscores the value of behavioural meta-analytic synthesis: if neural measures detect L1–L2 differences that behavioural measures miss, then moderator analyses may reveal hidden structure in the behavioural literature.

1.3. The emotional Stroop paradigm

The Stroop colour-word paradigm, introduced by Stroop (1935), remains the most robust tool for investigating attentional control. In its canonical form, participants name the ink colour of printed words while ignoring semantic content; when word and colour conflict, response times increase, revealing the automaticity of reading and the effort required for suppression (MacLeod, 1991). The emotional Stroop task adapts this logic by replacing colour words with emotionally valenced words (positive, negative or neutral) and measuring delay in colour identification for emotional relative to neutral stimuli (Williams et al., 1996). Unlike the classic Stroop, where interference arises from semantic competition between two colour representations, emotional Stroop interference reflects attentional capture by affectively salient stimuli that recruit limbic and executive systems (Williams et al., 1996).

Clinical research established that anxious patients show exaggerated interference for threat-relevant words, supporting the interpretation that the effect indexes prioritized threat processing (Williams et al., 1996). A meta-analysis of 70 emotional Stroop studies confirmed that the effect is driven by slow disengagement from threat-related stimuli rather than by fast, automatic attentional capture, with the largest effects observed under blocked presentation conditions (Phaf & Kan, 2007). In non-clinical

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

populations, negative words typically produce greater interference than positive or neutral words, consistent with evolutionary threat-detection accounts (Citron et al., 2020; Harris et al., 2003). Recent evidence further indicates that negative and high-arousal words interact to amplify interference, suggesting that stimulus properties modulate the effect beyond simple valence categorization (Martynova & Lyusin, 2026).

For bilingual research, the critical adaptation is the within-subject L1–L2 comparison, which isolates the cross-linguistic affective gap (Eilola et al., 2007; Sutton et al., 2007). To explain this variability, three theoretical frameworks have been invoked to predict when and why emotional interference should be stronger or weaker across different bilingual populations.

1.4. Theoretical frameworks for bilingual emotional interference

Three frameworks inform predictions about bilingual emotional Stroop effects. The Adaptive Control Hypothesis (ACH) posits that bilingual language-use contexts – single-language, dual-language, and dense code-switching – impose distinct demands on cognitive control, including goal maintenance, conflict monitoring and interference suppression (Green & Abutalebi, 2013). In dense code-switching contexts, demands shift from competitive inhibition to cooperative integration, reducing the need for interference control while increasing cross-linguistic activation (Green, 2018). Applied to emotional processing, the ACH predicts that balanced bilinguals should exhibit more integrated affective-linguistic networks and greater susceptibility to emotional distraction (Green & Abutalebi, 2013).

Neurodevelopmental models challenge the ACH by emphasizing the timing of L2 acquisition. The critical-period hypothesis suggests that childhood L2 acquisition produces more native-like neural and behavioural profiles, including automatized emotional associations in L2 (Grundy, 2020; Jafari et al., 2022). Early bilinguals are hypothesized to develop overlapping emotional representations across languages, whereas late bilinguals retain more segregated mappings (Jafari et al., 2022). This predicts stronger emotional Stroop interference in early-acquired bilinguals, because integrated networks render emotional words more attentionally salient in either language (Grundy, 2020).

Proficiency-based models introduce a temporal dimension distinct from both the ACH and neurodevelopmental accounts. As L2 proficiency increases, emotional words acquire richer affective representations, narrowing the L1–L2 emotional gap (Weimer et al., 2022; Zhang & Fan, 2023). However, this relationship may be non-linear: intermediate learners may show little L2 interference due to weak affective associations, whereas highly proficient bilinguals may approach L1 levels (Weimer et al., 2022). Measurement complexity arises because objective tests and self-report questionnaires yield potentially divergent conclusions about true L2 competence (Zhang & Fan, 2023). These three frameworks generate competing predictions, yet the primary literature has not systematically tested which boundary condition dominates.

1.5. Inconsistencies and research gaps

The bilingual emotional Stroop literature is characterized by three sources of inconsistency. First, effect sizes vary dramatically – from substantial interference to null or reversed facilitation – likely reflecting genuine population differences rather than methodological noise, yet the specific boundary conditions remain unidentified (Eilola et al., 2007; Sharif & Mahmood, 2023; Sutton et al., 2007; Winskel, 2013). Second, moderator variables have been operationalised inconsistently: age-of-acquisition cutoffs range from 5 to 12 years, proficiency is assessed via standardised tests, self-report or not at all, and the emotion-label versus emotion-laden distinction has been largely ignored (Gu & Chen, 2024; Kazanas et al., 2019). Third, the literature disproportionately samples European and East Asian language pairs, with limited representation from understudied populations such as Uyghur–Chinese or Thai–English bilinguals (Sharif & Mahmood, 2023).

These three gaps create a pressing need for systematic synthesis. Although meta-analytic evidence has consolidated findings on the emotional Stroop paradigm in monolingual and clinical populations (Phaf & Kan, 2007), no quantitative synthesis has examined whether bilingual language experience moderates these effects, and prior reviews of bilingual emotional processing have been qualitative rather than meta-analytic (Sharif & Mahmood, 2023). The present study addresses these three gaps by testing three boundary conditions that may partition the observed heterogeneity: neurodevelopmental timing (age of acquisition), valence structure (whether the design included positive words or only negative and neutral words) and proficiency operationalisation (objective test versus self-report) (Borenstein et al., 2021; Higgins et al., 2003).

1.6. The present meta-analysis

The present meta-analysis quantitatively synthesizes the available evidence on emotional Stroop interference in bilinguals. Three research questions guide the synthesis:

RQ1. Does age of L2 acquisition moderate the magnitude of emotional Stroop interference, with early-acquired bilinguals showing stronger effects than late-acquired bilinguals?

RQ2. Does the design-level inclusion of positive valence predict between-study differences in the L2–L1 interference contrast?

RQ3. Does L2 proficiency operationalisation (objective standardised tests versus self-report questionnaires) account for heterogeneity in the pooled effect?

Additional potential moderators, including stimulus type (emotion-label versus emotion-laden), language dominance and task design nuances, were coded descriptively but could not be subjected to formal statistical testing due to insufficient cell sizes or inconsistent reporting across the primary literature (Kazanas et al., 2019). The effect-size metric is the standardised mean difference in colour-identification performance between emotional and neutral conditions, computed as L2 minus L1, with positive values indicating greater interference in L2 (Borenstein et al., 2021). By identifying the population boundaries within which emotional Stroop interference is reliable, this synthesis aims to advance theoretical models of bilingual affective-cognitive interactions and inform the design of future primary studies.

2. Methods

2.1. Literature search

A systematic literature search was conducted across six electronic databases following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines: PubMed, EBSCOhost (PsycINFO, MEDLINE, ERIC, Academic Search Premier, and OpenDissertations), ProQuest, Google Scholar (via Publish or Perish, version 8), CNKI (China National Knowledge Infrastructure), and Wanfang Data. The search was restricted

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

to peer-reviewed articles, dissertations, and theses published between 2007 and 2026. No language restrictions were applied to the search strings; however, only records with full-text availability in English or Chinese were screened for eligibility, with one additional European-language record assessed using translation assistance.

The temporal boundary of 2007 requires explicit justification. The emotional Stroop paradigm has been employed in monolingual populations since 1985 (McKenna & Sharma, 2004; Stroop, 1935). However, direct within-subject comparisons of L1 and L2 emotional word processing using the standardised colour-word Stroop task did not emerge in the indexed literature until 2007, represented by Eilola et al. (2007) and Sutton et al. (2007). These two studies represent the earliest empirical investigations of emotional interference across bilinguals' two languages using manual button-press colour identification, the methodological standard subsequently adopted in this domain. Because no primary studies meeting our inclusion criteria – specifically, bilingual participants, within-subject L1–L2 design, manual button-press response, and extractable trial-level means and standard deviations – were published prior to 2007, extending the search window earlier would not have yielded additional eligible records.

