Psychological Review

Do Bilingual Advantages in Domain-General Executive Functioning Occur in Everyday Life and/or When Performance-Based Measures Have Excellent Psychometric Properties?

Kenneth R. Paap, John Majoubi, Regina T. Anders-Jefferson, Rin Iosilevsky, and Charlotte Ursula Tate

Online First Publication, February 13, 2025. https://dx.doi.org/10.1037/rev0000537

CITATION

Paap, K. R., Majoubi, J., Anders-Jefferson, R. T., Iosilevsky, R., & Tate, C. U. (2025). Do bilingual advantages in domain-general executive functioning occur in everyday life and/or when performance-based measures have excellent psychometric properties? *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000537

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2025 American Psychological Association
ISSN: 0033-295X

https://doi.org/10.1037/rev0000537

Do Bilingual Advantages in Domain-General Executive Functioning Occur in Everyday Life and/or When Performance-Based Measures Have Excellent Psychometric Properties?

Kenneth R. Paap, John Majoubi, Regina T. Anders-Jefferson, Rin Iosilevsky, and Charlotte Ursula Tate
Department of Psychology, San Francisco State University

Psychologists have sought to understand individual differences in the ability to control thoughts, emotions, and actions during goal-directed behavior. Issues include whether the ability is unitary or componential and whether it is domain-general or task-specific. If domain-general, is it highly heritable with scant room for environmental influence or can it be enhanced by the right type of life experience or formal training? These questions have triggered one of the most heated debates in cognitive science, viz., is there a bilingual advantage in executive functioning (EF)? The empirical part of this study reports a substantially improved test of the bilingual advantage hypothesis in three respects. First, it tests for relationships between bilingualism and EF at the latent-variable level. Second, it extracts a latent-variable for performance-based measures of EF that are psychometrically strong. Third, it also includes a latent-variable based on self-rating scales of self-control/impulsivity that have enjoyed considerable success in predicting real-world outcomes. The results provide no evidence for a bilingual advantage on EF performance and a small, but significant, relationship to self-ratings. However, the relationship to self-ratings is no longer significant, when social desirability is taken into account. The correlation between the latent variables for performance-based and self-ratings was near 0, suggesting that they are separate constructs.

Keywords: executive functioning, self-control, bilingualism, impulsivity, self-ratings

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

For more than a decade, there has been a lively debate regarding a hypothesized bilingual advantage in executive functioning (EF; Bialystok, 2019, 2024, Paap, 2019, 2023). As reviewed in more detail below, the empirical evidence has been mixed at best with more null results than significant effects. However, there have been good reasons to question the plethora of null results. One concern is that many studies inadequately measured bilingualism. Early studies often compared groups of bilinguals to groups of monolinguals by partitioning the sample on the basis of a small number of self-reported facets of bilingual experience. A general consensus has emerged that bilingualism should be measured as a multidimensional and continuous construct. A second concern is that performance-based measures of EF usually have poor psychometric properties and lack the convergent validity and test–retest reliability needed for powerful tests of the hypothesis. A third concern is that there appears to be a dissociation between performance-based measures of EF and self-rating scales of controlled behavior in everyday life. Furthermore, there has been a paucity of tests of the bilingual-advantage hypothesis that use self-rating scales and focus on real-world performance. The empirical focus of this study addresses these problems by deriving coherent latent variables for bilingualism, performance-based EF, and ratings-of-EF in everyday life. Furthermore, we use a structural equation model (SEM) model to test for the effects of bilingualism on each of the two measures of EF. If a wholistic latent-variable measure of bilingualism has no effect on either laboratory measures of EF or self-ratings in everyday life, then perhaps we can advance much further down the road to a consensus that, as a practical matter, there are no meaningful effects of bilingualism on EF.

Lisa Chinn served as action editor.

Kenneth R. Paap https://orcid.org/0000-0001-9528-930X

This study was not preregistered. The JASP datafile used for the structural equation model and Bayes factor analyses are available at https://osf.io/j8gvh/?view_only=9ffe42c84278487d9f541efd3ef1a44c.

The authors have no known conflicts of interest to disclose. The authors thank Nithya Balakrishnan, Cassia Reddig, Katherine Rivera-Franco, Joohee Won, and Cassandra Geraty for their contributions to this project. This project received no external funding. The Language, Attention, and Cognitive Engineering lab appreciates the internal support from the Department of Psychology, San Francisco State University.

Kenneth R. Paap played a lead role in conceptualization, formal analysis, supervision, and writing–original draft and an equal role in data curation, methodology, and validation. John Majoubi played a lead role in data curation and investigation, a supporting role in supervision and writing–review and editing, and an equal role in formal analysis, methodology, project administration, and software. Regina T. Anders-Jefferson played a supporting role in data curation and writing–review and editing and an equal role in formal analysis and software. Rin Iosilevsky played a supporting role in writing–review and editing and an equal role in investigation. Charlotte Ursula Tate played a supporting role in validation and writing–review and editing and an equal role in formal analysis.

Correspondence concerning this article should be addressed to Kenneth R. Paap, Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, United States. Email: kenp@sfsu.edu

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Definition of Key Terms

In Table 1, we present conceptual and operational definitions of several key terms. In the vast literature on the control of thoughts, emotions, and behavior, the terms executive-functioning, self-control, and AC are sometimes used synonymously and sometimes to signal important distinctions. The table clarifies how we use these terms in this article and also mentions influential researchers associated with some of the terms.

Executive functions (EF) are conceptually defined as the cognitive ability to control thoughts, emotions, and behavior in goal-directed behavior. Throughout the article, we will reserve EF to refer to performance-based measures of EF. The operational definition for our study is a latent variable based on the five measures described in the Method section. Self-control (SC) is conceptually defined as the ability to control thoughts, emotions, and behaviors in the face of temptations and impulses that distract from goal attainment. Although there is a slight difference in emphasis compared to EF, the theoretical framework guiding the design of our study is that EF and SC are trying to capture the same cognitive ability. However, it is useful to use the distinctive term SC, so that it is always clear that SC refers to a self-rating measure and EF to a performance-based measure. One might also note that many of the commonly used self-rating scales for assessing this ability refer to it as self-control, for example, the highly cited and influential Brief Self-Control (BSC) Scale developed by Tangney et al. (2004).

The Engle-lab (Burgoyne et al., 2023; Draheim et al., 2021) uses the term attention control (AC) rather than EF. AC refers to the domain-general ability to regulate information processing in service of goal-directed behavior, and in this respect, it has the same conceptual definition as EF. However, AC is treated as a unitary ability (despite often referring to maintenance and disengagement as two complementary functions) that makes it distinct from the separable components of the early Miyake et al.'s (2000) model. As discussed in detail in the Method section, we operationally define performance-based EF as a single latent variable. Arguably the better label for our performance-based EF might be performance-based AC, but this discussion clarifies exactly what is meant by performance-based EF.

Finally, Bialystok and Craik (2022) and Bialystok (2024) also refer to a cognitive ability that adapts to bilingualism as AC. Although its conceptual definition is very similar to the Engle construct, it is useful to refer to it as AC2 to make it clear that it has not been operationalized as a latent variable in the manner of Draheim et al. (2021) or Burgoyne et al. (2023).

A Brief Review of the Meta-Analyses

Seven recent meta-analyses of the bilingual-advantage hypothesis converge on the same conclusion as might be discerned from either the original publications (Donnelly et al., 2019; Grundy & Timmer, 2017; Gunnerud et al., 2020; Lehtonen et al., 2018; Lowe et al., 2021; Monnier et al., 2022; Paap, 2019; von Bastian et al., 2017) or from systematic reviews of the meta-analyses themselves (Paap, 2023; Paap, Majoubi, et al., 2024; Paap, Mason, et al., 2020). Although Grundy and Timmer (2017) may consider their results somewhat different, the other six conclude that the overall bilingual advantage is very small and that when corrected for publication and other biases by precision effect test or precision effect estimate with standard errors, the results are not distinguishable from 0 and in some cases now trend in the direction of monolingual advantages. Another telltale pattern across the meta-analyses is that, for the most part, the significant moderators show that bilingual advantages are more likely to occur when either study quality is low (in a variety of ways ranging from small sample sizes to failure to measure or match potentially confounding variables) or when conducted by Ellen Bialystok's lab (or by her former students or post docs). In contrast, to the lore often surrounding this debate, there are very few instances of significant moderation across different components of EF, different ages, or different types of bilinguals (e.g., early vs. late). This dovetails nicely with Gobet and Sala's (2023) recent second-order meta-analyses showing no evidence for cognitive training (e.g., working-memory, video gaming, exergames, chess, music training), when active control-groups are part of the study design. As suggested in the title of

Table 1
Definitions of Key Terms

Term	Construct	Operational
Executive functions (EF)	The cognitive ability to control thoughts, emotions, and behavior in goal-directed behavior.	Latent variable based on performance in the three squared tasks, a cued switching task, and an antisaccade task.
Self-control (SC)	The ability to control thoughts, emotions, and behaviors in the face of temptations and impulses that distract from goal attainment.	Latent variable based on self-rating scales: BSC, BDEFS, ACS, premeditation, perseverance, and urgency.
Bilingualism (Bi)	The ability to speak at least two languages.	Latent variable based on self-reports of number of languages spoken, proficiency in L2, L2/L1 proficiency ratio, balance of use, switches per day, and "entropy."
Attention control (AC)	Engle-lab term similar to EF. Refers to the domain-general ability to regulate information processing in service of goal-directed behavior. It highlights two complementary functions: maintenance and disengagement.	Latent variables based on Draheim et al.'s (2021) toolbox or Burgoyne et al.'s (2023) squared tasks.
Attention control 2 (AC2)	Bialystok and Craik (2022): The ability to maintain goals in an active state, to suppress interference, and switch processing resources when beneficial.	No latent variables demonstrated.

Note. BSC = Brief Self-Control; BDEFS = Barkley's Deficits in Executive Functioning Scale; ACS = Attention Control Scale.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Paap, Majoubi, et al. (2024), *it all fits together.* That is, if EF is highly heritable or if cognitive skills are acquired to high levels primarily through gains in task-specific automaticity (rather than recruitment and enhancement of domain-general EF), then no type of cognitive training will lead to benefit.

Are the Meta-Analyses Compromised by Poor Measure of EF?

Although there is a fair amount of diversity across the meta-analyses, all the included studies used performance-based measures of EF that typically measure speed, accuracy, or some combination of the two. Unfortunately, these measures are often psychometrically inadequate. The most important problem is a lack of convergent validity between purported measures of the same construct. This lack of convergent validity occurs frequently but convergence does pop up on occasion. For example, the seminal work by Miyake et al. (2000) showed that three quite different inhibitory-control tasks (antisaccade, stop-signal, Stroop) loaded on a latent variable (.57, .33, .44, respectively) and that the latent variables for shifting, updating, and inhibition showed intercorrelations of .56, 63, and .43. In contrast, it has been known for decades that the letter and arrow versions of the flanker task do not correlate with each other (Salthouse, 2010), $r = +.03$. Likewise, six pair-wise correlations between four versions of the Stroop task were nonsignificant and ranged from −.13 to +.22 (Shilling et al., 2002). Rey-Mermet et al. (2018) reviewed and reported newer latent-variable analyses of EF that led them to conclude that inhibitory-control measures “do not measure a common, underlying construct but instead measure the highly task-specific ability to resolve the interference arising in each task.” For them, the “… inevitable implication is that studies using a single laboratory paradigm for assessing inhibition do not warrant generalizing beyond the specific paradigm studied” (p. 15). Proponents of the bilingual-advantage hypothesis might reasonably suggest that many of the null results captured in the meta-analyses may not have used valid measures and, consequently, did not test for differences in domain-general EF.

