Research Article

Bilingualism Affords No General Cognitive Advantages: A Population Study of Executive Function in 11,000 People

Emily S. Nichols[1,2] (iD), Conor J. Wild[2,3], Bobby Stojanoski[2,3], Michael E. Battista[3], and Adrian M. Owen[2,3,4]

[1]Faculty of Education, The University of Western Ontario; [2]The Brain and Mind Institute, The University of Western Ontario; [3]Department of Psychology, The University of Western Ontario; and [4]Department of Physiology and Pharmacology, The University of Western Ontario

Abstract

Whether acquiring a second language affords any general advantages to executive function has been a matter of fierce scientific debate for decades. If being bilingual does have benefits over and above the broader social, employment, and lifestyle gains that are available to speakers of a second language, then it should manifest as a cognitive advantage in the general population of bilinguals. We assessed 11,041 participants on a broad battery of 12 executive tasks whose functional and neural properties have been well described. Bilinguals showed an advantage over monolinguals on only one test (whereas monolinguals performed better on four tests), and these effects all disappeared when the groups were matched to remove potentially confounding factors. In any case, the size of the positive bilingual effect in the unmatched groups was so small that it would likely have a negligible impact on the cognitive performance of any individual.

An enduring debate in the bilingualism literature is whether learning two or more languages affords benefits to cognition over and above the advantages of simply being able to speak a second language (Bialystok, 2017; Lehtonen et al., 2018). Much of this literature has focused on executive function, which refers to those cognitive processes, largely thought to depend on the frontal lobes of the brain, that are responsible for planning, managing, and executing goals (Owen, Downes, Sahakian, Polkey, & Robbins, 1990). The proposed mechanism underlying the bilingual advantage is that language joint activation, monitoring, and selecting rely on domain-general processes that in turn are strengthened through their use in bilingual language control (Bialystok, 2017). Over the years, several studies have shown that bilinguals outperform monolinguals on executive tasks, including tests of inhibition (Hernández, Costa, Fuentes, Vivas, & Sebastián-Gallés, 2010), cognitive control (Bialystok, Craik, & Luk, 2008), attention (Brito, Murphy, Vaidya, & Barr, 2016), working memory (Grundy & Timmer, 2017), and spatial processing (Morales, Calvo, & Bialystok, 2013). In contrast, other studies have reported no executive-function advantages in bilinguals relative to monolinguals (see Lehtonen et al., 2018, for a meta-analytic review).

The most comprehensive meta-analysis to date on the cognitive advantages of bilingualism examined inhibition, shifting, working memory, monitoring, attention, and verbal fluency but found no evidence for a bilingual advantage (Lehtonen et al., 2018). The results of this meta-analysis stand in contrast to those of a recent review, which concludes that bilinguals outperform

Corresponding Author:
Emily S. Nichols, The University of Western Ontario, Faculty of Education, London, Ontario, Canada N6A 5B7
E-mail: enicho4@uwo.ca

monolinguals on a wide variety of cognitive tasks, including inhibition, working memory, and attention, and that these advantages appear to extend into old age, protecting bilinguals from age-related diseases, such as Alzheimer's and other dementias (Bialystok, 2017). However, one problem with many previous studies is that they are based on relatively small sample sizes (Paap, Johnson, & Sawi, 2016). As a consequence, results are often disproportionately affected by other factors that are known to influence performance on tests of executive function (Noble, Norman, & Farah, 2005), such as socioeconomic status (SES; Morton & Harper, 2007), geographic background (Bialystok et al., 2008), and education (Perani et al., 2017). Additionally, a recent analysis of 104 conference abstracts on the topic of bilingualism and executive function revealed a systematic publication bias: Studies that supported the bilingual-advantage theory were more likely to be published subsequently as full journal articles than those that did not (de Bruin, Treccani, & Della Sala, 2015), casting doubt on the validity of any review or meta-analysis of the published literature.

Whether learning a second language is beneficial is not controversial; there are numerous advantages beyond those potentially afforded to cognition. For example, being able to communicate with a larger audience can lead to greater employment opportunities, more friendships, more potential to socialize, and easier travel in locations where those languages are spoken. All of these things are advantageous in and of themselves. However, whether these advantages extend to improvements in various aspects of executive functioning remains a contentious issue, requiring a larger and broader sample of monolinguals and bilinguals to resolve.

The Internet provides a unique opportunity for examining the relationship between bilingualism and executive function in the general population on a huge scale, allowing data to be sampled from participants from a broad range of socioeconomic, geographical, and educational backgrounds. If learning a second language affords advantages for executive function, then a population of bilinguals should outperform a population of monolinguals on a variety of tests of executive function.

To investigate whether speaking two or more languages is associated with improvement in executive function or protects against age-related cognitive decline, we invited participants to take part in an online study consisting of 12 tasks that compose the Cambridge Brain Sciences battery (www.cambridgebrainsciences.com). This executive battery assesses aspects of inhibition, executive function, selective attention, reasoning, verbal short-term memory, spatial working memory, planning, and cognitive flexibility. All participants also completed a detailed questionnaire, describing how many languages they speak, which languages they speak, which country they grew up in, their SES when growing up, the highest level of education they completed, and their age, gender, and handedness.