The complete search strings were as follows. For PubMed: ((bilingual*[Title/Abstract] OR multilingual*[Title/Abstract] OR "second language"[Title/Abstract] OR "foreign language"[Title/Abstract])) AND (Stroop[Title/Abstract]) AND (emotion*[Title/Abstract] OR affect*[Title/Abstract] OR valence[Title/Abstract] OR taboo*[Title/Abstract] OR "emotional word*"[Title/Abstract]). For EBSCOhost: (bilingual* OR multilingual* OR "second language" OR "foreign language" OR L1 OR L2 OR "native language") AND (Stroop) AND (emotion* OR affect* OR valence* OR arousal* OR taboo* OR "negative word*" OR "positive word*" OR "emotional word*"). For Google Scholar (via Publish or Perish): (bilingual OR multilingual OR "second language" OR "foreign language" OR L1 OR L2 OR "native language") AND (Stroop) AND (emotion OR emotional OR affective OR valence OR arousal OR taboo OR "negative words" OR "positive words" OR "emotion words" OR "emotion-laden"). For ProQuest: "emotional Stroop" bilingual. For CNKI: "双语 Stroop 情绪." For Wanfang: "双语 Stroop 情绪."

Web of Science Core Collection was additionally searched using the Topic field (Title + Abstract + Author Keywords + Keywords Plus) with the string: TS = ((bilingual* OR multilingual* OR "second language" OR "foreign language" OR L1 OR L2 OR "native language") AND (Stroop) AND (emotion* OR affect* OR valence* OR arousal* OR taboo* OR "negative word*" OR "positive word*" OR "emotional word*")). The search covered 2007–2026 and retrieved 89 records. After cross-referencing against the six-database pool and screening titles and abstracts, no additional eligible studies meeting all inclusion criteria were identified, confirming search saturation.

Search strings were adapted to the syntax and indexing conventions of each database to maximize retrieval efficiency while preserving identical conceptual coverage. PubMed employed [Title/Abstract] field tags, which preclude short alphanumeric terms such as L1 and L2 in field-restricted syntax; EBSCOhost used multi-field natural-language syntax across five sub-databases; Google Scholar relied on full-text Boolean queries without field restrictions via Publish or Perish, necessitating additional keyword variants to compensate for the absence of controlled vocabulary; ProQuest used a streamlined dissertation-indexing syntax given its narrower coverage; and CNKI and Wanfang used Chinese-language keywords in single-box interfaces. Despite surface differences in keyword coverage and syntax, all strings targeted the same three conceptual domains: bilingual or multilingual populations, Stroop paradigm, and emotional or affective content. This database-specific adaptation is standard practice in PRISMA-compliant systematic reviews (Page et al., 2021).

The exact fields searched were as follows: PubMed searched Title and Abstract via [Title/Abstract] tags; EBSCOhost searched Title, Abstract, Keywords, and Subject Terms across five sub-databases; ProQuest searched Title, Abstract, and Keywords; Google Scholar searched indexed full text via Publish or Perish; CNKI and Wanfang searched Title, Abstract, and Keywords in single-box Chinese-language interfaces; and Web of Science searched the Topic field (Title + Abstract + Author Keywords + Keywords Plus). Explicit field restrictions were applied only where required by database syntax; all other databases used the platform's default search interface.

Backward and forward citation tracking was conducted manually by screening the reference lists of all included studies and all relevant review articles identified during title/abstract screening. No additional eligible records were identified through citation tracking. Snowballing via dedicated citation indices was not performed because institutional subscriptions for those platforms were unavailable; this limitation is acknowledged in the search transparency statement. Registers of clinical trials were not searched because the target literature consists entirely of experimental behavioural studies rather than intervention trials.

De-duplication was performed manually prior to title/abstract screening by comparing author names, publication years, and titles across database exports. No reference management software was used. Cross-database overlap was documented, and 21 duplicate records were removed. The full PRISMA 2020 flow diagram (Page et al., 2021) is presented in Figure 1.

2.2. Inclusion and exclusion criteria

Studies were included if they met the following criteria: (1) involved bilingual or multilingual participants who actively used two languages in daily life; (2) employed a standard emotional Stroop paradigm requiring manual button-press responses to identify the ink colour of emotionally valenced and neutral words presented visually; (3) reported sufficient statistical data (means, standard deviations and sample sizes per condition, or equivalent test statistics) to compute standardised mean differences; (4) conducted direct within-subject comparisons of emotional word processing in the first language (L1) and the second language (L2) and (5) used emotionally valenced words (positive, negative or neutral) as stimuli. Studies that presented taboo words as the sole emotional category were excluded because taboo stimuli engage distinct neural and behavioural mechanisms that are not directly comparable to standard valence manipulations.

Studies were excluded if they: (1) used monolingual or non-bilingual samples, or compared bilinguals against monolinguals without within-subject L1–L2 data; (2) employed modified or non-standard Stroop tasks (e.g., auditory Stroop, face-word Stroop, working-memory Stroop, or eye-tracking fixation paradigms); (3) reported reaction time data in incompatible metrics (e.g., block-sum seconds, aggregated response times) from which trial-level means and standard deviations could not be extracted; (4) were review articles, theoretical papers, or conference abstracts without primary empirical data; (5) were published exclusively in non-peer-reviewed venues with insufficient methodological detail to verify