It merits mentioning that the cause, or an additional cause, of the lack of convergent validity is that many measures are based on reaction time (RT) difference scores. For example, inhibitory control ability is typically inferred from a flanker task by taking the difference in mean RT between the easy congruent trials where the irrelevant flanker information supports the correct response and the more difficult incongruent trials where the flanking information creates a conflict that needs to be resolved. Subtracting the mean RT on the congruent trials from the incongruent trials is very attractive (since Donders, 1969 introduced his subtraction method), because it provides a possible way of canceling out the shared encoding and response processes and isolating the targeted conflict–resolution (inhibitory control) process. However, as explained and documented by Paap and Sawi (2016) and Hedge et al. (2018), the correlation between any two difference-score measures is capped by the lower of the two test–retest reliabilities. Furthermore, the test–retest reliability of the differences between congruent and incongruent trials in any of the commonly used nonverbal interference tasks (e.g., Simon, flanker, spatial Stroop) is usually inadequate, because the two trial types tend to correlate highly and leave little systematic variability in the wake of the subtraction.

The Engle Lab Fixes the Psychometrics of EF Measures

The Engle lab has developed two sets of performance-based measures with strong psychometric properties (Burgoyne et al., 2023; Draheim et al., 2021). Their theoretical framework is different from Miyake and Friedman’s (2012) unity and diversity approach to EF. AC, their preferred name for the construct, is construed as being more unitary than componential. As recently advocated by Brysbaert (2024), the primary tool that the Engle lab has been using is to sift and winnow through new and existing measures of AC to identify a set (e.g., antisaccade, flanker deadline, visual arrays) that strongly loads on a latent variable, thereby demonstrating convergent validity. Beyond establishing a robust latent variable, the Engle collaborative also demonstrated that AC[1] plays a fundamental role in driving the relationship between two related cognitive abilities that each load on their own latent variable: working-memory capacity (WMC) and general fluid intelligence (gF). In isolation these latent variables correlate with each other and invite the interpretation that WMC directly supports gF. A structural model that includes AC can statistically mediate the relationship between WMC and gF, if the latent variable for AC is based on measures with good reliability and convergent validity. This implies that AC was the *active ingredient* such that the covariance between WMC and gF is caused by the influence of AC on both.

A more recent advance in the Engle toolkit (Burgoyne et al., 2023) introduced a trio of modified *squared* tasks that were based on the classic flanker, Simon, and Stroop tasks that relied on the logic of subtraction and, hence and unfortunately, the unreliable difference score. The ensemble of modifications start with eliminating RT as a dependent variable and gamifying the task by instructing participants that they could earn one point for each correct response (and be penalized one point for each error) in a 90-s game. The games are called *Squared*, because they include a double shot of congruency on every trial. Each trial displays one target at the top and two choice alternatives in the bottom row. For example, in Flanker Squared, the target array at the top may look like a typical incongruent trial in an arrow flanker task, < < > < <. But the player must pay attention to the flanking arrows (e.g., pointing left) and ignore the conflicting center arrow. The bottom row has two sets of arrows: { < < > < < } { > > < > > }. What makes the squared task so devilish is that the relevant and irrelevant information switches as the player moves from encoding the target on the top to selecting the designated response in the bottom row. That is, the *flanking arrows* from the top, must be matched to the *centered arrow* on the bottom right. The correct response for this trial is to push the right key.

The *Squared* tasks have excellent psychometric properties (Burgoyne et al., 2023). The test–retest reliabilities ranged from +.53 to +.75. The intertask correlations ranged from +.50 to +.53. The factor loading on the AC latent variable ranged from +.66 to +.71. The *Squared* tasks accounted for 75% of the variance in multitasking at the latent level. The toolkit developed by the Engle group also successfully predicts performance in complex real-world tasks (Draheim et al., 2022). This study will test for the effects of bilingualism on EF using the three *Squared* tasks to anchor a latent variable for EF.

[1] We treat *attention control* (AC) and *executive functioning* (EF) as synonyms for the domain-general construct that is typically the intended target of performance-based measures. When discussing Engle’s theory or results *attention control* is used in deference to their preference.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

The Mysteries of Measuring Bilingualism

The null results dominating the meta-analyses may have been further nurtured by the fact that many of the tests for bilingual advantages treated bilingualism as a dichotomy. That is, monolinguals with little or no exposure to a foreign language (L2) are pitted against bilinguals who regularly use two languages. If any aspect of bilingual language control recruits domain-general EF, then these two groups should differ markedly on these critical aspects. The following makes the argument as to why this is likely to be true.

If a bilingual estimates the percentage of time they use each language, then 100 minus the percentage of the most used language provides a simple measure of *balanced use.* A bilingual who uses each language half the time has a score of 50 (100–50). In contrast, a bilingual who uses his dominant language 75% of the time has a *balance use* score of only 25 (100–75). The bottom of the scale is anchored at 0 by pure monolinguals: 100–100 = 0. A strength of the *balance use* score is that it is based on straightforward questions that participants find easy to answer (viz., *What percentage of the time do you speak English?*). As described in more detail later, the *balance use* scale correlates strongly with five other dimensions of bilingualism: the ratio of language proficiencies (r = +.54), the number of languages spoken (r = +.71), "entropy" across seven contexts (r = +.74), language switches per day (r = +.77), and the reverse score on a test of productive English vocabulary (r = +.45). The point of this exercise is to show that most measures of bilingualism correlate with one another (at least in the heterogeneous language population of San Francisco) and, as shown later, coherently load on a latent variable. If multiple dimensions of bilingualism are active ingredients (capable of enhancing EF), then there should be a strong likelihood of showing benefits at the latent-variable level. In contrast, if there is only one idiosyncratic control process that is domain-general and strengthens like a muscle when exercised, then tests at the latent level will not be a panacea. We note in passing that most theoretical frameworks for investigating the effects of bilingualism on domain-general EF: (a) presume that the control processes are general-purpose but provide no evidence that they are not task-specific and (b) do not speculate as to what mechanism might be generating a domain-general benefit: for example, an increase in general processing capacity (Kahneman, 1973).

The Paradoxical Absence of a Relationship Between EF and SC

The meta-analyses of bilingual advantages in EF reviewed above rely exclusively on performance-based measures based on speed, accuracy, or some composite. However, vast tracts of published studies refer to a very similar construct as SC and measure it subjectively using self-report Likert scales (e.g., Paap, 2023; Paap, Anders-Jefferson, et al., 2024). Encouragingly, (a) the SC scales tend to correlate with one another, showing convergent validity (e.g., Tangney et al.'s brief self-control [BSC], 2004 and Barkley's deficits in executive functioning [BDEFS], 2011), (b) positively correlate with desirable outcomes (e.g., academic achievement) and (c) negatively correlate with undesirable outcomes (e.g., overeating). They also have the intuitive appeal of face validity as illustrated, for example, by these two items from the BSC: (a) *I am good at resisting temptation* and (b) *I am able to work effectively toward long-term goals.* Discouragingly, the self-report measures and performance-based measures are weakly correlated at best (Duckworth & Kern, 2011; Friedman et al., 2020; Mason et al., 2021; Necka et al., 2012; Paap, Anders-Jefferson, et al., 2020; Stahl et al., 2014). The critical point for the design of this study is that the absence of consistent and nontrivial correlations between performance-based measures of EF and either bilingualism or self-report measures of SC may have been driven by the inadequate psychometric properties of the performance-based measures. Thus, using a set of EF measures with good reliability and validity that successfully loads on a latent variable may reveal these elusive associations and breathe new life into the bilingual-advantage in EF hypothesis and the expectation that self-report measures of SC and performance-based measures of EF are tapping into the same important psychological construct. From this optimistic perspective, one predicts that, at the latent-variable level, a SEM model will show that bilingualism predicts both SC and EF which, in turn, covary with one another.

Necka et al. (2012)

A precursor to our study was reported by Necka et al. (2012) who extracted latent variables for performance-based measures of EF, self-rating measures of SC, and fluid intelligence. bilingualism was not examined. As anticipated for our study, most of the SC measures strongly loaded ($r \geq +.81$) on the SC latent variable, but that latent variable had no relationship to performance-based EF (e.g., Stroop, N-back, stop-signal), r = +.01. Although the five performance-based measures each loaded on the latent variable, the associations ranged from modest (r = .26) to moderate (r = +.63). Nevertheless, the regression coefficient between the EF and fluid-intelligence latent variables was a robust r = .74. Apparently, the EF abilities that predict fluid intelligence do not predict SC, consistent with the possibility the EF and SC latent variables are tapping into different psychological constructs.

Stahl et al. (2014)

The study by Stahl et al. has impressive scope. The focus was on interference control rather than a broader conception of EF. Their six-factor model includes separable factors for interference emanating from: (Factor 1) memory, (Factor 2) competing stimuli, (Factor 3) competing responses, and (Factor 4) the inhibition or cancelation of responses that have already been selected or even initiated. The latent variable of information sampling (Factor 5) was cleverly derived from two tasks by using Ratcliff's diffusion model to estimate the response–caution criterion (Ratcliff & Rouder, 1998). For each of the two indicator tasks a score was obtained from each participant that reflected the degree to which they made decisions rapidly and based on little evidence or slowly and only after accumulation of considerable evidence. Delay discounting (Factor 6) assumed that impulse control is closely related to motivational processes such as delay of gratification (Mischel et al., 2011). A familiar pattern was repeated in the Stahl et al.'s study. The Urgency, Premeditation, Perseverance, Sensation-Seeking scale was not related to any of the six latent factors derived from performance-based tasks! "There was no evidence for a relation between the present behavioral impulsivity factors and self-reported impulsivity" p. 867.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Friedman et al. (2020)

In another landmark study, Friedman et al. (2020) examined the relationship between the three latent variables in their unity and diversity model (viz., common EF, shifting, and updating) and five facets of the Urgency, Premeditation, Perseverance, Sensation-Seeking-Positive Urgency scales. Consistent with the results reviewed earlier, the latent factors (based on objective performance-based measures) and self-report measures were weakly related.

Because their two data sets came from "twin" studies that included monozygotic and dyzygotic pairs, Friedman et al. (2020) could partition the variance associated with each EF factor and each impulsivity dimension into: genetic, shared environmental, and nonshared environmental variances. Provocatively, self-report and performance-based measures of impulsivity dissociated with respect to the relative influence of genetics. Consistent with their earlier work (Friedman et al., 2008), they report that the individual differences in performance-based EFs were almost entirely genetic in origin (>80%), while the individual differences in the self-report impulsivity scales ranged from 21% to 45%.