Method

The experimental protocol was approved by the University of Western Ontario Office of Human Research Ethics (Protocol No. 109196), and all participants gave written informed consent.

Materials

Sociodemographic questionnaire. To obtain information about the number of languages spoken, which languages were spoken, and demographic variables (such as age, country of origin, SES, and education), we asked participants to complete a detailed questionnaire. The questions used in the present study are available in Appendix S1 in the Supplemental Material available online.

Cognitive tests. Twelve cognitive tests were used to assess a broad range of executive functions, such as inhibition, working memory, problem-solving, and planning. An issue that is often raised in bilingualism research is whether the tests being used are sensitive to cognitive changes. These 12 tests have been validated in patients with anatomically specific frontal-lobe lesions (e.g., Bor, Duncan, Lee, Parr, & Owen, 2006; Owen et al., 1990), in neurodegenerative populations with frontostriatal cognitive impairments (Owen, Sahakian, Semple, Polkey, & Robbins, 1995), and in pharmacological intervention studies (e.g., Mehta et al., 2000). Functional-neuroimaging studies in healthy adults (e.g., Hampshire, Highfield, Parkin, & Owen, 2012) and in neuropathological populations (e.g., Williams-Gray, Hampshire, Robbins, Owen, & Barker, 2007) have shown these tests to be associated with activity in frontal or frontostriatal circuitry. The individual tests are described in detail below, and test-retest reliability measures are given in Table S1 in the Supplemental Material.

Double Trouble is a novel and challenging variant of the Stroop test (Stroop, 1935), a test of inhibition that has been widely used in the bilingualism literature (Bialystok et al., 2008; Blumenfeld & Marian, 2014; Rosselli, Ardila, Lalwani, & Vélez-Uribe, 2016). A target word (either "RED" or "BLUE") is displayed on the screen in either the color red or the color blue. The participant must select the probe word that correctly describes the color that the target word is drawn in. The problem's color mappings can be congruent (if every word correctly describes the color it is displayed in), incongruent (if either the target word or both probe words are displayed in the opposite color), or doubly incongruent (if the target and probes are both written

in the colors opposite to what they describe). Participants have 90 s to complete as many trials as possible. A correct response increases the total score by 1 point, and an incorrect response decreases the score by 1 point.

Spatial Planning is based on the Tower of London task (Shallice, 1982), which is widely used to measure executive function and has been used in the bilingualism literature (Festman, Rodriguez-Fornells, & Münte, 2010; Gunzenhauser, Karbach, & Saalbach, 2019). Numbered beads are positioned on a tree, and participants must rearrange the beads in ascending numerical order. They have 3 min to solve as many puzzles as possible, and the puzzles become progressively harder—requiring more moves and more complex planning. Trials are aborted if the participant makes more than twice the number of moves required to solve the problem. A successfully completed puzzle increases the final score by (2 × minimum number of moves required – the number of moves made).

Odd One Out is based on a subset of reasoning problems from the Cattell Culture Fair Intelligence Test (Cattell, 1949), which has also been used in bilingual-advantage research (Kempe, Kirk, & Brooks, 2015; Macnamara & Conway, 2014). Nine groups of colored shapes are displayed in a grid. The features (color, shape, number of items) define each group and are related to each other according to a set of rules. Participants must deduce the rules that relate these features and select the group with contents that do not correspond to those rules. They have 90 s to solve as many problems as possible, and the puzzles become progressively more difficult. A correct response increases the final score by 1 point, whereas an incorrect response decreases the score by 1 point.

Grammatical Reasoning is based on Baddeley's 3-min grammatical-reasoning test (Baddeley, 1968). On each trial, a written statement regarding two shapes is displayed on the screen, and the participant must indicate whether the statement correctly describes the shapes pictured below it. The participant has 90 s to complete as many trials as possible. A correct response increases the total score by 1 point, and an incorrect response decreases the score by 1 point.

Feature Match is based on classic feature-search tasks used to measure attentional processing (Treisman & Gelade, 1980). Attention has widely been used to study bilingual advantages (Bialystok, 2015; Brito et al., 2016). On each trial, two groups of items (each with n items) are displayed beside each other. The groups either are identical in their contents (and item positions) or differ by just one item. Participants have 90 s to complete as many trials as possible, indicating whether the groups match. A correct response increases the final score by n, and the subsequent trial has groups of $n + 1$ items.

If the response is incorrect, the total score decreases by n, and the next trial has groups of $n - 1$ items.

Polygons is based on the Interlocking Pentagons task, a test of visuomotor ability often used for assessing age-related disorders (Folstein, Folstein, & McHugh, 1975). It was included here to assess age-related cognitive decline in our two samples. On each trial, two overlapping wire-framed polygons are displayed on the left side of screen, and participants must indicate whether the shape to the right is identical to one of the two overlapping ones. A correct response increases the total score by the difficulty level, and the subsequent trial will be more difficult (i.e., differences between polygons will be subtler). An incorrect response decreases the total score by the difficulty level, and the next trial will be slightly easier. Participants have 90 s to complete as many trials as possible.