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

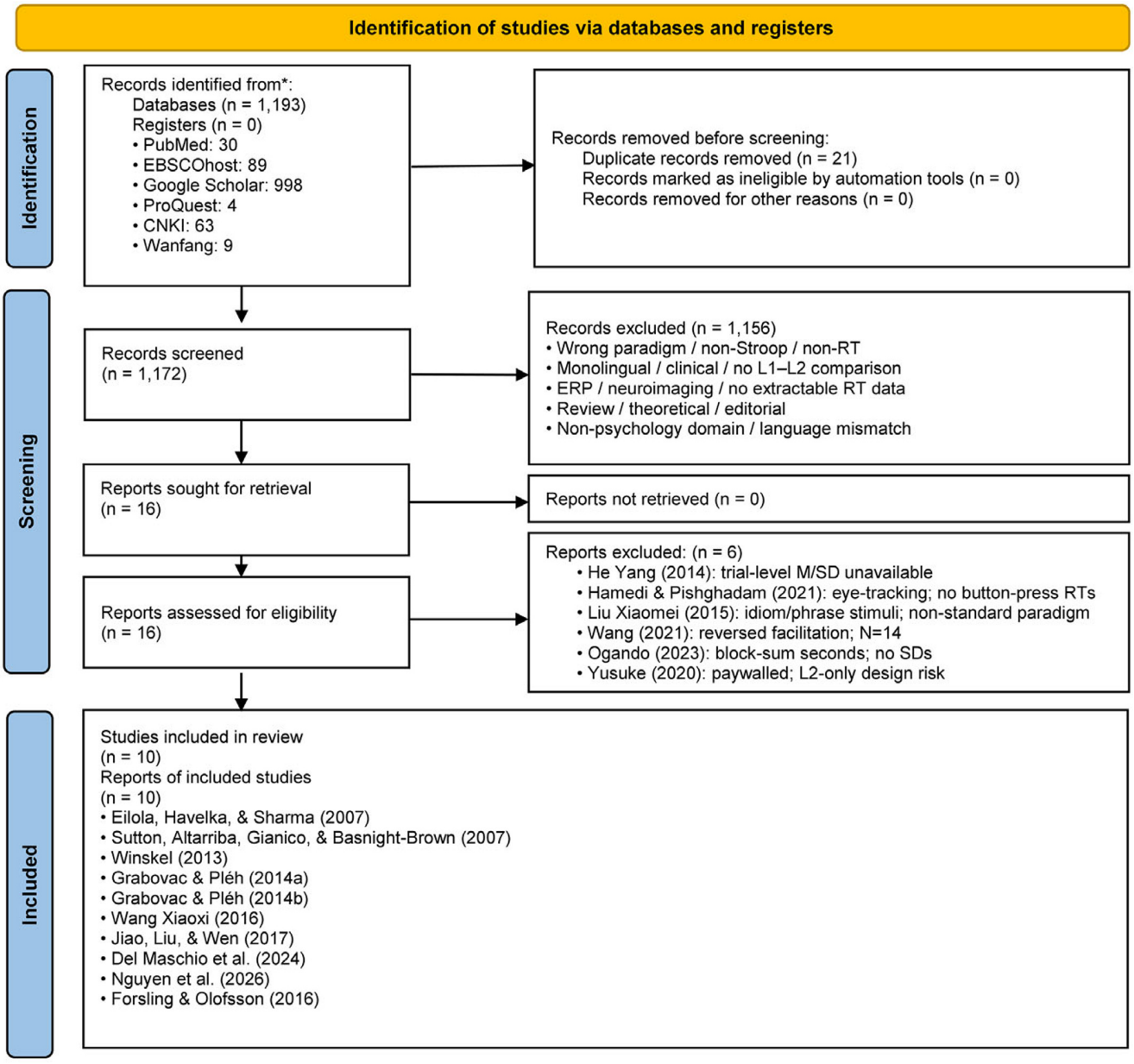


Figure 1. PRISMA 2020 Flow Diagram for Study Selection. *Note:* Google Scholar was searched via Publish or Perish (v8) with a maximum of 1,000 records retrieved per query. Deduplication was performed manually across the six databases. One full-text (Nguyen et al., 2026) was obtained via direct author contact after initial identification in Google Scholar during the search phase; the paper has since been published in *Bilingualism: Language and Cognition.* Scopus was not searched due to institutional subscription unavailability. Web of Science Core Collection was additionally searched as a supplementary verification; it retrieved 89 records, but after deduplication and screening, no additional eligible studies were identified. k = 12 effect sizes (S10 contributes 3 subgroups: comprehensive, reading, listening).

inclusion criteria or (6) did not report direct L1–L2 comparisons. Studies published prior to 2007 were not excluded by date restriction during screening; rather, the database search start date of 2007 reflects the temporal boundary of the indexed literature, and no eligible pre-2007 records were identified during citation tracking.

The distinction between emotion-label words (words that name emotions, e.g., "fear," "joy") and emotion-laden words (words that carry emotional connotation without naming an emotion, e.g., "war," "cancer," "holiday") was systematically coded during data extraction. This distinction was included because prior theoretical work demonstrates that these stimulus classes engage partially dissociable affective systems, and their conflation in previous syntheses has obscured meaningful variation in emotional Stroop effects (Table 1).

2.3. Data extraction and coding

A structured coding protocol was developed to extract moderator variables and effect-size data from each included study. The protocol

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

Table 1. Characteristics of included studies

Study	Language pair	*N*	AoA	Prof. method	Valence	Stimulus
Eilola et al. (2007)	Finnish–English	34	Late	SR	P/N/Neu/T	Laden
Sutton et al. (2007)	Spanish–English	64	Early	SR	N/Neu	Label
Winskel (2013)	Thai–English	54	Late	Obj	N/Neu	Mixed
Grabovac and Pléh (2014a)	Hungarian–Serbian	39	Early	SR	P/N/Neu	Laden
Grabovac and Pléh (2014b)	Hungarian–Serbian	20	Late	SR	P/N/Neu	Laden
Wang (2016)	Chinese–English	29	Late	Mix	P/N/Neu	Mixed
Forsling and Olofsson (2016)	Swedish–English	40	Late	SR	N/Neu	Mixed
Jiao et al. (2017) – Comprehensive	Uyghur–Chinese	20	Early	Mix	P/N/Neu	Laden
Jiao et al. (2017) – Reading	Uyghur–Chinese	20	NR	Mix	P/N/Neu	Laden
Jiao et al. (2017) – Listening	Uyghur–Chinese	20	NR	Mix	P/N/Neu	Laden
Del Maschio et al. (2024)	Italian–English	36	Early	Mix	N/Neu	Laden
Nguyen et al. (2026)	Spanish–English	74	NR	NR	P/N/Neu	Laden

Note: k = 12 effect sizes from 10 studies. Jiao et al. (2017) contributed three independent effect sizes from proficiency subgroups. Abbreviations: N = sample size; AoA = age of acquisition; Prof. Method = proficiency assessment method (SR = self-report only; Obj = objective test only; Mix = mixed methods; NR = not reported); Valence (P = positive; N = negative; Neu = neutral; T = taboo); Stimulus (Laden = emotion-laden; Label = emotion-label; Mixed = mixed stimulus types). All studies employed within-subject L1–L2 designs with manual button-press response modality. Taboo words in Eilola et al. (2007) were excluded from valence moderation analyses. One full-text (Nguyen et al., 2026) was obtained via direct author contact during the search phase; the paper has since been published in Bilingualism: Language and Cognition. Forsling and Olofsson (2016) is an unpublished undergraduate thesis (Examensarbete 30 hp, Umeå University) that met all inclusion criteria and provided extractable trial-level data; it was retained because it contributes rare late-acquisition data for a Scandinavian language pair.

addressed: (1) bibliographic information (author, year, publication type); (2) participant characteristics (sample size, mean age, language pair); (3) bilingual experience variables – age of L2 acquisition (early versus late), language dominance profile, L2 proficiency level, and the method by which proficiency was assessed; (4) task parameters – design (blocked versus mixed), response modality, stimulus language, valence categories presented, and stimulus type (emotion-label, emotion-laden, or mixed) and (5) statistical results – means, standard deviations, and sample sizes for emotional and neutral conditions in L1 and L2.

The age-12 cutoff was selected because it aligns with the critical-period literature in bilingualism research (Abrahamsson & Hyltenstam, 2009; Hartshorne et al., 2018) and because the majority of primary studies in this literature either explicitly used this cutoff or reported mean ages of acquisition that fell clearly on one side of it. When a study reported a mean age of acquisition with a range spanning 12 years, the mean was used to assign the category. When age of acquisition was not reported, the variable was coded as missing for that effect size.

Language dominance was coded as L1-dominant, balanced, or L2-dominant based on the study authors' own classification, verified against self-report or objective measures of relative proficiency and daily usage. "Balanced" bilingualism was defined as the authors' explicit designation of roughly equivalent proficiency and daily use across both languages, supported by quantitative evidence where available. L2 proficiency was coded at two levels: (1) the proficiency level reported by authors (intermediate, advanced, or native-proficient); and (2) the method used to assess proficiency (self-report only, objective test only, or mixed methods combining both). The objective-test variable was coded as "Yes" if the study administered a standardised language test and reported quantitative thresholds or scores, and "No" if proficiency was assessed solely via questionnaire. All extraction was performed manually from full-text articles, Supplementary Materials, and author correspondence where necessary; one full-text was obtained via direct author contact after initial identification in Google Scholar, and this record remained part of the database-identified pool rather than being counted as an "other source" in the PRISMA flow diagram.

2.4. Inter-coder reliability

Because moderator coding is central to the validity of meta-analytic conclusions, inter-coder reliability was assessed for all moderator variables across all 12 effect sizes. A second coder, blind to the primary coder's decisions and trained through an iterative three-round process, independently coded the full dataset. The coding scheme included definitions, examples, and decision rules for each variable. Discrepancies were resolved by consensus discussion, with the primary coder's decision retained when primary-study evidence was ambiguous.

Inter-coder reliability was excellent to perfect across all variables (Cohen, 1960; Cohen's κ = 0.74–1.00, M = 0.95). Agreement was perfect for design, response modality, source type, stimulus type, age of acquisition and proficiency objective (κ = 1.00); excellent for valence type (κ = 0.83); and good for proficiency method (κ = 0.74). The proficiency method variable showed the lowest agreement, reflecting the diversity of assessment tools across primary studies (e.g., standardised tests such as MHK, WRMTR and TEM-4 versus self-report questionnaires such as LHQ and Borg CR-10). Stimulus type was coded with perfect reliability (κ = 1.00), indicating that the distinction between emotion-label and emotion-laden words was applied consistently across coders, and full documentation of the training process, raw coding sheets and disagreement resolution is archived and available upon request (Table 2).