There are two quite important implications for this dissociation. First, it confirms the substantial body of evidence already presented that latent variables for SC and EF are tapping into different psychological constructs. Second, if high-quality performance-based measures can be identified and if they continue to show no relationship to bilingualism, then the literature on brain-training and the bilingual-advantage will all fit together (Paap, Majoubi, et al., 2024). To wit—in their second-order meta-analysis (based on 10 different meta-analyses), Gobet and Sala (2023) showed that all variants of brain-training yield effects indistinguishable from 0 when considering only studies using active control (i.e., controlling for placebo effects) and when correcting for publication bias. If an ability is highly heritable, then there are limited opportunities for experiences or activities to alter the individual differences initially constrained by genetics. Gobet and Sala conclude that the effects of brain training have led "… to a crystal-clear conclusion: The overall effects of far transfer is null and there is little or no true variability between the types of cognitive training" (p. 125). If individual differences in cognitive ability are highly heritable, then experiences or activities that many people pursue in everyday life, for example, performing music, playing chess or video games, navigating a taxi, controlling air traffic, or a class full of kindergarteners will not lead to measurable advantages in the abilities. Why should managing two languages fair differently?

This is not to say that individual differences in specific domains are not great or that professionals and experts are born and not made. We close this section by reminding readers that skills are acquired in stages whereby early stages involve the slow and intentional recruitment of domain-general control processes that give way to new task-specific structures that are fast and automatic (Chein & Schneider, 2012). In a discussion of his controlled dose hypothesis, Paap (2018) argues that it may make more sense to predict bilingual-advantages in domain-general EF in the early stages of acquiring an L2, when domain-general EF is actually needed and recruited.

The Contribution of Neuroscience to the Debate

It has become increasingly common for studies to include neuroscience measures and to focus on evidence of brain plasticity as a consequence of immersion in a language community. When we (Paap et al., 2014) first started to consider the neuroscience data surrounding the bilingual advantage we were surprised to discover serious issues in what we termed alignment, valence, and kind ambiguity. Alignment refers to the fact that differences between bilinguals and monolinguals in neural-data frequently do not align with the behavioral differences. Valence refers to frequent disagreements regarding whether "higher" neural scores imply better or worse cognitive functioning (see Cespón & Carreiras, 2020, for several event-related potential examples). Kind ambiguity refers to disagreements regarding what type of cognitive events the pattern of neural activation in a region-of-interest actually reflects. Our kind ambiguity echoes the "reverse inference" problem elegantly introduced by Poldrack (2006).

The alignment problem continues to receive attention. We agree with de Bruin et al. (2021), when they "… argue that the move from behavioral studies to a focus on brain plasticity is not going to solve the debate on cognitive effects, especially not when brain changes are interpreted in the absence of behavioral differences" (p. 433). Similarly, García-Pentón et al. (2016) asserted that "supporting evidence for an advantage should involve showing that these differences are accompanied by unambiguous behavioral data substantiating a cognitive gain."

As de Bruin et al. (2021) asserted, the interpretation of the direction of activation differences becomes "even more difficult," when they are accompanied by a behavioral disadvantage. Consider the surprising case reported by Mohades et al. (2014) who scanned both bilingual and monolingual children, when they were performing a Stroop task and a Simon task. In contrast to Bialystok et al. (2004) and Coderre and van Heuven (2014), bilinguals in the Mohades et al. study performed significantly worse than the monolinguals.

In summary, bilingualism induces changes to brain structure and function, but their interpretation requires alignment with simultaneous behavioral measures. To date, misalignments clearly outnumber alignments. As noted by one reviewer, such failures of alignment prohibit any strong interpretation of language-dependent changes in brain morphology as evidencing a bilingual advantage in EF, and the existence of such changes should not be taken as evidence contradicting the meta-analyses reviewed earlier or the results of this study reported later.

Method

San Francisco State University students were recruited to participate in a two-part study in exchange for course credit. The first part was completed online via a Qualtrics survey, while the second part required completing the performance-based measures in a university laboratory. The survey included basic demographics, detailed questions about language experience, five self-report scales of SC, four self-report scales of quality of life, a measure of general fluid intelligence (gF), a scale of social desirability, and two quality-of-responding probes that were first used in Paap, Anders-Jefferson, et al. (2024). Of the 353 completed surveys, 23 (6.5%) were eliminated, because the participants indicated that they did not always pay attention to the items, and 3 (0.8%) indicated they did not always answer honestly. One participant failed both checks. Thus, the measures derived from the survey were based on 328 participants. Although there were very little missing data, participants could skip an item if they would prefer not to answer.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

The use of human participants was approved by the San Francisco State University Institutional Review Board. We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Participants

Bilingualism Measures

The effects of bilingualism were evaluated as both a grouping variable and a continuous variable in regression and SEM models. The measures or facets of bilingualism are shown in Table 2. If a participant indicated no exposure to any L2, they were assigned a proficiency score of 0. For each language a participant had contact with, self-ratings of proficiency were obtained on the separate dimensions of speaking, listening, reading, and writing on a scale of 1–7 using the scale developed and shown by Paap and Greenberg (2013). Each of the seven points was labeled with a description referring to the range of topics and comparison group, for example, *Level 4 Advanced Intermediate: Can converse with little difficulty with a native speaker on most everyday topics, but with less fluency than a native speaker*.

A participant was considered bilingual, if the mean of their self-rated proficiency in comprehending and speaking a non-English language was at least 3.5 but was assigned to the ambiguous group, if they used their L2 less than 10% of the time. A participant was considered monolingual, if their self-rated proficiency in comprehending and speaking a non-English language was 2.0 or less but was assigned to the ambiguous group, if they used an L2 more than 10% of the time.

As shown in Table 2, this resulted in 135 bilinguals, 120 monolinguals, and 73 in the ambiguous group. The mean proficiency in English for monolinguals was 6.42, about midway between *as fluent as a native speaker* and *super fluent*, but their proficiency in a non-English language was near 0 ($M = 0.29$). In contrast, for bilinguals the mean proficiency in English was 6.06 compared to 5.76 in their non-English language. Further, for bilinguals, the P2/P1 ratio was high ($M = .89$), the mean number of language switches per day was *about a dozen*, and the less used language was typically used about one-third of the time. There were few "bilinguals" who used more than two languages ($M = 2.11$). In an entropy-like measure, bilinguals indicated the number of languages used in each of seven contexts: where you live, with family, with friends, at the university, at work, in your community, and in the media you use. The mean number used by the bilinguals was 1.66, which is very close to the average when a bilingual indicates two languages for five of the seven contexts. In summary, the bilingual group is strikingly different from the monolingual group, and their profile approximates what Green and Abutalebi (2013) described as a dual-language context.

Demographics

Table 3 compares bilinguals to monolinguals on several demographic variables. There were no statistically significant differences in age, Raven's advanced matrices test of general fluid intelligence (Raven et al., 1977) or the proportion of males. socioeconomic status (SES) was captured with four measures: *Which of the following best describes the highest educational level obtained by your* (1) *father* or (2) *mother?* Responses were selected from an 8-point Likert scale ranging from 1 (no formal education) to 8 (graduate or professional degree); (3) *relative to others in the country you currently reside, what socioeconomic class do you identify with?* (4) *Relative to other families in the country I grew up in, my family's income would be considered:* Responses to (3) and (4) were selected from a 5-point Likert scale ranging from 1 (low) to 5 (high). A composite measure of SES was calculated by taking the mean of the standardized scores. Mother's education, father's education, current SES, and the composite SES scores all showed significant monolingual advantages, but there were no significant differences in childhood SES.

Although this gives the impression that bilinguals may have been disadvantaged in comparisons of SC or EF, the composite measure of SES did not significantly correlate with any of the five measures of self-rated SC or any of the five performance-based measures of EF. For example, the correlation between the composite SES and the BSC scores was $r = -.056$ and with the Stroop Square was $r = +.079$. Given that SES is a robust correlate of childhood EF ability, this may seem surprising, but Paap's (2023) recent review of the relationship between SES and EF showed that the association surprisingly disappears in most samples older than 20 years of age.

A much higher proportion of individuals in the bilingual group indicated that they were not born in the United States and, in that

Table 2
Language Characteristics of the Three Groups

Language characteristic	Bilingual	Monolingual[a]	Ambiguous
N (number of participants)	135	120	73
English proficiency	6.06	6.42	6.22
Non-English proficiency	5.76	0.29	2.87
P1 (proficiency in best language)	6.47	6.42	6.30
P2 (proficiency in next best language)	5.72	0.28	2.81
P2/P1 ratio	0.89	0.05	.45
Entropy (mean languages per seven contexts)	1.66	1.04	1.40
Switches per day (7-point Likert: 1 = *none or rarely*, 2 = *a few times*, 3 = *about a dozen*, 4 = about two dozen …)	3.13	0.28	1.55
Balance of use (100 − % of most used language)	33.6	0.6	9.3
Number of languages used now	2.11	1.03	1.42
Age-of-acquisition of L2 (years old)	4.8	10.5 (n = 26)	4.9
Years speaking L2	16.9	9.5 (n = 26)	17.1

[a] 94 monolinguals indicated no exposure to a foreign language.

Table 3

Demographics of Bilinguals and Monolinguals

Demographic	Bilingual	Monolingual	Test
Age	21.7	21.2	$t(249) = 0.80$, $p = .212$
Mother's education	4.9	6.2**	$t(251) = -6.08$, $p < .001$
Father's education	4.9	5.8**	$t(250) = -3.91$, $p < .001$
Childhood SES	2.7	2.9	$t(251) = -1.63$, $p = .052$
SES now	2.5	2.8**	$t(250) = -3.06$, $p = .001$
SES standardized composite	−.22	+.22**	$t(252) = -4.88$, $p < .001$
Ravens matrices (gF)	9.1	9.4	$t(250) = -0.98$, $p = .165$
GPA	3.4	3.4	$t(240) = +0.52$, $p = .302$
Immigrant proportion	.37**	.07	$z = 5.77$, $p < .001$
Male proportion	.21	.25	$z = -0.75$, $p = .23$

Note. SES = socioeconomic status; GPA = grade point average.
$^{**}p < .01$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

sense, were immigrants. Any "healthy immigrant effect" (Vang et al., 2017) would favor the bilingual group.

Self-Rating Measures of Self-Control

Five commonly used self-rating measures of self-control were included: (a) the BSC Scale (Tangney et al., 2004), (b) the Urgency, (c) Perseverance, and (d) Premeditation subscales of Whiteside and Lynam's (2001) Urgency, Premeditation, Perseverance, Sensation-Seeking scale of impulsivity, and the (e) Attention Control Scale (ACS, Derryberry & Reed, 2002). The overall ACS consists of a nine-item subscale labeled Focusing (e.g., "My concentration is good even if there is music in the room around me") and an 11-item subscale labeled Shifting (e.g., *It is easy for me to read or write while I'm also talking on the phone*). Each subscale has five or six reverse-coded items. The overall ACS scale is internally consistent (α = .99), positively related to positive emotionality ($r = +.40$) and inversely related to negative emotionality ($r = -.55$).

The BSC is widely used, and the seminal article by Tangney et al. (2004) has more than 9,000 Google-scholar citations. It enjoys a remarkable record in correlating positively with desirable outcomes and negatively with undesirable outcomes in everyday life. The BSC consists of 13 items including 4 that are positively worded (e.g., *I am able to work effectively toward long-term goals.*) and 9 that are reverse scored (e.g., *I have a hard time breaking bad habits.*). Tangney et al. (2004) reported that the original BSC enjoyed excellent reliability in two large samples of college students. Cronbach's α showed an internal consistency of .83 ($N = 351$) in Study 1 and .85 ($N = 255$) in Study 2. Test–retest reliability over a 1- to 3-week period was measured in Study 1 and yielded an impressive $r = 0.87$.