Digit Span is based on the verbal working memory component of the revised Wechsler Adult Intelligence Scale (WAIS-R; Wechsler, 1981) and has been used in bilingual-advantage research (Ratiu & Azuma, 2015; Rosselli et al., 2016). A sequence of digits is displayed one at a time in green in the center of the screen. Participants must then repeat the sequence of digits by selecting them on the on-screen keyboard. Difficulty is dynamically varied, as in previous tests, and the test ends after three mistakes. The resulting score is the length of the longest digit sequence successfully remembered.

Rotations is a task that measures the ability to spatially manipulate objects in mind (Silverman, Choi, Mackewn, Fisher, & Olshansky, 2000). On each trial, two groups of colored squares (each with n squares) are displayed beside each other. One of the groups is rotated by a multiple of 90°. The groups either are identical (when unrotated) or differ by the position of just one item, and participants must indicate whether the groups match. They have 90 s to complete as many trials as possible. A correct response increases the final score by n, and the subsequent trial has groups of $n + 1$ squares. If the response is incorrect, the total score decreases by n, and the next trial has groups of $n - 1$ squares.

Token Search is based on a test that is widely used to measure strategy during search behavior (Collins, Roberts, Dias, Everitt, & Robbins, 1998), and similar spatial tasks have been used previously in bilingualism research (Kerrigan, Thomas, Bright, & Filippi, 2017; Morales et al., 2013). A set of boxes, one of which contains a hidden green token, is displayed on a grid. Participants must find the token by clicking the boxes one at a time. Once found, the token is hidden within another box. The token will not appear within the same box twice, so the participant must search the boxes until the token has been found once within each box. An error is committed if the participant checks a box

that has already been clicked while trying to find the token or a box that previously contained the token. If the participant makes an error, a new trial begins with one box fewer to search. If the participant finds the token once in each box without making any errors, a new trial begins with one box more to search. The test ends after three errors. The resulting score is the maximum level completed.

Paired Associates is based on a test commonly used to assess memory impairments in aging clinical populations (Gould et al., 2005) and was included here to assess age-related cognitive decline in our two samples. Memory has also been shown to be impaired in patients with neurosurgical removals of frontal-lobe tissue (Owen, Sahakian, Semple, Polkey, & Robbins, 1995). Sets of boxes are displayed at random locations on a grid. The boxes open one after another to reveal an icon, after which they close. The icons are then displayed sequentially in the center of the screen, and the participant must select the box that contained that icon. If the participant remembers all the icon–location pairs correctly, then the next trial will have one box more. If an error is made, the next trial will have one box less. The test ends after three errors. The participant's score is the maximum number of pairs successfully remembered.

Spatial Span is based on the Corsi block-tapping task—a tool for measuring spatial short-term memory capacity. Researchers have widely studied spatial processing when examining the bilingual advantage (Kerrigan et al., 2017; Morales et al., 2013; Rosselli et al., 2016); the version of the test used here is associated with frontal-lobe activity in healthy participants and is sensitive to frontal-lobe removals in patients (Bor et al., 2006). Sixteen purple boxes are displayed in a grid. A sequence of randomly selected boxes turn green one at a time (900 ms per green square). Participants must then repeat the sequence by clicking boxes in the same order. Difficulty is varied dynamically: Correct responses increase the length of the next sequence by one square, and an incorrect response decreases the sequence length. The test ends after three errors. The score is the length of the longest sequence successfully remembered.

Monkey Ladder is based on a task from the nonhuman-primate literature (Inoue & Matsuzawa, 2007), and similar spatial tasks have been used previously in bilingualism research (Kerrigan et al., 2017; Morales et al., 2013). Numbered boxes are displayed simultaneously at random locations within a grid. After a variable interval (number of squares × 900 ms), the numbers disappear, leaving only the boxes. Participants must click the boxes in ascending numerical sequence. Difficulty is varied dynamically, as in Spatial Span. The test ends after three errors, and the resulting score is the length of the longest sequence successfully remembered.

Experimental design

Data were collected via the Cambridge Brain Sciences online platform (www.cambridgebrainsciences.com). The accuracy of online data has been found to be high (Wesnes et al., 2017), and this particular platform has been used in previous large-scale studies (Hampshire et al., 2012; Wild, Nichols, Battista, Stojanoski, & Owen, 2018). After reaching the website, participants were asked to give informed consent and to register with an e-mail address. They next completed a detailed questionnaire inquiring about demographic and lifestyle items (available in Appendix S1), which took approximately 10 min. They were then asked to complete 12 cognitive tests measuring a broad range of cognitive abilities, including inhibition, selective attention, reasoning, verbal short-term memory, spatial working memory, planning, and cognitive flexibility. This testing period took approximately 35 to 40 min.

Only data from the participants who completed all relevant questionnaire items and all 12 tests were included in the analysis. In accordance with local ethical guidelines, we did not include participants below the age of 18. In total, 11,213 participants met these requirements. Data were then cleaned to remove impossible and improbable questionnaire responses. Test scores were filtered for outliers in two passes: Scores greater than 6 standard deviations from the mean were assumed to be technical errors and were first removed, eliminating 32 participants. Then scores greater than 4 standard deviations from the recalculated mean, which were assumed to be performance outliers, were removed, eliminating 140 participants. Consequently, 11,041 participants were included in the final analysis.