2.5. Risk of bias assessment

Risk of bias was assessed at the study level using a domain-based approach adapted from the Cochrane risk-of-bias framework for cross-over and within-subject designs. Domains included:

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

Table 2. Inter-coder reliability of moderator coding

Variable	Cohen's κ	Agreement (%)	Interpretation
Design	1.000	100	Perfect
Response modality	1.000	100	Perfect
Source type	1.000	100	Perfect
Stimulus type	1.000	100	Perfect
Age of acquisition	1.000	100	Perfect
Proficiency objective	1.000	100	Perfect
Valence types	0.833	91.7	Excellent
Proficiency method	0.742	83.3	Good
Valence mod eligible	1.000	100	Perfect

Note: Two independent coders assessed all 12 effect sizes. Discrepancies were resolved by consensus. The lowest agreement occurred for Proficiency Method, reflecting the diversity of assessment tools across primary studies (e.g., standardised tests such as MHK, WRMTR and TEM-4 versus self-report questionnaires such as LHQ and Borg CR-10).

(1) adequacy of randomisation or counterbalancing of stimulus presentation; (2) standardisation of stimulus presentation and response recording procedures; (3) completeness of outcome data; (4) selective reporting and (5) other sources of bias (e.g., unvalidated stimulus norms, post-hoc outlier exclusion without pre-registration). Each domain was rated as low risk, some concerns, or high risk; two studies were rated as high risk due to very small sample sizes ($N < 30$) and unvalidated stimulus norms, while the remainder showed moderate or low risk. Risk-of-bias ratings were not used to exclude studies but were considered in sensitivity analyses.

Publication bias was assessed through funnel-plot visual inspection, Egger's regression test for funnel-plot asymmetry, and trim-and-fill analysis. Sensitivity was further examined via leave-one-out analysis to identify disproportionately influential studies.

2.6. Effect size calculation

The effect-size metric was Hedges' g, the standardised mean difference in colour-identification performance between emotional and neutral word conditions, computed separately for L1 and L2 and then contrasted as L2 minus L1. Positive values indicate greater emotional interference in L2 relative to L1; negative values indicate relative facilitation or reduced interference in L2. Standard errors were computed from the reported standard deviations and sample sizes. Where studies reported multiple valence categories or proficiency subgroups, each independent contrast was treated as a separate effect size. The nesting of effect sizes within studies was modelled using multilevel meta-analytic specifications to preserve statistical independence.

2.7. Meta-analytic procedures

All analyses were conducted in R (version 4.5.2) using the *metafor* package (Viechtbauer, 2010), following standard procedures for meta-analysis in R (Harrer et al., 2021). Random-effects models with restricted maximum-likelihood estimation (REML) were used for the primary analysis and all subgroup analyses. The multilevel structure was accounted for by specifying study-level random effects when multiple effect sizes were nested within a single study. Moderator analyses were conducted as mixed-effects meta-regressions. Subgroup differences were tested using the omnibus Q-test. The proportion of heterogeneity explained by moderators was quantified using R^2 analogues. All significance tests were two-tailed with $\alpha = .05$.

Heterogeneity was quantified using $^*I^{*2}$ (Higgins & Thompson, 2002), which estimates the percentage of total variation attributable to between-study differences rather than sampling error, and τ^2, which estimates the absolute between-study variance. Prediction intervals were computed to indicate the range within which the effect size of a future study would be expected to fall. Where subgroup analyses were not viable due to insufficient cell sizes, the variable was reported descriptively rather than statistically.

3. Results

3.1. Overall effect size

The random-effects meta-analysis of 12 effect sizes from 10 studies yielded a non-significant pooled estimate of emotional Stroop interference in bilinguals, Hedges' $g = -0.14$, 95% CI [−0.48, 0.21], $z = -0.78$, $p = .435$. The effect-size distribution was characterized by extreme heterogeneity, $I^2 = 91.91\%$, $\tau^2 = 0.33$, $Q(11) = 52.93$, $p < .001$, indicating that the vast majority of observed variance reflected genuine between-study differences rather than sampling error. The 95% prediction interval ranged from −1.27 to 0.99, encompassing both substantial interference and substantial facilitation, which underscores the importance of identifying the moderators that partition this variability. Figure 2 presents the forest plot of individual and pooled effect sizes.

3.2. Moderator analyses

Age of acquisition emerged as the sole significant between-study moderator of emotional Stroop effects. Among the nine effect sizes with codable acquisition timing, mixed-effects meta-regression revealed a significant omnibus effect of AoA category, $Q(1) = 4.97$, $p = .026$, accounting for 63.10% of the between-study heterogeneity ($R^2 = .63$) and reducing residual heterogeneity to $I^2 = 27.52\%$. Early bilinguals showed a small positive interference effect, $g = 0.20$, 95% CI [0.01, 0.39], $p = .039$, whereas late bilinguals exhibited significantly attenuated effects relative to their early-acquired counterparts, $\beta = -0.30$, 95% CI [−0.56, −0.04], $p = .026$. Three effect sizes lacked explicit AoA information – the reading and listening subgroups from Jiao et al. (2017) and Nguyen et al. (2026) – and were excluded from this analysis to preserve coding rigour.

Proficiency assessment method did not significantly account for between-study variability. The contrast between effect sizes that employed objective tests or mixed methods ($k = 6$) and those relying exclusively on self-report ($k = 5$) was non-significant, $Q(1) = 0.89$, $p = .346$, and explained none of the heterogeneity ($R^2 = .00$). Residual heterogeneity remained extreme, $I^2 = 92.96\%$. It is noteworthy that the six effect sizes using objective measures were drawn from more diverse language pairs (Thai–English, Chinese–English, Uyghur–Chinese, Italian–English) than the five self-report effect sizes, which clustered in European populations.

Similarly, between-study variation in valence eligibility – contrasting the eight effect sizes that presented positive, negative, and neutral words against the four effect sizes presenting only negative and neutral words – did not reach significance, $Q(1) = 0.48$, $p = .490$, and explained no heterogeneity ($R^2 = .00$).

Stimulus type could not be subjected to formal subgroup analysis because the distribution was heavily skewed: eight effect

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

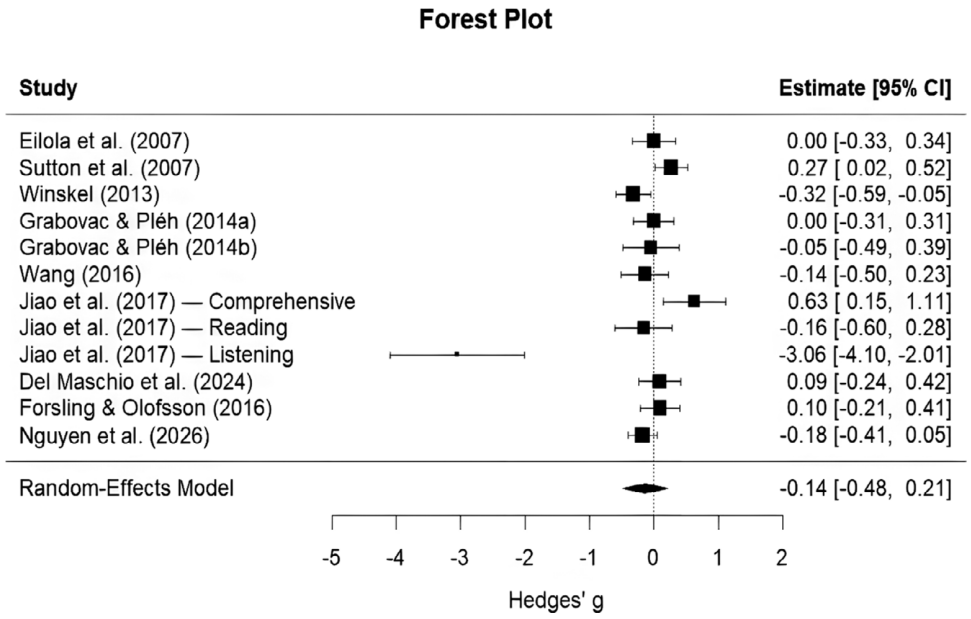


Figure 2. Forest plot of individual and pooled effect sizes. *Note*: Positive Hedges' *g* values indicate greater emotional interference in L2 relative to L1.

sizes used exclusively emotion-laden words, three effect sizes used mixed stimuli, and only one effect size (Sutton et al., 2007) used emotion-label words. With a single observation in the emotion-label category, any statistical contrast would be uninformative and potentially misleading.