The Whiteside and Lynam (2001) impulsivity scales have more than 5,000 Google-scholar citations and each consists of 10 to 12 items (e.g., *My thinking is usually careful and purposeful* for Premeditation, *I have trouble controlling my impulses* for Urgency, and *I finish what I start* for Perseverance). No more than two items in any one scale need to be reverse-coded. The internal consistency coefficients for the three scales ranged from .82 to .91.

As systematically reviewed by Paap (2023), self-rating scales of self-control show good convergent validity with one another but correlate weakly with performance-based measures EF (Allom et al., 2016; Cyders & Coskunpinar, 2011; Duckworth & Kern, 2011; Friedman et al., 2020; Mason et al., 2021; Necka et al., 2018; Stahl et al., 2014). As anticipated and reported in the results, the five self-rating measures used in this study correlate with one another and load on a coherent latent variable.

Performance-Based Measure of EF

The focal performance-based tasks are the three Squared tasks developed by the Engle lab (Burgoyne et al., 2023) and described in the introduction. These three new tasks, with excellent reliability and validity, are joined with an antisaccade task and a cued color-shape switching task. The antisaccade task was selected because it has an outstanding record in generating measures of inhibitory control that load on a latent for inhibition. In fact, as Rey-Mermet et al. (2018) pointed out, not only does antisaccade performance load on these latent variables, but it also tends to dominate them. The antisaccade task was precisely the same as that used by Paap and Greenberg (2013) which, in turn, was based on the antisaccade task developed by Kane et al. (2001) who showed that individual differences in WMC predicted performance on antisaccade blocks, but not prosaccade blocks. The task on each trial was to identify the target stimulus (i.e., "B," "P," or "R"). The briefly presented target was followed by a visually similar mask ("8"). The target and mask subtended about 0.9° of visual angle. In the antisaccade block, a distractor stimulus is always blinked just before, and on the, from, the target. The distractor appeared about 2.0° to one side of fixation and the target 2° to the opposite side. Because the eventual target is always presented on the opposite side, the best strategy is to inhibit the natural predisposition to attend to (and/or saccade toward) any peripheral stimulus with an abrupt onset. Individuals with superior inhibitory control should respond faster and more accurately than those with lesser ability. Other details about trial blocks and number of trials are available in Paap and Greenberg (2013).

Switch costs (mean RT on switch trials minus mean RT on repeat trials) were derived from the color-shape task because Paap et al. (2017) showed that the Switch costs derived from three very different task dyads (viz., color-shape, letter-digit, animacy-size) significantly correlated with one another, with intertask correlations ranging from $r = .25$ to $r = .34$. Thus, it was anticipated that these five performance-based measures would show convergent validity and load on a coherent latent variable. The color-shape task was identical to that described and used by Paap and Greenberg (2013). In the critical

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

mixed block, each trial is initiated with a precue (rainbow or △○) that specifies that the upcoming target needs to be judged as green or red with two fingers of the left hand or as a triangle or circle with two fingers of the right hand. Across the block the number of repeat trials (color-color or shape-shape) was equal to the number of switch trials (color-shape or shape-color).

Raven's Advanced Matrices Test of Fluid Intelligence

Fluid intelligence (gF) was assessed using Set 1 of Raven's Advanced Progressive Matrices (Raven et al., 1977). The test consisted of 12 items. Each item was composed of a pattern with a missing piece in the lower right. Participants were instructed to *look at the pattern, think what the missing part must be like to complete the pattern correctly, both across the rows and down the columns*. Participants selected from a set of eight alternatives. The task was computerized and controlled by Qualtrics. Participants were given a maximum of 2 min to respond to each item. Most responses, regardless of correctness, in this self-paced computer-controlled version were made well within the deadline. The manual states that with self-pacing, Set 1 can be used as a short 10-min test. The 12-item test has a decent Cronbach α, for example, .81 (Partchev, 2020) and .73 (Bors & Stokes, 1998; N = 506 University of Toronto students). Arthur et al. (1999) reported a test–retest r = 0.76 for 71 participants at a 1-week interval.

Raven's scores were included in this study because our previous work (Paap, Anders-Jefferson, et al., 2020) showed that gF was the most consistent predictor of performance-based measures of EF. Furthermore, some researchers (Salthouse, 2005, 2010) have argued that EF and gF may be two names for the same ability. However, a significant correlation was not anticipated for self-rating scales used in this study, because we had never observed a significant correlation between the BSC and Raven's using samples of university students (Paap et al., 2019, r = –0.07; Mason et al., 2021, r = –0.10; Paap et al., 2022, r = +0.05). Similarly, Erceg et al. (2019) reported a correlation of r = –0.14 for a sample of 159 college students (M = 21.3 years old). Finally, Mazza et al. (2021) reported a correlation of r = +0.07 based on a sample of 522 Mturk participants (M = 33.6 years) paid considerably more than is typical ($60 plus an average of $10 in bonuses) for completing a 10-h battery of surveys and cognitive tasks. A deviation from this pattern was observed by Paap et al. (2022), because their two studies using M-Turk adults (rather than college students) showed small, but significant, correlations (r = +.26** and +.17*) between Raven's scores and BSC scores. This partial disconnect between the relationship of self-report and performance-based measures to Raven's scores is not surprising given our earlier discussion that self-ratings of SC and performance-based measures of EF do not correlate with each other.

Quality-of-Life Outcome Scales

As reviewed in Paap, Anders-Jefferson, et al. (2024), the BSC consistently correlates with Rosenberg's (1965) Self-Esteem Scale, Goldberg and Williams's (1991) General Health Questionnaire, Diener et al.'s (1985) Satisfaction With Life scale, and Lyubomirsky and Lepper's (1999) Subjective Happiness Scale. Paap, Anders-Jefferson, et al. (2024) also cited studies showing that these scales enjoy good internal consistency and test–retest reliability. The Self-Esteem Scale consists of 10 items (e.g., *I feel I am a person of worth*), of which, five need to be reverse scored (e.g., *I feel I do not have much to be proud of*). The General Health Questionnaire consists of 12 items (e.g., *Have you been able to concentrate well on what you were doing?*), of which, six need to be reverse scored (e.g., *Have your worries made you lose a lot of sleep?*). The classic Satisfaction with Life Scale consists of four items (e.g., *In most ways my life is close to my ideal.*). Self-rated Happiness was measured with four items (e.g., *In general, I consider myself: 1* "not a very happy person" *to 7* "a very happy person."

Social Desirability

As noted above, self-rating scales of SC and performance-based measures of EF are weakly correlated. Among several important differences (see Paap, Anders-Jefferson, et al., 2024 for a more detailed review), self-ratings are more vulnerable to response biases stemming from acquiescence, social desirability, and other sources that may differentially inflate the self-control scales of individuals with certain personality types or groups with different cultural values or occupying different positions in the socioeconomic hierarchy.

In our earlier work (Paap, Anders-Jefferson, et al., 2024), Stöber's (2001) Social Desirability Scale (SDS) was primarily used as a covariate when treating BSC as a predictor of the outcome variables described above. SDS scores will correlate with BSC scores to the extent that an individual is biased to over report on items with positive valence or under report on those with negative valence. The original SDS has nine positive-valence items (*I always admit my mistakes openly and face the potential negative consequences*) and seven with negative valence (*I sometimes litter*). One point is scored for each *true* response to a positive item and one point for a *false* response to a negative item. Thus, larger totals indicate a greater bias to give socially desirable answers. Stöber reported test–retest correlations over 0.80 across intervals from 2 to 6 weeks and Cronbach αs of either .74 or .75 across three college student samples and a large community sample.

Prior studies show strong correlations between BSC and social desirability scores. Bertrams and Dickhäuser (2012) reported r = +0.46 for a sample of 150 undergraduates. Similarly, Kwapis and Bartczuk (2020) reported r = +0.45 for a sample of 141 adolescents (M = 17.7 years old). Uysal and Knee (2012) reported similar correlations when social desirability was measured with the Marlowe Crowne Scale: r = +0.43 for 160 undergraduates in Study 1, r = +0.59 for 74 undergraduates in Study 2, and r = +0.51 for 55 undergraduates in Study 3. Collectively, these substantial correlations highlight the possibility that positive correlations between BSC scores and other desirable outcomes may be mediated by social desirability.

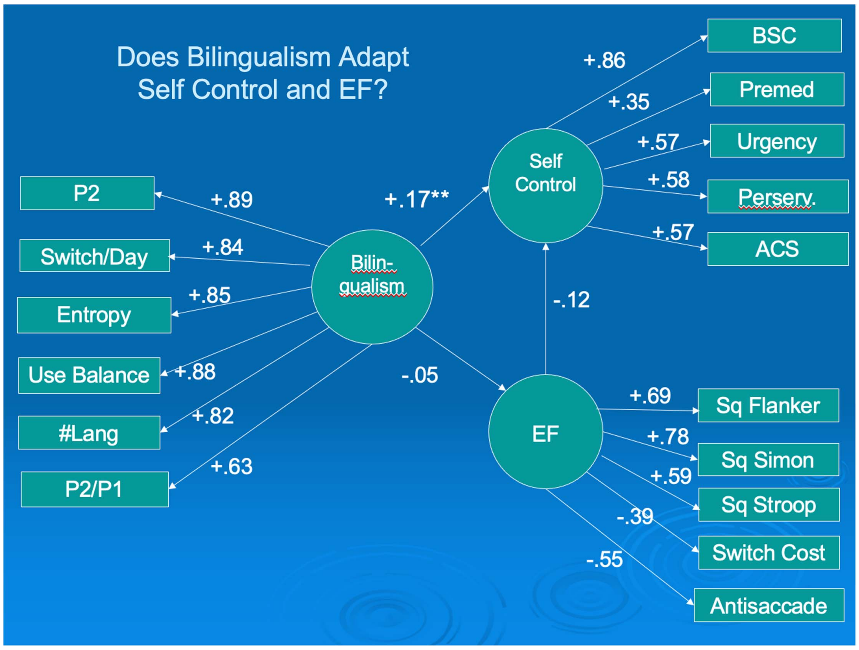
Transparency and Openness

Our primary research question is instantiated in the SEMs shown in Figures 1 and 2 that focus on the latent variables for bilingualism, SC, and performance-based EF. The sample size, especially for the laboratory measures, was determined by Kline's (2016) recommendation that a sample size of 200 should be regarded as a minimum threshold for SEM. The JASP datafile used for the SEM and Bayes factor (BF) analyses are available at https://osf.io/j8gvh/?view_only=9ffe42c84278487d9f541efd3ef1a44c. The study's design and

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 1

A SEM That Assumes That Bilingualism Adapts and Enhances EF and Self-Control at the Latent-Variable Level



Note. It further assumes that the EF construct represents an ability that directly affects self-control in everyday life. EF = executive function; SEM = structural equation model; BSC = Brief Self-Control; ACS = Attention Control Scale; P1 = proficiency of most proficient language; P2 = proficiency of second-most proficient language. See the online article for the color version of this figure.

its analysis were not preregistered. The SEM model based on a smaller sample size was presented at the 64th Annual Meeting of the Psychonomic Society.