It should be noted that because the data were collected from volunteers who self-selected and were not randomly assigned to groups, the present study is observational rather than experimental. Although we attempted to control for well-known confounding factors by including them in our regression analysis and by matching our samples, there are of course potential unknown confounds that have not been considered and that may explain our findings.

Statistical analysis

Data were analyzed using the R statistical toolbox (Version 3.6.1; R Core Team, 2019), and all figures were constructed using the R package *ggplot2* (Version 2.2.1; Wickham, 2016). Chi-square tests were used to assess proportions of SES, handedness, gender, and education between groups, and a two-sided *t* test was used to compare age between groups.

Table 1. Descriptive Statistics for the Two Groups in the Demographically Matched Sample

Variable	Monolinguals	Bilinguals	Comparison	p
n	372	372		
Gender			$\chi^2(2, N = 744) = 1.41$	.494
Female	68.82%	66.13%		
Male	30.38%	32.26%		
Other	0.80%	1.61%		
Age (years)	$M = 34.66$ ($SD = 11.26$)	$M = 34.75$ ($SD = 11.28$)	$t(742) = -0.10$	.920
Highest level of education completed			$\chi^2(4, N = 744) = 0.01$	.999
None	0.00%	0.00%		
High school	23.12%	23.12%		
Postsecondary	48.39%	48.12%		
Master's	20.43%	20.70%		
Doctoral/professional	8.06%	8.06%		
Socioeconomic status			$\chi^2(1, N = 744) = 2.71$	.100
At or above poverty line	96.24%	93.28%		
Below poverty line	3.76%	6.72%		
Handedness			$\chi^2(1, N = 744) = 0.08$	.448
Right	89.78%	91.67%		
Left	10.22%	8.33%		

Note: Welch's *t* test was used to compare age.

Given that the data in this study were observational in nature, group imbalances in demographic variables and other potential confounding factors might drive any observed group differences in cognitive performance. To control for such factors, we constructed two groups (monolingual and bilingual) matched in age, education, SES, gender, and handedness using the R package *MatchIt* (Version 3.0.2; Ho, Imai, King, & Stuart, 2011) with the nearest-neighbor-matching method. Prior to creating the matched samples, we also removed participants who may have masked any positive effects of bilingualism on task performance. Non-English speakers (who were more likely to be bilingual) may have been at a disadvantage, given that the tests and their instructions were provided in English, so only participants who selected English as one of their spoken languages were included in the matching processes. Similarly, participants who indicated that they were bilingual but selected only a single language were not included in data analysis on the assumption that this was an error or that they did not consider themselves fully bilingual. Finally, participants from some countries (Portugal and "other") were much more likely to be bilingual, and so the matched samples were constructed from individuals only in Canada, the United States, the United Kingdom, and Australia. Descriptive information for the matched samples, with 372 monolinguals and 372 bilinguals, is in Table 1. Descriptive information for the unmatched sample, with 5,994 monolinguals and 5,047 bilinguals, is in Table S2 in the Supplemental Material.

Factor analysis. Imaging studies have underscored the fact that there is rarely a one-to-one mapping between cognitive functions and the brain areas (or networks) that underpin them. One approach to this issue is to examine the complex statistical relationships between performance on any one cognitive task (or group of tasks) and changes in brain activity to reveal how one is related to the other. To do this most effectively, researchers must include large amounts of data because of the natural variance in cognitive performance (and brain activity) across tests and across individuals. In the age of computerized Internet testing and "big data," this problem becomes much easier to solve. Hampshire et al. (2012) collected data on the 12 Cambridge Brain Sciences tasks from approximately 45,000 participants. These data were then subjected to a factor analysis, and three discrete factors relating to overall cognitive performance were identified. Each one of these factors is something that no single test can assess; each represents an independent aspect of cognitive function that is best described by performance on a combination of tests. They were labeled, for convenience, as encapsulating aspects of short-term memory, reasoning, and verbal abilities, respectively. This technique allows an individual's performance to be compared with a very large normative database in terms of these descriptive factors rather than in terms of performance on a single test.

Here, the same 12 tests were used to create three factor scores reflecting performance in three cognitive domains (memory, reasoning, and verbal ability) identified by

Hampshire and colleagues (2012). The three cognitive-domain scores were calculated using the formula $Y = X(Ar+)^T$, where Y is the resulting $N \times 3$ matrix of domain scores, X is the $N \times 12$ matrix of test z scores, Ar is the 12×3 matrix of varimax-rotated principal component weights (i.e., factor loadings) from Hampshire et al., and T means "transpose." All 12 tests contributed to each domain score, as determined by their component weights. The resulting factor scores (i.e., principal component analysis scores) are standardized (i.e., population $M = 0$, $SD = 1.0$), so a score above zero indicates that someone is above average.

Matched sample.