3.3. Publication bias and sensitivity analyses

Visual inspection of the funnel plot revealed marked asymmetry, with the extreme negative effect size from Jiao et al. (2017) – Listening subgroup ($g = -3.06$) falling well below the lower bound of the expected distribution. Egger's regression test for funnel-plot asymmetry (Egger et al., 1997) was significant, $z = -3.71$, $p = .0002$, corroborating the visual impression of asymmetry. Trim-and-fill analysis (Duval & Tweedie, 2000) estimated zero missing studies on the right side of the funnel, indicating that the asymmetry was driven by the presence of an extreme negative outlier rather than by suppression of small positive studies (Figure 3).

Leave-one-out sensitivity analysis identified the listening-proficient subgroup from Jiao et al. (2017) as disproportionately influential. Removing this single effect size reduced the pooled estimate to $g = 0.004$, 95% CI [−0.13, 0.14], $z = 0.06$, $p = .956$, and attenuated heterogeneity to $I^2 = 49.17\%$, a level conventionally considered moderate (Higgins et al., 2003). No other study exerted comparable leverage; removing any of the remaining 11 effect sizes produced pooled estimates ranging from $g = -0.19$ to $g = -0.15$, all non-significant. The corresponding forest plot is presented in Figure 4.

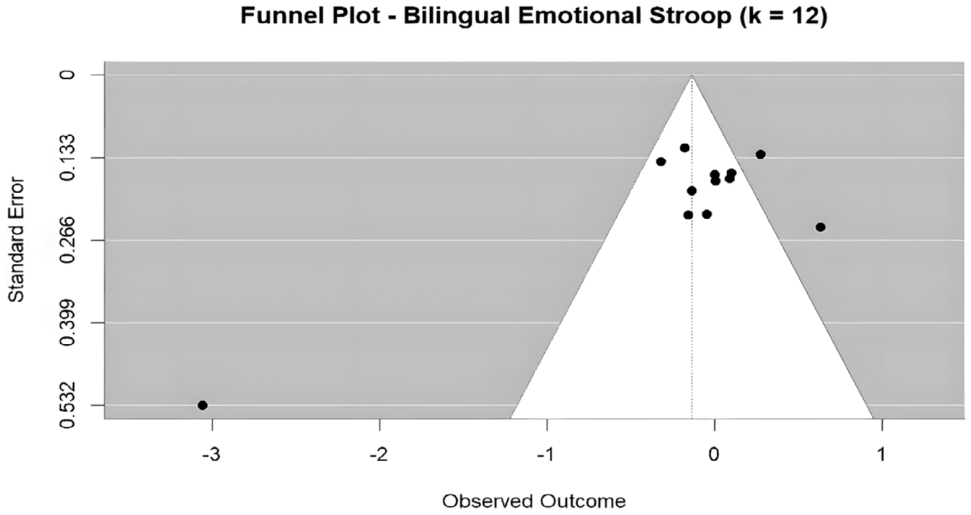


Figure 3. Funnel plot of effect sizes against standard errors. *Note:* The asymmetry was driven by an extreme negative outlier (Jiao et al., 2017, listening-proficient subgroup, $g = -3.06$). Trim-and-fill analysis estimated zero missing studies.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

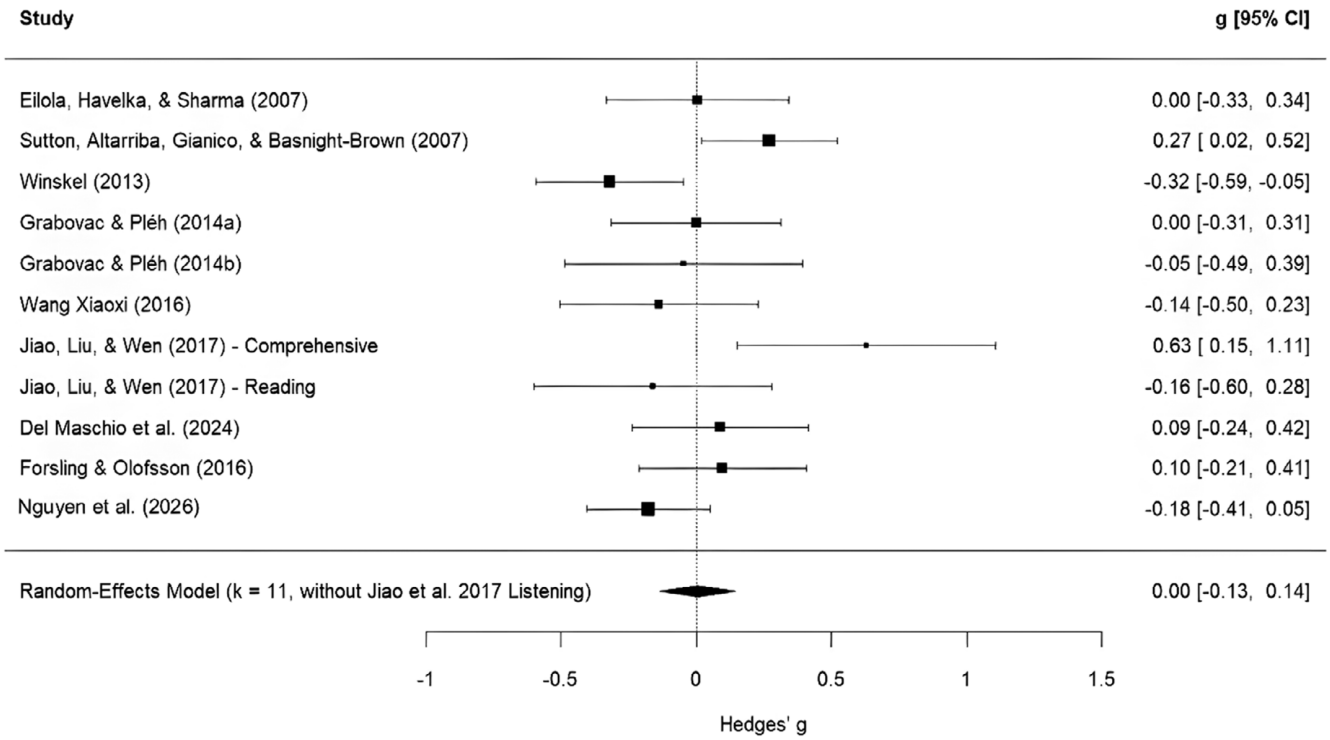


Figure 4. Sensitivity-analysis forest plot after exclusion of the Jiao et al. (2017) listening-proficient outlier. *Note:* The pooled estimate approached zero (Hedges' g = 0.004, 95% CI [−0.13, 0.14], p = .956), and heterogeneity attenuated to moderate levels (I^2 = 49.17%), confirming that the null overall effect is robust and not driven by the extreme outlier.

4. Discussion

4.1. Age of acquisition as a neurodevelopmental boundary condition

The first research question addressed whether age of L2 acquisition moderates emotional Stroop interference. The meta-analytic evidence supports a clear affirmative answer: early bilinguals exhibited a small but significant interference effect, whereas late bilinguals showed attenuated effects that did not differ significantly from zero. This finding aligns with neurodevelopmental models proposing that childhood acquisition of two languages fosters more integrated language–emotion neural networks, which paradoxically heighten vulnerability to affective distraction through more automatic emotional processing in both language systems (Grundy, 2020; Jafari et al., 2022). The present results extend these models by demonstrating that the degree of integration is not merely a neural correlate but manifests behaviourally as heightened Stroop interference when emotionally salient words compete with colour-identification demands.