Results

Self-Rating and Performance-Based Measures of Control

Square-Task Correlations

Task-level correlations are presented in Table 4. As can be seen, the three *squared* tests correlated very highly with each other (mean $r = .486$), correlations ranged from $r = .419$ to $r = .539$, demonstrating convergent validity. The values were very similar to those reported by Burgoyne et al. (2023).

Exploratory Factor Analyses

Because we are keenly interested in assessing the degree to which self-rating scales and performance-based tasks measure the same construct we started with an exploratory factor analysis (EFA) of the 10 control measures shown in Table 5 to directly understand the underlying variance–covariance matrix across all measures. Consistent with Burgoyne et al. we used principal axis factoring with an oblique promax rotation and pairwise deletion. As shown in Figure 3, a parallel analysis with simulated random-data prescribed the extraction of two factors.

As shown in Table 5, this EFA mapped perfectly onto the five self-rating scales (Factor 2) and the five performance-based tests (Factor 1). The EFA shows that there is no better partition of the 10 tasks for distinguishing the two factors.

Does Bilingualism Yield a Coherent Latent Variable?

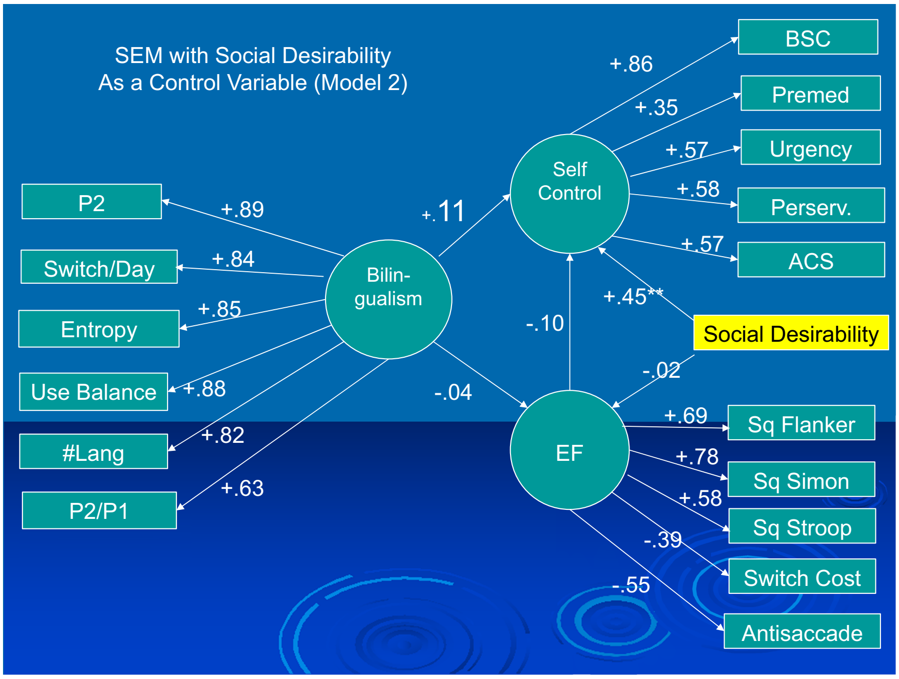
The main purpose of this study was to test the hypothesis that bilingualism enhances performance-based EF and/or self-rated SC at the latent variable level using psychometrically strong measures. But what might a coherent latent variable for bilingualism capture? It demonstrates that bilinguals tend to have a similar profile across the six measures. Failure to cohere into a single factor would indicate that there are different types of bilinguals in the population that was sampled. In contrast, if a coherent latent variable emerges, then it can capture individual differences in the overall intensity of bilingual experience. This is the underlying logic of studies that test for bilingualism effects using some type of composite score based on multiple continuous dimensions of bilingualism.

An additional complication emerges if some aspects of managing two languages recruit domain-general components of EF, but others are specific to the language module. To take a concrete, but somewhat arbitrary example, suppose that language switching recruited and strengthened a domain-general switching mechanism, but that selection of words in the target language was governed by language-specific inhibitory connections from language nodes

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 2

A SEM That Controls for the Effects of Social Desirability



Note. The social desirability measure predicts self-control and leads to the attenuation of the regression coefficient from bilingualism to self-control. SEM = structural equation model; BSC = Brief Self-Control; ACS = Attention Control Scale; P1 = proficiency of most proficient language; P2 = proficiency of second-most proficient language. See the online article for the color version of this figure.

to specific lexical representations. Under this scenario, facets (measures) of bilingualism affected by L2 proficiency and age-of-acquisition (AoA) of L2 may have played critical roles in managing the competition between the lexicons, but would have made no contribution to enhancing domain-general EF. However, if L2 proficiency and AoA correlate with dimensions like frequency of language switching and "entropy," they will load on the latent variable and only add noise to the ability of the bilingualism latent variable to predict EF. For this reason, our SEM analyses are supplemented by regression tests of individual measures of bilingualism.

An EFA of Bilingualism Measures

To examine the underlying variance–covariance structure of the measures used, an EFA was done using eight of the variables that have face validity in capturing facets of bilingualism: self-rated proficiency in speaking and listening to an L2 (P2), the ratio of L2 to L1 proficiency (P2/P1 ratio), the number of languages used now (NumUsedNow), an "entropy" measure calculated as the mean number of languages used in each of seven context (Entropy), balance of use (BalanceUse), number of switches per day (SwithPerDay), age-of-acquisition of L2 [L2_AoA]), and years speaking L2 (YearsL2). Principal axis factoring with an oblique promax rotation and pairwise deletion was used. As shown in Figure 4, a parallel analysis with simulated random-data prescribed the extraction of two factors.

Table 4

Bivariate Correlations Between the Square Measures

Task	1	2	3
Flanker Square	—	.505**	.539**
Stroop Square		—	.419**
Simon Square			—

** p < .001.

The path diagram is shown in Figure 5. Five of the six measures loading on Factor 1 have loadings greater than .82. The lowest loading is for the P2/P1 ratio at .61. The only two measures that load on Factor 2 reflected the onset (L2_AoA) and duration of bilingualism (Years L2). The loadings were +.85 and −.83. The correlation between the two factors was −.23.

Having only two measures for a latent variable can raise concerns about reliability, validity, generalizability, dimensionality, and statistical power. These possible concerns, coupled with a desire to keep our focal SEM model simple, led us to explore L2 AoA and Years L2 as separate predictors of each of the 10 measures of control. All 20 correlations had $r < .11$, and despite large *N*s, none came close to significance at $p < .05$. Using JASP, we then conducted parallel Bayesian correlations that calculated BF_{01}. All 20 BFs favored the null hypothesis, and, in fact, all exceeded the guideline of $BF_{01} = 3.0$ as substantial evidence for the null. The mean of all 20 BF_{01} ($M = 8.0$) indicates that, on average, there is eight times more evidence for the null than the alternative hypothesis across these 20 correlations.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 5

EFA Factor Loadings for 10 Measures of Self-Control/Executive Functioning

Measure	Factor 1 performance-based	Factor 2 self-ratings
Simon Square	.74*	−.02
Flanker Square	.68*	−.01
Stroop Square	.60*	−.12
Antisaccade RT	−.58*	−.06
Switch cost RT	−.47*	−.07
Brief Self-Control Scale	−.14	.84*
Attention Control Scale	+.14	.62*
Urgency (UPPS)	+.04	.59*
Perseverance (UPPS)	−.16	.55*
Premeditation (UPPS)	+.05	.36*

Note. EFA = exploratory factor analysis; RT = reaction time; UPPS = Urgency, Premeditation, Perseverance, Sensation-Seeking.
$^{*}p < .01$.

Given that neither L2 AoA nor Years L2 provide even a hint that they are associated with greater EF or self-control, they are set aside in favor of a single latent variable for bilingualism based on the six measures shown for Factor 2 in Figure 5.

The SEM for a Model Assuming That Bilingualism Enhances EF and Self-Control

Figure 1 depicts a SEM that assumes that bilingualism adapts and enhances both EF and SC at the latent-variable level. It further assumes that the EF construct represents an ability that directly affects self-control in everyday life. The logic underlying this pathway starts with observing that the nature of many self-control items describes a type of event where control either succeeds (*I am good at resisting temptation*; *I refuse things that are bad for me*; *When concentrating I ignore feelings of hunger or thirst*) or fails (*I say inappropriate things*; *Pleasure and fun sometimes keep me from getting work done*; *I have trouble carrying on two conversations at once*). In contrast the performance-based tasks are artificial but designed deliberately to be sensitive to domain-general AC (the *square* tasks), inhibition (the antisaccade task), and switching (color-shape switching task). The performance-based tasks are novel for the participants, and performance should not be affected by automated skills and strategies that have been acquired to cope with specific tasks and situations where control is needed. Thus, domain-general capacity (as measured by the performance-based tasks) should facilitate self-control in everyday life, but it seems less likely that the reverse is uniformly reciprocated.

Figure 3

Simulated Random Data From a Parallel Analysis Show That the Eigenvalues for Simulated Data Exceed the Actual Data at Three Factors and That a Two-Factor Solution Is Best

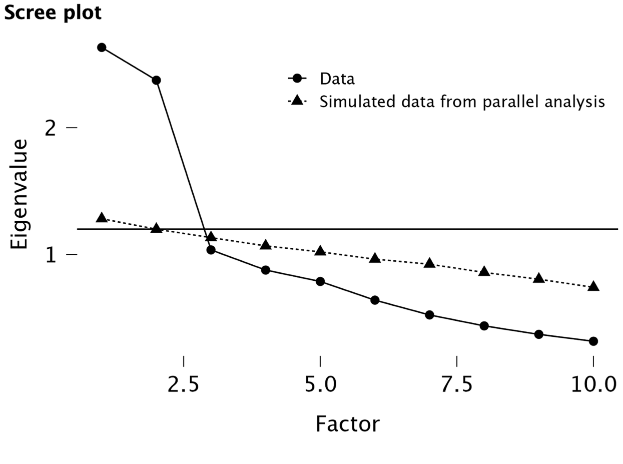


Figure 1 includes the factor loading and regression coefficients for this SEM. The model adequately fit the data, $\chi^2(101) = 214.87$, $p < .001$; comparative fit index (CFI) = .945, Tucker–Lewis index (TLI) = .935, root-mean-square error of approximation (RMSEA) = .059, 90% CI [.048, .069], standardized root-mean-square residual (SRMR) = .064.[2] The critical results are the regression coefficients between the latent variables. As shown in Figure 1, bilingualism does not predict performance-based EF $\beta = -.05$, $p = .56$). However, the latent variable for bilingualism did predict the latent variable for self-control ($\beta = +.17$, $p < .01$). This pattern suggests that bilingualism directly promotes self-control without being mediated through EF. Although surprising, this relationship could be very important and lead to the widespread adoption of self-reported measures of self-control under the assumption that they provide a pipeline to self-control ability that performance-based measures are insensitive to.

However, these self-reports using Likert scales are vulnerable to a host of biases such as acquiescence and social desirability and to subtle changes in wording (Paap, Anders-Jefferson, et al., 2024). Furthermore, Paap, Anders-Jefferson, et al. (2024) showed that this is true for the exact set of self-report measures used in the present study. To investigate the possible role of social desirability, this measure was added to the SEM as shown in the pathway model shown in Figure 2.