Linear regression. To investigate the effect of bilingualism on performance on each test as well as on our three factors, we performed linear regression separately for each of the 15 scores. Models were constructed as follows: bilingualism (monolingual vs. multilingual), SES (below poverty line vs. at or above poverty line), and handedness (left vs. right) were constructed as binary regressors. Education, gender, country of origin, and languages spoken at home were treated as categorical, with $n - 1$ regressors. Participants' age (mean-centered across the entire sample) was also included, as was an Age × Group interaction term. This was done to verify that a second well-studied effect could be replicated in this sample and to further test the hypothesis that bilingualism might provide a cognitive protective effect against aging. The regression models were built and estimated using the R packages *stats* (R Core Team, 2019) and *lmSupport* (Version 2.9.13; Curtin, 2018). Bayes factor estimates that compared a model including the bilingualism regressor (i.e., the full model) with a model that did not (i.e., the reduced model) were computed using an approximation based on the Bayesian information criterion (BIC) from these two models, as specified by Wagenmakers (2007). This calculation was similarly performed for the age regressor. All statistical tests were corrected for multiple comparisons using a false-discovery rate (FDR) across scores (12 tests and three factors), and separately for each effect (group, age, and Age × Group). Because large sample sizes will inherently produce significant results in some statistical tests, we included measures of standardized and unstandardized effect sizes, confidence intervals, and Bayes factors to put effects into meaningful context. Including the intercept term, the final design matrix contained 22 columns and 744 rows.

Model selection. Following the initial set of linear regressions, we performed model selection to assess whether any effects (or lack thereof) were due to which regressors we chose to include in the model. For this, we used the R package *MuMIn* (Version 1.43.6; Barton, 2019). Model selection was performed on each of the 12 tests and our three factors as follows. First, the global model was specified, with all predictors including the Age × Group interaction term. Next, models were estimated for every possible nested version of the global model but always with the interaction term, yielding 64 models with unique combinations of regressors. From this, the model with the best fit was selected on the basis of the lowest BIC. We then extracted all parameter estimates and p values for age, group, and the interaction term from each of the models, and we calculated the percentage of models that led to a significant result ($p < .05$, corrected for multiple comparisons using the FDR). To avoid overcorrection, we calculated the FDR separately for each model variation—that is, for a single iteration of regressors, 15 p values (12 tests and three factors) were extracted and then FDR corrected. This procedure was performed on each of the 64 models. Finally, we determined which regressors were likely to be included in a significant model. Using this methodology, we were able to assess how much the variables included in the model were likely to influence the outcome.

Unmatched sample. We also performed follow-up analyses using the entire unmatched sample (5,994 monolinguals and 5,047 bilinguals) to investigate whether any effects of bilingualism would be observed using a significantly larger, though arguably less controlled, data set. Linear regression models for the 15 scores were constructed just as in the matched sample analysis. With the intercept term, the final design matrix for the global model contained 22 columns and 11,041 rows. The model was then selected in the same manner as specified above in order to determine the set of predictors that led to the highest model fit. FDR correction was again performed separately for each iteration of regressors.

Results

Matched sample

Of the 40,105 participants who registered for the study, 11,213 (age range = 18–87 years) completed all 12 cognitive tasks and all of the questions pertaining to bilingualism, country of birth, SES, and education; 744 participants were included in the final sample after data cleaning and matching were completed (see the Method section). Descriptive information for this subsample is available in Table 1, and distributions including medians, quartiles, and ranges are shown in Figures 1 and 2.

For each of the 15 scores of cognitive performance, the model including only age, group, and the interaction term provided the best fit; none of them showed a significant group effect (Table 2) or a significant Age × Group interaction (Table 3, Figs. 3 and 4). Bayes
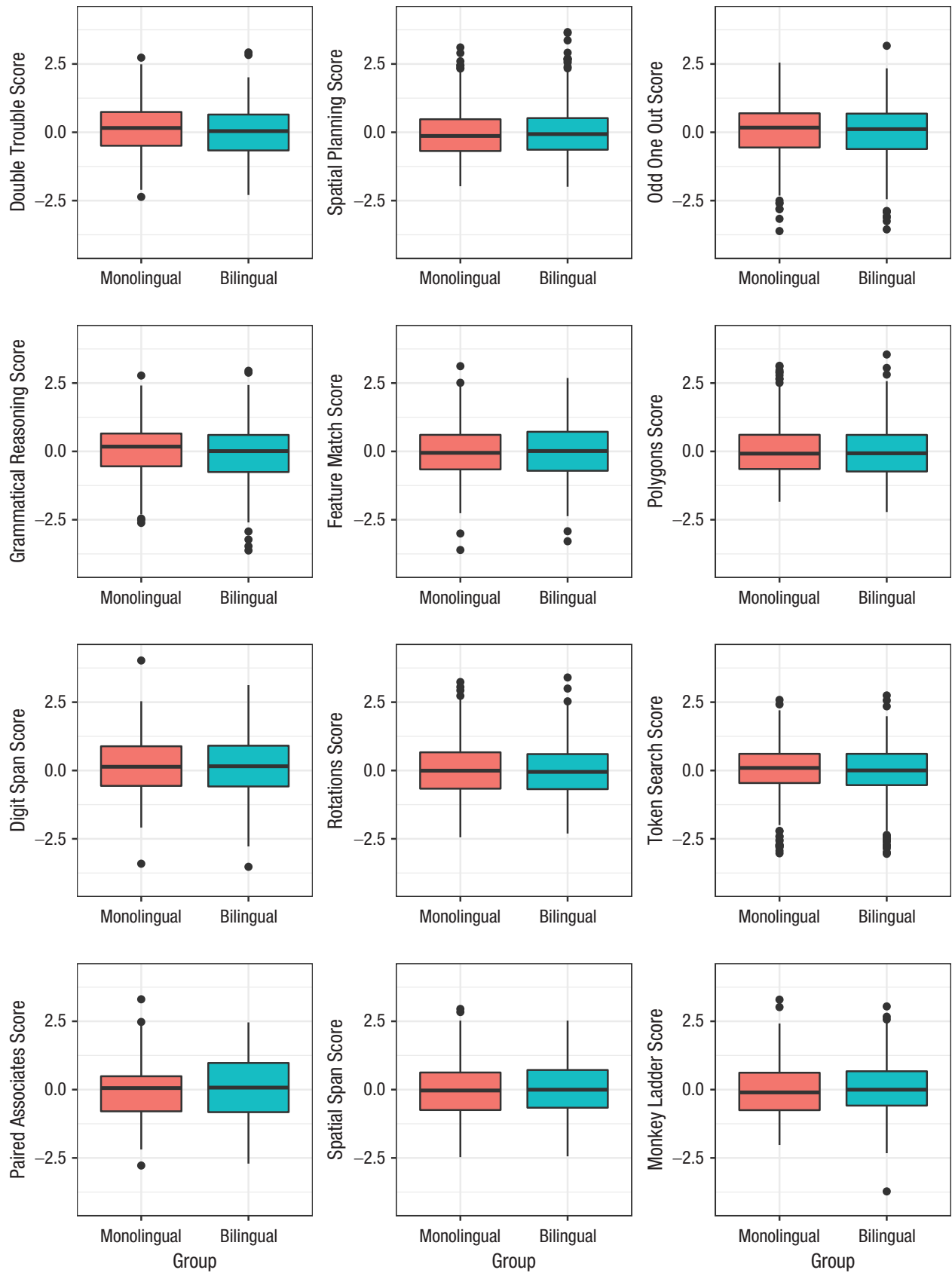
Fig. 1. Distribution of scores for both of the demographically matched groups on each of the 12 tests. Medians are indicated by thick black horizontal lines. The first and third quartiles are marked by the lower and upper edges of the boxes, respectively. Lower and upper whiskers extend to the smallest and largest value, respectively, within 1.5 times the interquartile range. Outlying values beyond these ranges are plotted individually.