Theoretically, this pattern reflects the neurodevelopmental consequences of childhood bilingual acquisition. Early bilinguals in the present pool – acquiring their L2 before age 12 and in some cases during kindergarten – likely developed more automatized affective associations in both languages during the putative critical period for emotional encoding, rendering emotional words more salient irrespective of the language context (Dang et al., 2023). In contrast, late bilinguals, who acquired their L2 after this window, may retain more segregated language–emotion representations, resulting in dampened cross-linguistic interference. The exclusion of the three effect sizes that lacked explicit AoA information – the reading and listening subgroups from Jiao et al. (2017) and Nguyen et al. (2026) – strengthens confidence that the significant AoA effect is not an artefact of ambiguous coding.

4.2. Valence effects and the level-of-analysis problem

The second research question examined whether the design-level inclusion of positive valence predicts between-study differences in the L2–L1 interference contrast. The between-study comparison of valence eligibility – contrasting effect sizes that included positive, negative, and neutral words against those presenting only negative and neutral words – did not account for heterogeneity and did not reach significance. This null finding should be interpreted carefully: it addresses whether the *design-level inclusion* of positive valence predicts the L2–L1 interference contrast across studies, not whether negative words produce larger within-study interference than positive or neutral words within the same participants. The latter question requires trial-level valence-specific means and standard deviations, which the primary studies reported inconsistently; consequently, the present meta-analysis cannot adjudicate whether negative valence produces the strongest interference *within* individuals – a question that remains open and requires primary studies with fully crossed valence-by-language designs.

Descriptively, however, the studies that reported valence-specific contrasts showed a consistent directional pattern: negative words elicited numerically larger interference than positive or neutral words where such comparisons were available (Eilola et al., 2007; Grabovac & Pléh, 2014a, 2014b; Jiao et al., 2017). This descriptive consistency, combined with the non-significant between-study moderator, suggests that valence effects may operate primarily at the within-subject level rather than as a between-study design feature. Evolutionary accounts of threat detection posit that negatively valenced stimuli command prioritized attentional

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

resources because they carry survival-relevant information that must be processed rapidly (Citron et al., 2020; Harris et al., 2003). If negative words indeed produce stronger interference within individuals, as the descriptive pattern suggests, this would support neurobiological hypotheses according to which affective salience directly engages executive control systems, especially in bilinguals who must simultaneously manage competition between two active language systems (Dang et al., 2023). Future primary research should report complete valence-by-language descriptives to enable a more powerful meta-analytic test of this hypothesis.

4.3. Proficiency, dominance, and the measurement gap

The third research question addressed whether L2 proficiency operationalisation accounts for heterogeneity in the pooled effect. Neither proficiency assessment method (objective test versus self-report) nor the broader dominance-by-proficiency interaction could be reliably evaluated in the current pool. The proficiency-objective contrast was non-significant, explaining none of the heterogeneity. More fundamentally, formal language-dominance subgroup analysis was precluded because only one effect size in the pool represented L2-dominant bilinguals, and several studies that would have contributed to dominance coding were excluded at the full-text stage due to incompatible paradigms or insufficient data. This distribution is not a statistical inconvenience but a substantive finding: the bilingual emotional Stroop literature has disproportionately sampled L1-dominant and balanced bilinguals, leaving L2-dominant populations severely under-represented.

The null proficiency-objective result also highlights a measurement problem. The six effect sizes that administered standardised tests (e.g., WRMTR, MHK, TEM-4, EPT) drew from more diverse language pairs than the five effect sizes relying on self-report questionnaires (e.g., LHQ, LEAP-Q, Borg CR-10). The heterogeneity in assessment tools – reflected in the lowest inter-coder reliability for this variable – may have introduced noise that obscured a true proficiency effect (Zhang & Fan, 2023). The field would benefit from adopting a common proficiency battery to reduce measurement error and enhance cross-study comparability. Furthermore, theoretical work suggests that as L2 proficiency increases, emotional words in the second language acquire richer affective representations, narrowing the gap between L1 and L2 emotional processing (Weimer et al., 2022); however, the present literature base is too sparse to test this developmental prediction meta-analytically.

The non-significant effects of proficiency assessment method and valence eligibility are theoretically informative, though small-pool power limitations preclude definitive conclusions. They suggest that the field's inconsistencies are unlikely to be driven primarily by superficial methodological differences (self-report versus objective testing; positive-inclusive versus negative-neutral designs), and instead point towards a deeper population-level variable: the timing of L2 acquisition. Ruling out these alternative explanations strengthens confidence that age of acquisition is the primary boundary condition of bilingual emotional Stroop interference.

4.4. The emotion-label versus emotion-laden imbalance

A systematic distinction between emotion-label words (e.g., "fear," "joy") and emotion-laden words (e.g., "war," "cancer," "holiday") was incorporated into the coding protocol following theoretical arguments that these classes engage partially dissociable affective systems. Emotion-label words directly name emotional states and may therefore activate conceptual representations of affect more strongly than emotion-laden words, which carry emotional connotation without explicit semantic reference to emotion categories (Gu & Chen, 2024). However, the distribution of stimulus types in the identified literature precluded formal subgroup analysis: eight of twelve effect sizes used exclusively emotion-laden words, three used mixed stimuli, and only one effect size (Sutton et al., 2007) used emotion-label words. With a single observation in the emotion-label category, any statistical contrast would be uninformative and potentially misleading.

This imbalance is itself a noteworthy characteristic of the field. Researchers in bilingual emotional Stroop research have preferentially selected emotion-laden materials, perhaps because they are easier to norm for valence and arousal across languages, or because they are perceived as more ecologically valid in communicative contexts. The near-absence of emotion-label studies means that the theoretical prediction – derived from monolingual work and highlighted in recent bilingual research (Gu & Chen, 2024) – that emotion-label words produce stronger interference because they directly name affective states remains untested in bilingual meta-analytic research. Future primary studies should include both stimulus types within the same design to enable a direct comparison.

4.5. Methodological implications

The leave-one-out sensitivity analysis demonstrated that no single study drove the null overall effect; rather, the identified literature base as a whole does not support a reliable average emotional Stroop interference effect in bilinguals. The one extreme outlier – the listening-proficient subgroup from Jiao et al. (2017), which showed massive facilitation – was retained in the primary analysis because it met all inclusion criteria. This massive facilitation is theoretically interpretable: listening-proficient bilinguals possessed advanced auditory L2 comprehension but underdeveloped visual-semantic L2 links, rendering printed emotional words less attentionally salient in L2 than in L1 and producing relative facilitation rather than interference. Its exclusion reduced heterogeneity to moderate levels but did not render the overall effect significant, confirming that the null finding is robust and not an artefact of a single unusual sample.

The publication bias assessment, which indicated funnel-plot asymmetry driven by this outlier rather than by suppression of small positive studies, further supports the conclusion that the literature's variability is substantive. Trim-and-fill estimated zero missing studies, suggesting that the asymmetry is not a file-drawer problem but a genuine reflection of diverse population effects. These methodological patterns underscore the importance of moving beyond pooled estimates towards conditional, moderator-informed conclusions in bilingualism meta-analyses.

This level of heterogeneity (I^2 = 91.91%) is not atypical for bilingualism meta-analyses. Lehtonen et al. (2018) reported I^2 > 80% in their executive-function synthesis, and Donnelly et al. (2019) found substantial residual heterogeneity even after moderator testing. Such variability reflects the genuine diversity of bilingual populations – spanning simultaneous childhood bilinguals to late adult learners – rather than methodological noise. The present finding that a single moderator (age of acquisition) explains 63% of this variance is therefore methodologically noteworthy, indicating that neurodevelopmental timing partitions a literature that has previously pooled incompatible populations.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

4.6. Limitations and future directions

Several limitations bound the present conclusions. First, the pool of 12 effect sizes, while rigorously screened and independently coded, remains small. This small pool reflects the genuine immaturity of the bilingual emotional Stroop literature rather than incomplete searching; synthesizing 10 verified studies with independent coding and extractable data is methodologically preferable to synthesizing many studies with coding errors and ambiguous classifications. With a small pool, mixed-effects meta-regression has limited statistical power to detect small-to-moderate moderator effects; the non-significant results for proficiency method and valence eligibility should therefore be interpreted as inconclusive rather than as evidence for genuine null effects. Second, three effect sizes lacked codable age-of-acquisition data, and one lacked codable proficiency information, reducing the effective sample for those moderators. Third, all included studies used cross-sectional designs, precluding causal inferences about how emotional Stroop effects change as proficiency or dominance shift longitudinally. Fourth, the response modality was uniformly manual button-press; although this standardisation reduces methodological heterogeneity, it also limits generalisability to oral colour-naming paradigms. Fifth, the strict inclusion criteria – requiring colour-word emotional Stroop, within-subject L1–L2 designs, manual button-press responses, and extractable means and standard deviations – yielded a final pool of only 10 studies (12 effect sizes) from 1,193 screened database records across six databases. While this rigour safeguards methodological homogeneity, it also excludes a substantial body of related evidence.