The new SEM shows that the Social Desirability strongly predicts ($\beta = +.45$) the self-control latent variable. Noteworthy, there is a concomitant reduction in the bilingualism to self-control regression-coefficient from a highly significant $\beta = +.170$, $p = .007$ to a nonsignificant $\beta = +.11$, $p = .053$.

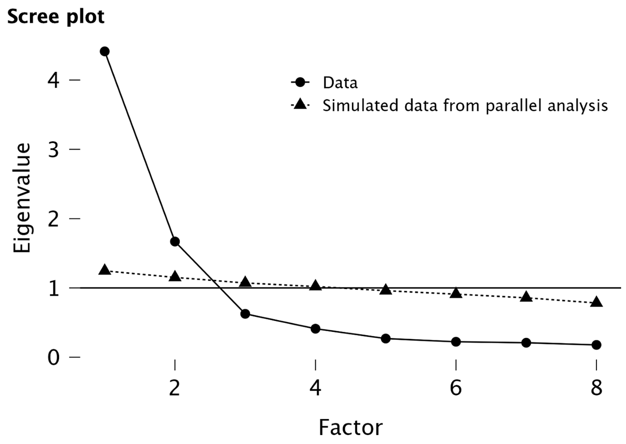
k-Means Clustering

To further probe this interpretation of the association between bilingualism and self-control, a k-means clustering analysis was performed. For this analysis, composite measures of bilingualism, self-control, and performance-based EF were formed by standardizing each measure and then computing the mean *z*-score for each composite. Thus, each participant contributes a data point in three-dimensional *z*-space. The algorithm is initialized by randomly selecting k locations to serve as "centroids," and each data point is assigned to the cluster whose centroid is closest. After assigning each data point to a cluster, the mean of each centroid is recalculated

[2] For interested readers, the separate measurement models within the SEM both showed an adequate fit for the data. Specifically, for the EF measures, the two-factor model provided a more adequate fit, $\chi^2(14) = 100.58$, $p = .001$; CFI = .884, TLI = .846, RMSEA = .077, 90% CI [.060, .095], SRMR = .070. Likewise, the bilingualism measurement model also adequately fit the data, $\chi^2(9) = 62.84$, $p < .001$; CFI = .964, TLI = .940, RMSEA = .135, 90% CI [.105, .167], SRMR = .029.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Figure 4
Scree Plot for Bilingualism

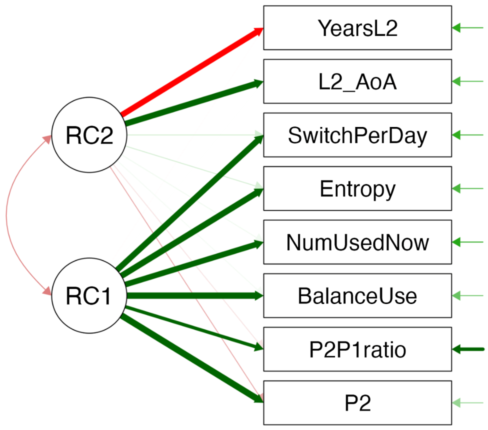


Note. Simulated random data from a parallel analysis show that the eigenvalues for simulated data exceed the actual data for the bilingualism measures at three factors and that a two-factor solution is best.

and becomes the new centroid for their respective clusters. The process is updated iteratively until the centroids no longer change significantly or a predetermined maximum number is reached. The results for $k = 5$ are reported, because larger values yield at least one very small cluster and smaller k's tend to produce clusters that poorly distinguish bilinguals from monolinguals (i.e., degree of bilingualism).

Table 6 shows the final cluster centers (19 iterations were required for convergence) for $k = 5$ and for each dimension: bilingualism, self-control, and performance-based EF. The number of participants in each cluster is 60, 73, 55, 77, and 63, respectively. The largest cluster (Cluster 4, $n = 77$) is intriguing and informative. The mean bilingualism score for this cluster is almost (mean $z = +.86$) a standard deviation above the mean of 0 and also has a high mean self-control score (mean $z = +.66$). This prompts the expectation that this cluster of participants should also have strong performance-based EF scores. But these participants actually perform below the mean of 0 (mean $z = -.19$). This profile is often referred to as having *illusory superiority* (or suffering from the Lake Wobegone effect).

Figure 5
The Path Diagram for the EFA of Eight Measures of Bilingual Experience



Note. EFA = exploratory factor analysis; BSC = Brief Self-Control; ACS = Attention Control Scale; P1 = proficiency of most proficient language; P2 = proficiency of second-most proficient language; RC1 = Residual Covariance 1; RC2 = Residual Covariance 2. See the online article for the color version of this figure.

Low-performing monolinguals (Cluster 3, mean $z = -.61$) rate their self-control as average (mean $z = +.01$), indicating a similar tendency to overrate their self-control, but this cluster is the smallest in size. Thus, it appears that inflated self-ratings tend to be associated with higher bilingualism scores. The consequences of including the social desirability measure in the SEM and these k-means clustering results are consistent with the interpretation that bilinguals in the population we are sampling from are more likely to engage in impression management and putting their best foot forward. This may be an effective strategy for members of minority communities that are not afforded their fair share of opportunities. It is the case that the participants classified as bilinguals have higher social desirability scores, $t(1, 251) = 4.83$, $p < .001$. They are more likely to be targets of ethnic or racial discrimination given that more of the bilinguals are immigrants ($M = .37$) compared to the monolinguals ($M = .03$), $t(1, 250) = 6.17$, $p < .001$. Bilinguals are more likely to be the target of class discrimination given that have lower composite SES z-scores ($M = -.23$) compared to monolinguals ($M = +.23$), $t(1, 251) = 4.83$, $p < .001$.

Supporting Evidence From Pure Groups Analyses

To this point, the bilingual advantage in EF hypothesis has been adjudicated with a correlation/regression approach that resonates with the SEM modeling that included latent variables for bilingualism, self-rated self-control, and performance-based EF. However, many bilingual researchers have expressed the belief that some measures of bilingual experience are more instrumental in adapting and improving domain-general EF than others. There is not agreement as to what these necessary or sufficient conditions may be, but if some experiences are causal and others are not, then the “inactive” ingredients in our bilingualism latent-variable may be creating noise that is drowning out the signal. There is also the possibility that a causal relationship between bilingualism and self-control/EF is not linear. Given these concerns, we compared our bilinguals to our monolinguals with t tests and BFs for each of the 10 measures of control. The BFs will be interpreted within the framework originally developed by Jeffreys (1961) and adapted by Wetzels et al. (2011). These tests are reported in Table 7.

Performance-Based EF

The three Square tasks all yield substantial evidence for the null, while switch costs and Antisaccade RT provided only “anecdotal” evidence for the null. Thus, the performance-based measures are clear in providing no evidence for a bilingual advantage in performance-based measures even when those measures are psychometrically sound.

Self-Control Scales

Three of the five self-control scales yielded bilingual advantages: BSC, urgency, and perseverance, with $BF_{10} = .2.40$, 2.04, and .1.44, respectively. However, the magnitudes are interpreted as only

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Table 6

The Mean z-Score on Each of the Three Dimensions for Each of the Five Clusters

	Cluster				
Measure	1	2	3	4	5
Bilingualism composite	−.62	.75	−.66	.89	−.73
Self-control composite	−.79	−.42	.01	.66	.53
EF performance composite	.15	.13	−.61	−.19	.25

Note. EF = executive function.

anecdotal evidence for the bilingual-advantage hypothesis. In contrast, the BFs for perseverance and the ACS show substantial evidence for the null hypothesis.

Discussion

No Evidence for a Bilingual Advantage

Our results show no evidence for a bilingual advantage in performance-based measures of EF. The null results cannot be attributed to EF measures with inadequate reliability and convergent validity. Although our first SEM (Figure 1) showed a significant pathway from the bilingualism latent variable to the self-control latent variable, the second SEM (Figure 2) showed that the pathway was no longer significant when a social-desirability measure was included. These results and a series of BFs analyses on an array of bilingual experiences and EF measures provided no evidence that bilingualism adapts and improves either self-rated SC or performance-based EF. Thus, the present results align well with the meta-analyses reviewed in the introduction that converged on the conclusion that the mean advantage is not distinguishable from 0, when corrected for publication bias. The contribution goes beyond mere replication of null results, because the advantage hypothesis was tested at the latent-variable level and extended to self-rating measures of self-control in everyday life that have been touted by some researchers as having greater face validity and predictive validity.

Table 7

Bayes Factor Analyses Testing for a Bilingual Advantage in Each of the 10 Control Measures

Performance-based measure	BF_{01}	Wetzel's evidence for null guideline
Antisaccade	2.08	Anecdotal
Switching cost	2.40	Anecdotal
Flanker Square	4.25	Substantial
Stroop Square	5.31	Substantial
Simon Square	6.07	Substantial
Self-rating measures	BF_{01}	Wetzel's evidence for null guideline
Brief Self-Control	0.42	Anecdotal
UPPS: Premeditation	7.10	Substantial
UPPS: Urgency	0.49	Anecdotal
UPPS: Perseverance	0.69	Anecdotal
ACS	4.14	Substantial

Note. BF = Bayes factor; UPPS = Urgency, Premeditation, Perseverance, Sensation-Seeking; ACS = Attention Control Scale.

Self-Rating Scales Versus Performance-Based Measures

The results of the present study confirmed earlier reports that self-ratings and performance-based measures weakly correlate at best. There are multiple reasons why they are clearly separable constructs.

Artificiality of Performance-Based EF

In Paap, Anders-Jefferson, et al. (2020), we suggested that the laboratory tasks are very sensitive to the participant's calibration of speed and accuracy, a skill that has little relevance to delaying gratification (urgency), planning before acting (premeditation), or having the grit to persist in the face of adversity (perseverance). The computerized EF tasks almost always encourage the participant to go as fast as possible without making more than an occasional error. But the mechanisms needed to filter out competing information in the nick of time and when there is little intrinsic value associated with a "correct" response, may be different from those needed to resist actions that affect laden and/or creatures of habit and have genuine and sometimes substantial costs and benefits. Moreover, competing information in the real world is not exquisitely tied to the onset of new task relevant information, and the conflict need not be resolved within the first couple of hundred milliseconds of the onset of the event. In fact, in the real world any rapid suppression of responses counter to long-term goals often needs to be sustained in order to be ultimately successful. In other words, the performance-based tasks are artificial and not tuned to goal-directed behavior in everyday life. Others have made similar points. For example, Barkley and Fischer (2011) "… suggest that low correlations between self-report and behavioral measures arise because they assess different constructs, both of which may be relevant for everyday behavior … rather than because one or both are invalid or dwarfed by measurement error" (p. 16). In summary, it was important to test for effects of bilingualism on self-report measures of self-control because performance-based measures of EF rely on "artificial" tasks that differ in several ways from how attention is controlled in everyday life.