Fig. 2. Distribution of scores for both of the demographically matched groups for the three factors. Medians are indicated by thick black horizontal lines. The first and third quartiles are marked by the lower and upper edges of the boxes, respectively. Lower and upper whiskers extend to the smallest and largest value, respectively, within 1.5 times the interquartile range. Outlying values beyond these ranges are plotted individually.
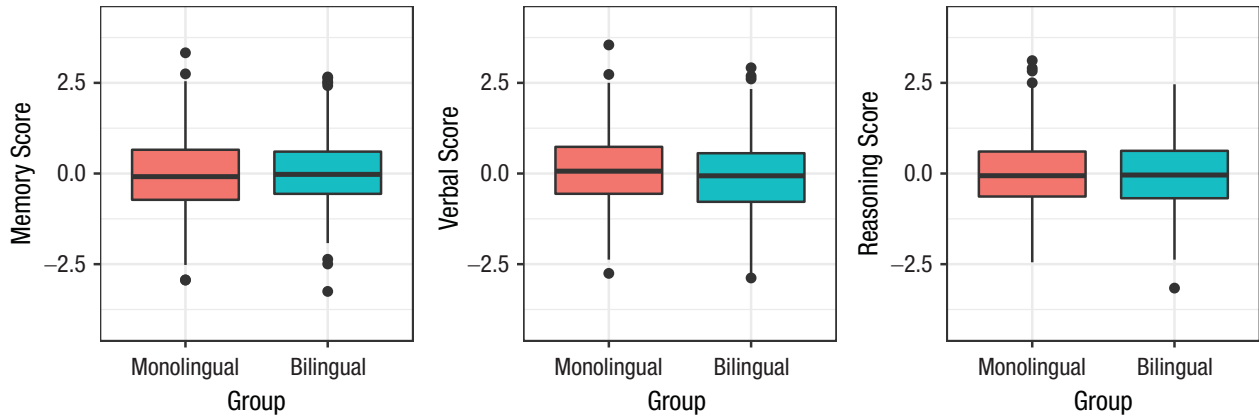
factors strongly supported the null hypotheses that there was no effect of group, or Age × Group interaction, for all 15 scores. A single exception to this was the Rotations test, in which the Bayes factor provided anecdotal evidence in support of the Age × Group interaction (though it was still not significant). All tests and factors showed a statistically significant effect of age, except for Odd One Out (see Table S3 in the Supplemental Material).