Future research should prioritize five targets. First, future meta-analyses could revise these criteria to incorporate additional Stroop variants (e.g., face-word, picture-word, spatial, and auditory paradigms), between-subjects designs, and emotional phrases or sentences, thereby drawing more generalisable and practically applicable conclusions. Second, the field needs more primary studies with L2-dominant bilinguals to enable a definitive test of dominance moderation. Third, researchers should report complete valence-by-language descriptive statistics to permit within-study valence contrasts in future meta-analyses. Fourth, the inclusion of emotion-label words alongside emotion-laden words within the same bilingual design would test the stimulus-type hypothesis that the present literature base could not adjudicate. Fifth, adopting standardised proficiency measures – ideally objective tests with published thresholds – would reduce measurement noise and clarify whether proficiency genuinely moderates emotional interference once assessment error is minimized.

5. Conclusion

The present synthesis advances the understanding of bilingual emotional processing by shifting the theoretical emphasis from whether bilinguals experience emotional Stroop interference to when and for whom such effects are reliable. The null overall effect, combined with the significant age-of-acquisition moderation, suggests that early- and late-acquired bilinguals represent distinct populations with respect to emotional Stroop interference, and that pooling them in previous narrative reviews may have obscured a conditional pattern. The identification of age of acquisition as a significant moderator – accounting for the majority of between-study heterogeneity – establishes neurodevelopmental timing as a boundary condition that previous primary studies have treated inconsistently. Early-acquired bilinguals, who developed overlapping language–emotion representations during childhood, showed robust interference; late-acquired bilinguals did not. This conditional pattern reconciles previously conflicting findings by showing that population characteristics, not methodological noise, drive the variability that has long characterized this literature.

The broader implication is that bilingual emotional processing cannot be understood through universal claims. The null moderation by proficiency assessment method and valence eligibility, alongside the structural impossibility of testing stimulus-type effects, reveals that the field's inconsistencies stem not from a lack of coherent underlying mechanisms but from a literature base that has sampled too narrowly from specific bilingual populations and stimulus categories. Emotional Stroop interference in bilinguals is therefore neither a fixed trait nor a methodological artefact; it is an emergent property of the interaction between neurodevelopmental history, language experience, and experimental design.

These findings carry practical weight for education and clinical practice. Educators designing bilingual programmes should recognize that the emotional salience of instructional materials may interact with learners' language histories in ways that affect attentional engagement. Clinicians assessing bilingual clients should consider that emotional reactivity profiles may differ systematically between early and late bilinguals, with implications for the choice of assessment language and the interpretation of affective responses. Moving forward, the field requires primary studies that span understudied language pairs, employ standardised proficiency measures, and report complete valence-specific data. Only through such systematic expansion can the next generation of research move beyond the question of existence towards a precise, conditional understanding of bilingual emotional cognition.

Data availability statement. All data analysed during this meta-analysis are derived from the primary sources cited in the References and summarized in Table 1. The coding sheets and inter-coder reliability documentation are archived and available upon request.

Author contributions. The corresponding author conceptualized the research topic and established the manuscript outline. The first author performed the literature search and data collection, meta-analytic synthesis, and drafted the manuscript under the instruction of the corresponding author. Both authors finalized the manuscript.

Funding statement. This study was supported by the National Social Science Fund of China (Grant No. 23BYY169).

Competing interests. The authors declare no conflicts of interest or competing financial interests.

Ethical compliance. This research is a meta-analysis synthesizing previously published studies. All primary studies included in this synthesis reported receiving ethical approval and informed consent from participants. No new primary data were collected for this study.

Artificial intelligence statement. During the preparation of this manuscript, generative AI tools (Kimi AI and DeepSeek-V3) were used for linguistic refinement, proofreading, and structural organisation of the revised draft. These tools were employed in May 2026 to resolve grammatical errors and inconsistencies identified during the co-author's review of the draft. All core scientific content, including the meta-analytic data synthesis, effect size calculations, statistical interpretations, and theoretical arguments, was produced solely by the authors without AI assistance. The authors take full responsibility for the final manuscript.

References

Abrahamsson, N., & **Hyltenstam, K.** (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, **59**(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

Anooshian, L. J., & **Hertel, P. T.** (1994). Emotionality in bilinguals: The use of multiple languages in emotional situations. *Journal of Psycholinguistic Research*, **23**(3), 217–233.

Bialystok, E. (2011). Reshaping the mind: The benefits of bilingualism. *Canadian Journal of Experimental Psychology*, **65**(4), 229–235. https://doi.org/10.1037/a0025406

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., **Craik, F. I. M.**, **Klein, R.**, & **Viswanathan, M.** (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging*, **19**(2), 290–303. https://doi.org/10.1037/0882-7974.19.2.290

Borenstein, M., **Hedges, L. V.**, **Higgins, J. P. T.**, & **Rothstein, H. R.** (2021). *Introduction to meta-analysis* (2nd ed.). Wiley. https://doi.org/10.1002/9781119558378.

Caldwell-Harris, C. L. (2015). Emotionality differences between a native and foreign Language. *Current Directions in Psychological Science*, **24**(3), 214–219. https://doi.org/10.1177/0963721414566268

Caldwell-Harris, C. L., **Tong, J.**, **Lung, W.**, & **Poo, S.** (2011). Physiological reactivity to emotional phrases in Mandarin–English bilinguals. *International Journal of Bilingualism*, **15**(3), 329–352. https://doi.org/10.1177/1367006910379262

Citron, F. M. M., **Cacciari, C.**, & **Kucharski, M.** (2020). When emotions are expressed figuratively: Psycholinguistic and affective norms for 619 German idioms expressing emotions. *Behavior Research Methods*, **52**(4), 1694–1711.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46. https://doi.org/10.1177/001316446002000104

Dang, Q., **Ma, F.**, **Yuan, Q.**, **Fu, Y.**, **Chen, K.**, **Zhang, Z.**, **Lu, C.**, & **Guo, T.** (2023). Processing negative emotion in two languages of bilinguals: Accommodation and assimilation of the neural pathways based on a meta-analysis. *Cerebral Cortex*, **33**(13), 8352–8367. https://doi.org/10.1093/cercor/bhad121

Degner, J., **Doycheva, C.**, & **Wentura, D.** (2012). It matters how much you talk: On the automaticity of affective connotations of first and second language words. *Bilingualism: Language and Cognition*, **15**(1), 181–189. https://doi.org/10.1017/S1366728911000095

Del Maschio, N., **Sulpizio, S.**, **Bellini, C.**, **Del Mauro, G.**, **Giannachi, M.**, **Buga, D.**, **Fedeli, D.**, **Perani, D.**, & **Abutalebi, J.** (2024). Neurocognitive mechanisms of emotional interference in native and foreign languages: Evidence from proficient bilinguals. *Frontiers in Behavioral Neuroscience*, **18**(1392005). https://doi.org/10.3389/fnbeh.2024.1392005

Dewaele, J.-M. (2008). The emotional weight of *I love you* in multilinguals' languages. *Journal of Pragmatics*, **40**(10), 1753–1780. https://doi.org/10.1016/j.pragma.2008.03.002

Dewaele, J.-M. (2010). *Emotions in multiple languages*. Palgrave Macmillan. https://doi.org/10.1057/9780230289505

Donnelly, S., **Brooks, P. J.**, & **Homer, B. D.** (2019). Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychonomic Bulletin & Review*, **26**(4), 1122–1147. https://doi.org/10.3758/s13423-019-01567-z

Duval, S., & **Tweedie, R.** (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56**(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., **Smith, G. D.**, **Schneider, M.**, & **Minder, C.** (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, **315**(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Eilola, T. M., **Havelka, J.**, & **Sharma, D.** (2007). Behavioural and physiological responses to the emotional and taboo Stroop tasks in native and non-native speakers of English. *International Journal of Bilingualism*, **11**(2), 123–143.