Predictive Validity of Self-Rating Scales

Self-ratings appear to enjoy greater face validity. But does face-validity translate to predictive validity? Paap (2023) suggested that "When two purported measures of the same construct fail to correlate with each other one way of figuring out which is the valid measure is to see how well they each predict anticipated outcomes" p. 34. Since the impactful introduction of the BSC (Tangney et al., 2004), researchers have been enthused with its apparent ability to predict a variety of outcomes in the real world. For example, they reported that in their two college-student samples, BSC scores predicted grade point average ($r = +0.39$ and $+0.15$). Duckworth et al. (2010) provide evidence that self-control (BSC scores) may causally influence academic achievement. This longitudinal study tracked 142 fifth graders (M = 10.5 years) for 4 years. A growth curve analysis showed that changes in self-control over time predicted subsequent changes in grade point average. Gordeeva et al. (2017) reported a statistically significant correlation ($r = +0.17$)

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

between BSC scores for first-year university science majors and average scores in the immediately following examination session.

Ferrari et al. (2009) studied 606 adults (407 men, 199 women, $M_{\text{age}} = 38.5$ years) who were in recovery and residing in self-governed, communal living, abstinent homes across the United States. BSC scores were positively related to length of abstinence, but the factor defined by the four items with positive valence was primarily responsible for the significant relationship. These specific examples bolster the prospects that BSC scores will correlate with objective measures of other outcome variables.

A meta-analysis of the predictive validity of the BSC conducted by de Ridder et al. (2011) showed that observed behaviors (drawn from eight studies) and self-reported behaviors (drawn from 29 studies) were equally related to self-control as measured by the BSC. This is promising, but, as parenthetically shown, most of the evidence relies on subjective measures and correlational studies. In summary, although self-report measures are touted as having better predictive validity than performance-based measures, much of this evidence has not used objective measures of the outcome variable.

Foibles of Self-Ratings Using Likert Scales

Paap, Anders-Jefferson, et al. (2024) recently expanded on traditional concerns that self-ratings are vulnerable to a variety of biases (e.g., acquiescence and social desirability) and that adding negative-valanced items that require reverse-coding can significantly alter the factor structure and predictive validity of the scale. For example, when the nine negative items of the BSC are rewritten to create a completely positive scale: (a) The outcomes with strong correlations (with self-esteem, mental health, fluid intelligence) in the original scale weakened and the weak correlations (with satisfaction-with-life and happiness) strengthened, and (b) the mean overall scores increased. In summary, the self-rating scales (especially the BSC) that formed our self-control latent variable have a track record of being influenced by social desirability, and we know that the strength of the correlations with a variety of outcome variables depend on the proportion of items with positive valence.

Predictive Validity in the Real World

As a quick reminder our latent variable for performance-based EF was extracted from the three *Squared* tasks, switching costs, and antisaccade performance. Thus, one might presume that the Engle group would feel comfortable in labeling our performance-based latent variable "attention control." There is compelling evidence for treating AC as separable from other cognitive variables that are correlated with AC and with each other: gF, WMC, and processing speed (Burgoyne et al., 2023; Draheim et al., 2021). The argument starts with the finding that the strong relationship between WMC and fluid intelligence can be eliminated when accounting for the variance that each shares with AC (Draheim et al., 2021). Thus, AC is not only separable, but it also appears to be the active ingredient that causes WMC and fluid intelligence to correlate with one another.

There is overwhelming evidence that a wide array of performance-based measures of cognitive ability (related to memory, attention, and intelligence) positively correlate with good outcomes and negatively correlate with bad outcomes (Draheim et al., 2022; Mashburn et al., 2023). The domains examined include academic achievement, job performance, public health, mortality, and psychological well-being.

Unfortunately, for present purposes, the long history of investigating the role of cognitive ability on real-world outcomes does not, as a rule, tease apart the effects of AC from other cognitive abilities. Despite this challenge, Draheim et al. (2022) crafted a review intended to support the conclusion that AC is more strongly predictive of behavior in everyday life than either WMC or fluid intelligence. Although there appears to be a substantial cumulative weight favoring the centrality of AC, their comprehensive review of existing research (much of it published prior to 2020) is compromised by the very problems that the Engle group has brought to light and then solved in Draheim et al. (2021) and Burgoyne et al. (2023). In summary, now that we have psychometrically strong measures of AC, a new round of research is needed to assess how well these measures predict real-world work.

Predictive Validity in Synthetic Work

The superior *squared* tasks (and another good set initially identified by Draheim et al., 2021) were used by Burgoyne et al. (2023) to simulate the predictive validity of AC on real-world work by forming a latent variable for multitasking based on three paradigms and four measures. "These multitasking paradigms challenge participants to manage multiple information processing demands simultaneously (or concurrently), including elements of visual and auditory processing, arithmetic, symbol substitution, telling time, and problem solving" (Burgoyne et al., 2023, p. 72).

In a straightforward SEM model, the AC latent variable (based on the three Squared tasks) had a pathway coefficient of .87 to the latent variable for multitasking. This indicates that AC accounted for 75.6% of the variance in multitasking. To gain a better understanding of the degree to which AC ability can uniquely predict multitasking ability, another SEM added latent variables for fluid intelligence, WMC, and processing speed as additional predictors of multitasking. The pathway coefficients were .32 for AC, .57 for gF, .07 for WMC, and .20 for processing speed. Based on the R^2 values, it appears that AC still accounts for 10.2% of the variance in multitasking ability while WMC, as usual, eroded to 0.5%. Fluid intelligence (32.5%) and processing speed (4%) accounted for the remainder of the variance in multitasking accounted for by this model.

In summary, if this multitasking latent variable is a reasonable proxy for work in the real world, then it appears that both AC and gF are important abilities. It would be informative to expand Burgoyne et al.'s (2023) paradigm for predicting synthetic work (e.g., the multitasking latent variable) to include a latent variable based on self-rating scales. This would enable one to determine if the two constructs make distinctive contributions in a paradigm, where the "outcome" variable is objective, but not overly artificial.

Conclusions

We asked in our title if bilingual advantages in performance-based measures of EF (AC) would emerge when the measures show good reliability and validity and form a coherent latent variable. There was no evidence supporting this relationship when using psychometrically sound tasks. We also asked if self-rating measures of SC might provide a better measure of the same construct or provide an excellent measure of a related construct. The new evidence generated in the present study is consistent with the substantial body of published work showing that performance-based measures and self-rating measures do not converge on the same construct. Bilingualism does not enhance either.

References

Allom, V., Panetta, G., Mullan, B., & Hagger, M. S. (2016). Self-report and behavioral approaches to the measurement of self-control: Are we assessing the same construct? *Personality and Individual Differences*, *90*, 137–142. https://doi.org/10.1016/j.paid.2015.10.051

Arthur, W. A., Jr., Tubre, T. C., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven advanced progressive matrices test. *Journal of Psychoeducational Assessment*, *17*(4), 354–361. https://doi.org/10.1177/073428299901700405

Barkley, R. A. (2011). *Barkley deficits in executive functioning scale (BDEFS)*. Guilford Press.

Barkley, R. A., & Fischer, M. (2011). Predicting impairment in occupational functioning in hyperactive children as adults: Self-reported executive function (EF) deficits vs. EF tests. *Developmental Neuropsychology*, *36*(2), 137–161. https://doi.org/10.1080/87565641.2010.549877

Bertrams, A., & Dickhäuser, O. (2012). Passionate thinkers feel better: Self-control capacity as a mediator of the relationship between need for cognition and affective adjustment. *Journal of Individual Differences*, *33*(2), 69–75. https://doi.org/10.1027/1614-0001/a000081

Bialystok, E. (2019). The signal and the noise: Finding the pattern in human behavior. In I. A. Sekerina, L. Spradlin, & V. Valian (Eds.), *Bilingualism, executive function, and beyond* (pp. 17–34). John Benjamin. https://doi.org/10.1075/sibil.57.02bia

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. (2022). How does bilingualism modify cognitive function? Attention to the mechanism. *Psychonomic Bulletin & Review*, *29*(4), 1246–1269. https://doi.org/10.3758/s13423-022-02057-5

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

Bors, D. A., & Stokes, T. L. (1998). Raven's advanced progressive matrices: Norms for first-year university students and for the development of a short form. *Educational and Psychological Measurement*, *58*(3), 382–398. https://doi.org/10.1177/0013164498058003002

Brysbaert, M. (2024). Designing and evaluating tasks to measure individual differences in experimental psychology: A tutorial. *Cognitive Research: Principles and Implications*, *9*(1), Article 11. https://doi.org/10.1186/s41235-024-00540-2

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General*, *152*(8), 2369–2402. https://doi.org/10.1037/xge0001408

Cespón, J., & Carreiras, M. (2020). Is there electrophysiological evidence for a bilingual advantage in neural processes related to executive functions? *Neuroscience and Biobehavioral Reviews*, *118*, 315–330. https://doi.org/10.1016/j.neubiorev.2020.07.030

Chein, J. M., & Schneider, W. (2012). The brain's learning and control architecture. *Current Directions in Psychological Science*, *21*(2), 78–84. https://doi.org/10.1177/0963721411434977

Coderre, E. L., & van Heuven, W. J. B. (2014). The effect of script similarity on executive control in bilinguals. *Frontiers in Psychology*, *5*, Article 1070. https://doi.org/10.3389/fpsyg.2014.01070

Cyders, M. A., & Coskunpinar, A. (2011). Measurement of constructs using self-report and behavioral lab tasks: Is there overlap in nomothetic span and construct representation for impulsivity? *Clinical Psychology Review*, *31*(6), 965–982. https://doi.org/10.1016/j.cpr.2011.06.001

de Bruin, A., Dick, A. S., & Carreiras, M. (2021). Clear theories are needed to interpret differences: Perspective on the bilingual advantage debate. *Neurobiology of Language*, *2*(4), 433–451. https://doi.org/10.1162/nol_a_00038

de Ridder, D. T. D., de Boer, B. J., Lugtig, P., Bakker, A. B., & van Hooft, A. J. (2011). Not doing bad things is not equivalent to doing the right thing: Distinguishing between inhibitory and initiatory self-control. *Personality and Individual Differences*, *50*(7), 1006–1011. https://doi.org/10.1016/j.paid.2011.01.015

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology*, *111*(2), 225–236. https://doi.org/10.1037/0021-843X.111.2.225

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

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

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

Draheim, C., Pak, R., Draheim, A. A., & Engle, R. W. (2022). The role of attention control in complex real-world tasks. *Psychonomic Bulletin & Review*, *29*(4), 1143–1197. https://doi.org/10.3758/s13423-021-02052-2

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*, *150*(2), 242–275. https://doi.org/10.1037/xge0000783

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. https://doi.org/10.1016/j.jrp.2011.02.004

Duckworth, A. L., Tsukayama, E., & May, H. (2010). Establishing causality using longitudinal hierarchical linear modeling: An illustration predicting achievement from self-control. *Social Psychological & Personality Science*, *1*(4), 311–317. https://doi.org/10.1177/1948550609359707

Erceg, N., Galić, Z., & Bubić, A. (2019). "Dysrationalia" among university students: The role of cognitive abilities, different aspects of rational thought and self-control in explaining epistemically suspect beliefs. *Europe's Journal of Psychology*, *15*(1), 159–175. https://doi.org/10.5964/ejop.v15i1.1696

Ferrari, J. R., Stevens, E. B., Jason, L. A., & Jason, L. A. (2009). The relationship of self-control and abstinence maintenance: An exploratory analysis of self-regulation. *Journal of Groups in Addiction & Recovery*, *4*(1/2), 32–41. https://doi.org/10.1080/15560350802712371

Friedman, N. P., Hatoum, A. S., Gustavson, D. E., Corley, R. P., Hewitt, J. K., & Young, S. E. (2020). Executive functions and impulsivity are genetically distinct and independently predict psychopathology: Results from two adult twin studies. *Clinical Psychological Science*, *8*(3), 519–538. https://doi.org/10.1177/2167702619898814

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

García-Pentón, L., Fernández García, Y., Costello, B., Duñabeitia, J. A., & Carreiras, M. (2016). The neuroanatomy of bilingualism: How to turn a hazy view into the full picture. *Language, Cognition and Neuroscience*, *31*(3), 303–327. https://doi.org/10.1080/23273798.2015.1068944

Gobet, F., & Sala, G. (2023). Cognitive training: A field in search of a phenomenon. *Perspectives on Psychological Science*, *18*(1), 125–141. https://doi.org/10.1177/17456916221091830

Goldberg, D. P., & Williams, P. (1991). *A user's guide to the general health questionnaire*. NFER-Nelson.