When examining the distribution of significant p values resulting from the 64 model variations, we found that no test showed a significant group effect in any model. The age term was significant in 100% of models for all tests and factors except Odd One Out. Finally, no test or factor showed a significant Age × Group interaction in any model. Distributions of group-level p values for each test and factor are shown in Figures 5 and 6. Distributions of group-level bilingualism

Table 2. Bilingualism Regression Parameters for the Best-Fitting Model Following Model Selection in the Matched Sample

Variable	ΔR^2	Group β	$t(740)$	p	99% CI β	BF_{01}	η_p^2
Tests							
Double Trouble	< .01	−0.11	−1.61	.402	[−0.30, 0.07]	157.05	< .01
Spatial Planning	< .01	0.09	1.25	.528	[−0.11, 0.26]	210.02	< .01
Odd One Out	< .01	−0.05	−0.64	.823	[−0.24, 0.14]	548.82	< .01
Grammatical Reasoning	< .01	−0.11	−1.49	.402	[−0.29, 0.08]	198.41	< .01
Feature Match	< .01	0.03	0.40	.852	[−0.15, 0.21]	379.17	< .01
Polygons	< .01	−0.02	−0.34	.852	[−0.21, 0.16]	280.09	< .01
Digit Span	< .01	< 0.01	0.06	.976	[−0.18, 0.19]	232.86	< .01
Rotations	.01	−0.07	−1.04	.528	[−0.26, 0.11]	13.46	< .01
Token Search	< .01	−0.11	−1.51	.402	[−0.29, 0.08]	61.38	< .01
Paired Associates	< .01	< 0.01	0.02	.976	[−0.18, 0.19]	37.56	< .01
Spatial Span	< .01	0.04	0.60	.823	[−0.14, 0.22]	45.13	< .01
Monkey Ladder	< .01	0.12	1.77	.402	[−0.06, 0.31]	23.65	< .01
Factors							
Memory	< .01	0.08	1.20	.528	[−0.10, 0.26]	296.42	< .01
Verbal	< .01	−0.11	−1.54	.402	[−0.30, 0.08]	38.41	< .01
Reasoning	< .01	−0.05	−0.75	.705	[−0.24, 0.13]	195.54	< .01

Note: BF_{01} is the Bayes factor showing the likelihood of the null over the alternative hypothesis. CI = confidence interval.

Table 3. Age × Group Interaction Regression Values for the 12 Cognitive Tasks and Three Factors in the Demographically Matched Sample

Variable	ΔR^2	Interaction β	$t(740)$	p	99% CI β	BF_{01}	η_p^2
Tests							
Double Trouble	< .01	< 0.01	0.70	.609	[−0.01, 0.02]	21.28	< .01
Spatial Planning	< .01	< 0.01	0.97	.551	[−0.01, 0.02]	16.93	< .01
Odd One Out	< .01	< −0.01	−0.44	.658	[−0.02, 0.01]	24.72	< .01
Grammatical Reasoning	< .01	< 0.01	0.65	.609	[−0.01, 0.02]	22.12	< .01
Feature Match	< .01	< 0.01	1.09	.520	[−0.01, 0.02]	15.06	< .01
Polygons	< .01	< 0.01	1.35	.441	[−0.01, 0.03]	10.88	< .01
Digit Span	< .01	0.01	1.52	.441	[−0.01, 0.03]	8.55	< .01
Rotations	< .01	0.02	2.64	.129	[< −0.01, 0.03]	0.84	< .01
Token Search	< .01	0.01	1.64	.441	[−0.01, 0.03]	7.03	< .01
Paired Associates	< .01	< 0.01	0.57	.609	[−0.01, 0.02]	23.15	< .01
Spatial Span	< .01	< −0.01	−0.82	.609	[−0.02, 0.01]	19.43	< .01
Monkey Ladder	< .01	< −0.01	−1.13	.520	[−0.02, 0.01]	14.43	< .01
Factors							
Memory	< .01	< −0.01	−0.63	.609	[−0.02, 0.01]	22.41	< .01
Verbal	< .01	0.01	1.88	.441	[< −0.01, 0.03]	4.62	< .01
Reasoning	< .01	0.01	1.45	.441	[−0.01, 0.03]	9.50	< .01

Note: BF_{01} is the Bayes factor showing the likelihood of the null over the alternative hypothesis. CI = confidence interval.

parameter estimates for each test and factor are shown in Figures S3 and S4 in the Supplemental Material.

Unmatched sample

In the unmatched sample, 5,047 participants reported speaking two or more languages, whereas 5,994 participants reported speaking only one language, as outlined in Table S2. On average, the two groups were well matched in terms of gender, $\chi^2(2, N = 11,041) = 3.65$, $p = .162$, and handedness, $\chi^2(1, N = 11,041) = 0.92$, $p = .338$. Bilinguals were younger than monolinguals, $t(10832) = 15.38$, $p < .001$, and a larger proportion of them were from high-SES backgrounds, $\chi^2(1, N = 11,041) = 15.10$, $p < .001$. The groups differed in their proportions of levels of education, $\chi^2(4, N = 11,041) = 380.00$, $p < .001$, but the overall pattern did not favor one group or the other (see Table S2). On average, bilinguals reported speaking 2.57 languages (range = 2–9).

Scores on each of the 12 tests and three factors were again submitted to linear regression, and the global model included all regressors and the Age × Group interaction. Distributions including medians, first and third quartiles, and ranges for each test are shown in Figures S1 and S2 in the Supplemental Material.