Ferré, P., **García, T.**, **Fraga, I.**, **Sánchez-Casas, R.**, & **Molero, M.** (2010). Memory for emotional words in bilinguals: Do words have the same emotional intensity in the first and in the second language? *Cognition & Emotion*, **24**(5), 760–785. https://doi.org/10.1080/02699930902985779

Forsling, A., & **Olofsson, I.** (2016). *Effekter av emotionellt innehåll i språk hos tvåspråkiga: En experimentell studie med emotionell Stroop av svenskengelsktalande* [Effects of emotional content in language in bilinguals: An experimental study with emotional Stroop in Swedish-English speakers] [Unpublished undergraduate thesis]. Umeå University.

Grabovac, I., & **Pléh, C.** (2014a). Emotional Stroop task in early and late Hungarian–Serbian bilinguals. *Journal of Psycholinguistic Research*, **43**(5), 613–625.

Grabovac, I., & **Pléh, C.** (2014b). Emotional Stroop task in late Hungarian–Serbian bilinguals. *Psihologija*, **47**(4), 431–444.

Green, D. W. (2018). Language control and code-switching. *Bilingualism: Language and Cognition*, **3**(2), 916–917. https://doi.org/10.3390/languages3020008

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

Grosjean, F. (2010). *Bilingual: Life and reality*. Harvard University Press. https://doi.org/10.4159/9780674056459

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

Gu, X., & **Chen, S.** (2024). Two languages, two emotional minds in one brain: Processing emotion-label and emotion-laden words by Chinese-English bilinguals. *International Journal of Bilingual Education and Bilingualism*, **27**(1), 1–21.

Harrer, M., **Cuijpers, P.**, **Furukawa, T. A.**, & **Ebert, D. D.** (2021). *ng meta-analysis with R: A hands-on guide*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003107347

Harris, C. L., **Aycicegi, A.**, & **Gleason, J. B.** (2003). Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language. *Applied PsychoLinguistics*, **24**(4), 561–579. https://doi.org/10.1017/S0142716403000286

Hartshorne, J. K., **Tenenbaum, J. B.**, & **Pinker, S.** (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, **177**, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Higgins, J. P. T., & **Thompson, S. G.** (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, **21**(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., **Thompson, S. G.**, **Deeks, J. J.**, & **Altman, D. G.** (2003). Measuring inconsistency in meta-analyses. *BMJ*, **327**(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Jafari, Z., **Villeneuve, C.**, **Thompson, J.**, & **Koravand, A.** (2022). Subcortical plasticity and enhanced neural synchrony in multilingual adults. *Bilingualism: Language and Cognition*, **25**(4), 678–691.

Jiao, J., **Liu, Y.**, & **Wen, S.** (2017). Automatic emotional access in emotional Stroop of different proficient type of bilinguals. *Acta Psychologica Sinica*, **49**(9), 1150–1157. https://doi.org/10.3724/SP.J.1041.2017.01150

Kazanas, S. A., **McLean, D.**, & **Altarriba, J.** (2019). Emotion word processing in bilinguals: The role of emotion, language, and task type. In J. Altarriba & R. R. Heredia (Eds.), *An introduction to bilingualism: Principles and processes* (2nd ed., pp. 215–238). Routledge.

Kroll, J. F., & **Bialystok, E.** (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, **25**(5), 497–514. https://doi.org/10.1080/20445911.2013.799170

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

Lindquist, K. A. (2017). The role of language in emotion: Existing evidence and future directions. *Current Opinion in Psychology*, **17**, 135–139. https://doi.org/10.1016/j.copsyc.2017.07.006

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, **109**(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

Martynova, O., & **Lyusin, D.** (2026). The role of stimuli valence and arousal in emotional word processing: Evidence from the emotional Stroop task. *Psihologija*, **59**(1), 1–18.

McKenna, F. P., & **Sharma, D.** (2004). Reversing the emotional Stroop effect reveals that it is not what it seems: The role of fast and slow components. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **30**(2), 382–392.

Nguyen, M. V. H., **Tamber-Rosenau, B. J.**, & **Hernandez, A. E.** (2026). *Emotional interference in cognitive tasks: The role of language experience and task demand in monolinguals and heritage bilinguals*. Cambridge University Press.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795

Nichols, E. S., **Wild, C. J.**, **Stojanoski, B.**, **Battista, M. E.**, & **Owen, A. M.** (2020). Bilingualism affords no general cognitive advantages: A population study of executive function. *Developmental Science*, **23**(5), e12989

Page, M. J., **McKenzie, J. E.**, **Bossuyt, P. M.**, **Boutron, I.**, **Hoffmann, T. C.**, **Mulrow, C. D.**, **Shamseer, L.**, **Tetzlaff, J. M.**, **Akl, E. A.**, **Brennan, S. E.**, **Chou, R.**, **Glanville, J.**, **Grimshaw, J. M.**, **Hróbjartsson, A.**, **Lalu, M. M.**, **Li, T.**, **Loder, E. W.**, **Mayo-Wilson, E.**, **McDonald, S.**, … **Moher, D.** (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, **372**(71)

Pavlenko, A. (2012). Affective processing in bilingual speakers: Disembodied cognition? *International Journal of Psychology*, **47**(6), 405–428. https://doi.org/10.1080/00207594.2012.743665

Phaf, R. H., & **Kan, K.-J.** (2007). The automaticity of emotional Stroop: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry*, **38**(2), 184–199. https://doi.org/10.1016/j.jbtep.2006.10.008

Sharif, H., & **Mahmood, S.** (2023). Emotional processing in bilinguals: A systematic review aimed at identifying future trends in neurolinguistics. *Humanities and Social Sciences Communications*, **10**(1), 1–12. https://doi.org/10.1057/s41599-023-01926-1

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, **18**(6), 643–662. https://doi.org/10.1037/h0054651

Sutton, T. M., **Altarriba, J.**, **Gianico, J. L.**, & **Basnight-Brown, D. M.** (2007). The automatic access of emotion: Emotional Stroop effects in Spanish–English bilingual speakers. *Cognition & Emotion*, **21**(5), 1077–1090. https://doi.org/10.1080/02699930601054133

Viechtbauer, W. (2010). Conducting meta-analyses inRwith themetaforPackage. *Journal of Statistical Software*, **36**(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Wang, X. (2016). *The time course of bilingual emotional word processing in Chinese-English late bilinguals: A behavioral and event-related potential study* [Unpublished master's thesis]. Ningbo University.

Weimer, A., **Koniakowsky, I.**, **Nazir, T. A.**, & **Huckauf, A.** (2022). Behavioral evidence for differences in emotional resonance during processing first and second language. *International Journal of Bilingualism*, **26**(6), 784–798. https://doi.org/10.1177/13670069211073578

Williams, J. M. G., **Mathews, A.**, & **MacLeod, C.** (1996). The emotional Stroop task and psychopathology. *Psychological Bulletin*, **120**(1), 3–24. https://doi.org/10.1037/0033-2909.120.1.3

Winskel, H. (2013). The emotional Stroop task and emotionality rating of negative and neutral words in late Thai–English bilinguals. *International Journal of Psychology*, **48**(6), 1090–1098. https://doi.org/10.1080/00207594.2013.793800

Yurtsever, A., **Anderson, J. A. E.**, & **Grundy, J. G.** (2024). Bilingual experiences and executive functions: A mouse-tracking study using Stroop and Flanker tasks. *Bilingualism: Language and Cognition*, **27**(1), 1–15.

Zhang, J., & **Fan, L.** (2023). The effects of executive functions on language control during Chinese-English emotional word code-switching. *Frontiers in Psychology*, **14**, Article 1087513. https://doi.org/10.3389/fpsyg.2023.1087513

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 07:39:11, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S1366728926101795