Gordeeva, E. N., Osin, D. D., Suchkov, T. Y., Ivanova, O. A., & Bobrov, V. V. (2017). Self-control as a personal resource: Determining its relationships to

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

success, perseverance, and well-being. *Russian Education & Society*, *59*(5–6), 231–255. https://doi.org/10.1080/10609393.2017.1408367

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., & Timmer, K. (2017). Bilingualism and working memory capacity: A comprehensive meta-analysis. *Second Language Research*, *33*(3), 325–340. https://doi.org/10.1177/0267658316678286

Gunnerud, H. L., Ten Braak, D., Reikerås, E. K. L., Donolato, E., & Melby-Lervåg, M. (2020). Is bilingualism related to a cognitive advantage in children? A systematic review and meta-analysis. *Psychological Bulletin*, *146*(12), 1059–1083. https://doi.org/10.1037/bul0000301

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

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press, Clarendon Press.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, *130*(2), 169–183. https://doi.org/10.1037/0096-3445.130.2.169

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kwapis, K., & Bartczuk, R. P. (2020). The development and psychometric properties of the Polish version of the self-control scale. *Annales*, *33*(3), 123–144. https://doi.org/10.17951/j.2020.33.3.123-144

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

Lowe, C. J., Cho, I., Goldsmith, S. F., & Morton, J. B. (2021). The bilingual advantage in children's executive functioning is not related to language status: A meta-analytic review. *Psychological Science*, *32*(7), 1115–1146. https://doi.org/10.1177/0956797621993108

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*, 137–155. https://doi.org/10.1023/a:1006824100041

Mashburn, C. A., Burgoyne, A. P., & Engle, R. W. (2023). Working memory, intelligence, and life success. In R. H. Logie, Z. Wen, S. E. Gathercole, N. Cowan, & R. W. Engle (Eds.), *Memory in science for society* (pp. 149–184). Oxford University Press. https://doi.org/10.1093/oso/9780192849069.003.0007

Mason, L. A., Zimiga, B. M., Anders-Jefferson, R., & Paap, K. R. (2021). Autism traits predict self-reported executive functioning deficits in everyday life and an aversion to exercise. *Journal of Autism and Developmental Disorders*, *51*(8), 2725–2750. https://doi.org/10.1007/s10803-020-04741-8

Mazza, G. L., Smyth, H. L., Bissett, P. G., Canning, J. R., Eisenberg, I. W., Enkavi, A. Z., Gonzalez, O., Kim, S. J., Metcalf, S. A., Muniz, F., Pelham, W. E., III, Scherer, E. A., Valente, M. J., Xie, H., Poldrack, R. A., Marsch, L. A., & MacKinnon, D. P. (2021). Correlation database of 60 cross-disciplinary surveys and cognitive tasks assessing self-regulation. *Journal of Personality Assessment*, *103*(2), 238–245. https://doi.org/10.1080/00223891.2020.1732994

Mischel, W., Ayduk, O., Berman, M. G., Casey, B. J., Gotlib, I. H., Jonides, J., Kross, E., Teslovich, T., Wilson, N. L., Zayas, V., & Shoda, Y. (2011). "Willpower" over the life span: Decomposing self-regulation. *Social Cognitive and Affective Neuroscience*, *6*(2), 252–256. https://doi.org/10.1093/scan/nsq081

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

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

Mohades, S. G., Struys, E., Van Schuerbeek, P., Baeken, C., Van De Craen, P., & Luypaert, R. (2014). Age of second language acquisition affects nonverbal conflict processing in children: An fMRI study. *Brain and Behavior*, *4*(5), 626–642. https://doi.org/10.1002/brb3.246

Monnier, C., Boiche, J., Armandon, P., Baudoin, S., & Bellocchi, S. (2022). Is bilingualism associated with better working memory capacity? A meta-analysis. *International Journal of Bilingual Education and Bilingualism*, *25*(6), 2229–2255. https://doi.org/10.1080/13670050.2021.1908220

Necka, E., Gruszka, A., Orzechowski, J., Nowak, M., & Wojcik, N. (2018). The (In)significance for the Trait of Self-Control: A psychometric study. *Frontiers in Psychology*, *9*, Article 1139. https://doi.org/10.3389/fpsyg.2018.01139

Necka, E., Lech, B., Sobczyk, N., & Smieja, M. (2012). How much do we know about our own cognitive control: Self-report versus performance measures of executive control. *European Journal of Psychological Assessment*, *28*(3), 240–247. https://doi.org/10.1027/1015-5759/a000147

Paap, K. R. (2018). Bilingualism in cognitive science: The characteristics and consequences of bilingual language control. In A. De Houwer & L. Ortega (Eds.), *The Cambridge handbook of bilingualism* (pp. 435–465). Cambridge University Press. https://doi.org/10.1017/9781316831922.023

Paap, K. R. (2019). The bilingual advantage debate: Quantity and quality of the evidence. In J. W. Schwieter (Ed.), *The handbook of the neuroscience of multilingualism* (pp. 701–735). Wiley-Blackwell. https://doi.org/10.1002/9781119387725.ch34

Paap, K. R. (2023). *The bilingual advantage in executive functioning hypothesis: How the debate provides insight into psychology's replication crisis* (1st ed.). Routledge. https://doi.org/10.4324/9781003308027

Paap, K. R., Anders-Jefferson, R., Majoubi, J., & Balakrishnan, N. (2022). *What the brief self-control scale predicts depends on the population one samples from* [Paper presentation]. Annual Meeting of the Cognitive Neuroscience Society, San Francisco, California, United States.

Paap, K. R., Anders-Jefferson, R., Mikulinsky, R., Masuda, S., & Mason, L. (2019). On the encapsulation of bilingual language control. *Journal of Memory and Language*, *105*, 76–92. https://doi.org/10.1016/j.jml.2018.12.001

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L., & Mikulinsky, R. (2020). Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, Simon, and spatial Stroop tasks? *Cognitive Research: Principles and Implications*, *5*(1), Article 7. https://doi.org/10.1186/s41235-020-0207-y

Paap, K. R., Anders-Jefferson, R. T., Balakrishnan, N., & Majoubi, J. B. (2024). The many foibles of Likert scales challenge claims that self-report measures of self-control are better than performance-based measures. *Behavior Research Methods*, *56*(2), 908–933. https://doi.org/10.3758/s13428-023-02089-2

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002

Paap, K. R., Majoubi, J., Balakrishnan, N., & Anders-Jefferson, R. T. (2024). Bilingualism, like other types of brain training, does not produce far transfer: It all fits together. *The International Journal of Bilingualism*. Advance online publication. https://doi.org/10.1177/13670069231214599

Paap, K. R., Mason, L., Zimiga, B., Ayala-Silva, Y., & Frost, M. (2020). The alchemy of confirmation bias transmutes expectations into bilingual advantages: A tale of two new meta-analyses. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *73*(8), 1290–1299. https://doi.org/10.1177/1747021819900098

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Paap, K. R., Myuz, H. A., Anders, R. T., Bockelman, M. F., Mikulinsky, R., & Sawi, O. M. (2017). No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *Journal of Cognitive Psychology*, *29*(2), 89–112. https://doi.org/10.1080/20445911.2016.1248436

Paap, K. R., & Sawi, O. (2016). The role of test–retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods*, *274*, 81–93. https://doi.org/10.1016/j.jneumeth.2016.10.002

Paap, K. R., Sawi, O. M., Dalibar, C., Darrow, J., & Johnson, H. A. (2014). The brain mechanisms underlying the cognitive benefits of bilingualism may be extraordinarily difficult to discover. *AIMS Neuroscience*, *1*(3), 245–256. https://doi.org/10.3934/Neuroscience.2014.3.245

Partchev, I. (2020). Diagnosing a 12-item dataset of Raven matrices: With Dexter. *Journal of Intelligence*, *8*(2), Article 21. https://doi.org/10.3390/jintelligence8020021

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. https://doi.org/10.1016/j.tics.2005.12.004

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Raven, J. C., Court, J. H., & Raven, J. (1977). *Manual for Raven's advanced progressive matrices: Sets I and II*. H. K. Lewis.

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501–526. https://doi.org/10.1037/xlm0000450

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press. https://doi.org/10.1515/9781400876136

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

Salthouse, T. A. (2010). Is flanker-based inhibition related to age? Identifying specific influences of individual differences on neurocognitive variables. *Brain and Cognition*, *73*(1), 51–61. https://doi.org/10.1016/j.bandc.2010.02.003

Shilling, V. M., Chetwynd, A., & Rabbitt, P. M. A. (2002). Individual inconsistency across measures of inhibition: An investigation of the construct validity of inhibition in older adults. *Neuropsychologia*, *40*(6), 605–619. https://doi.org/10.1016/S0028-3932(01)00157-9

Stahl, C., Voss, A., Schmitz, F., Nuszbaum, M., Tüscher, O., Lieb, K., & Klauer, K. C. (2014). Behavioral components of impulsivity. *Journal of Experimental Psychology: General*, *143*(2), 850–886. https://doi.org/10.1037/a0033981

Stöber, J. (2001). The Social Desirability Scale-17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment*, *17*(3), 222–232. https://doi.org/10.1027//1015-5759.17.3.222

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324. https://doi.org/10.1111/j.0022-3506.2004.00263.x

Uysal, A., & Knee, C. R. (2012). Low trait self-control predicts self-handicapping. *Journal of Personality*, *80*(1), 59–79. https://doi.org/10.1111/j.1467-6494.2011.00715.x

Vang, Z. M., Sigouin, J., Flenon, A., & Gagnon, A. (2017). Are immigrants healthier than native-born Canadians? A systematic review of the healthy immigrant effect in Canada. *Ethnicity & Health*, *22*(3), 209–241. https://doi.org/10.1080/13557858.2016.1246518

von Bastian, C. C., de Simoni, C., Kane, M., Carruth, N., & Miyake, A. (2017, November). *Does being bilingual entail advantages in working memory? A meta-analysis* [Paper presentation]. Meeting of the Psychonomic Society, Vancouver, BC, Canada.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*(4), 669–689. https://doi.org/10.1016/S0191-8869(00)00064-7

Received June 16, 2024
Revision received November 16, 2024
Accepted November 25, 2024 ■