As shown in Table 4, the set of regressors that provided the best fit differed depending on the test or factor. Regression coefficients of the best-fitting model

for each test and factor are summarized in Table 5 (describing the group term) and Table 6 (describing the interaction term). In five tests and two factors, the selected model showed a significant group effect, but only Digit Span showed a bilingual advantage, $\Delta R^2 < .01$, $\beta = 0.05$, $t(11031) = 2.52$, $p = .029$; Grammatical Reasoning, Feature Match, Rotations, and Token Search, and both the Verbal and Reasoning factors, showed a monolingual advantage. Similar to the findings in the matched sample, all tests and factors showed a significant effect of age except for Odd One Out (see Table S4 in the Supplemental Material). No tests or factors showed a significant Age × Group interaction (see Figs. 7 and 8).

When examining the distribution of significant p values resulting from the 64 model variations, we found that eight tests and two factors showed a significant group effect some proportion of the time, depending on the set of regressors (exact percentages are shown in Table 4). Bilinguals showed an advantage in Double Trouble and Digit Span, whereas monolinguals showed an advantage in Feature Match, Rotations, Token Search, and their overall Reasoning factor score. The direction of the advantage varied for Grammatical Reasoning (25% monolingual advantage and 37.5% bilingual advantage) and the Verbal factor score (12.5% monolingual advantage and 75% bilingual advantage), depending on the set of regressors
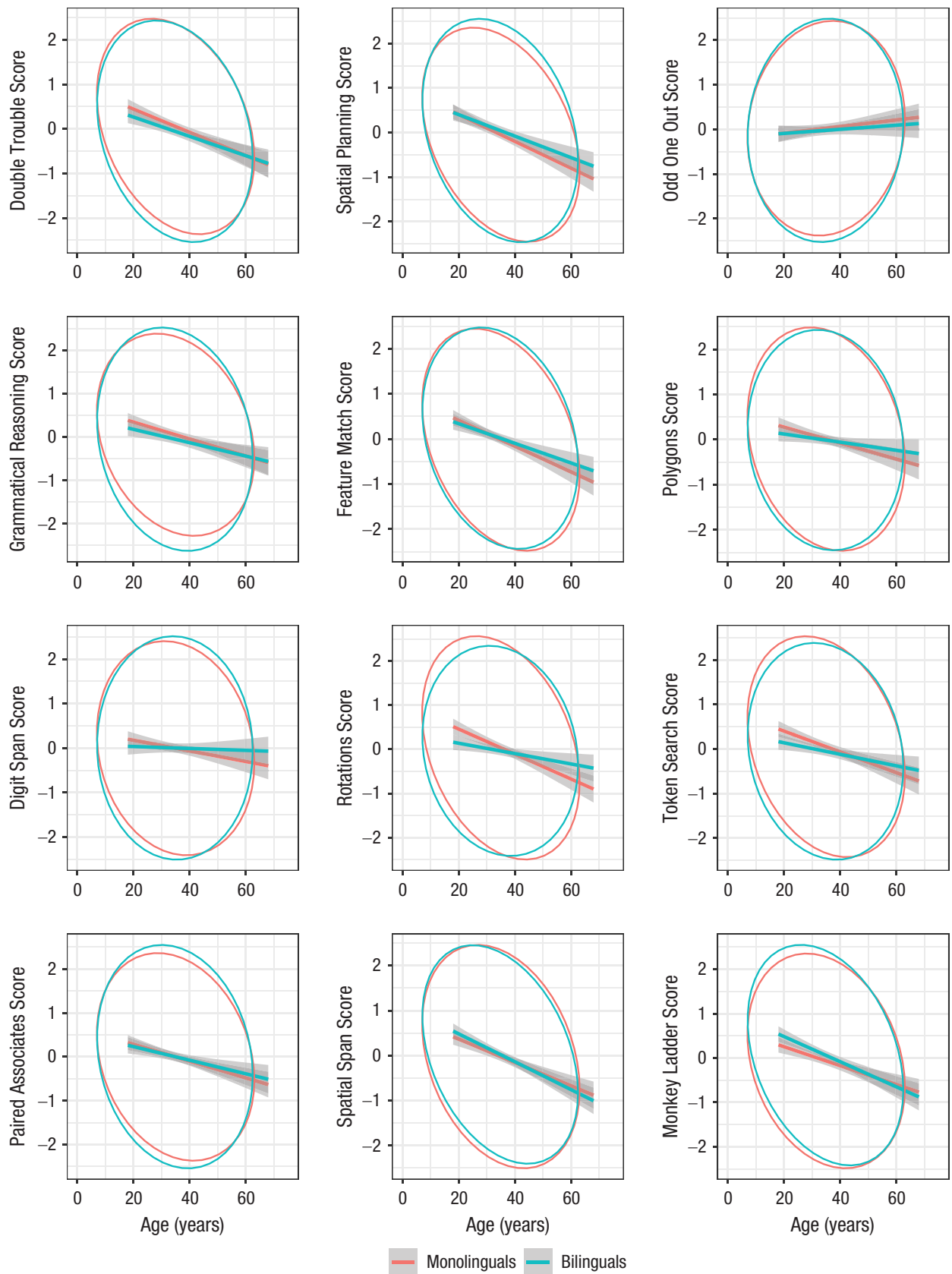
Fig. 3. Plots showing the linear relationship between age and scores for each of the tests in the matched sample. For each regression line, a 95% confidence ellipse and a 95% confidence interval is shown. Effect sizes are reported in Table S3 in the Supplemental Material. Individual data points have not been included because of the large sample size.
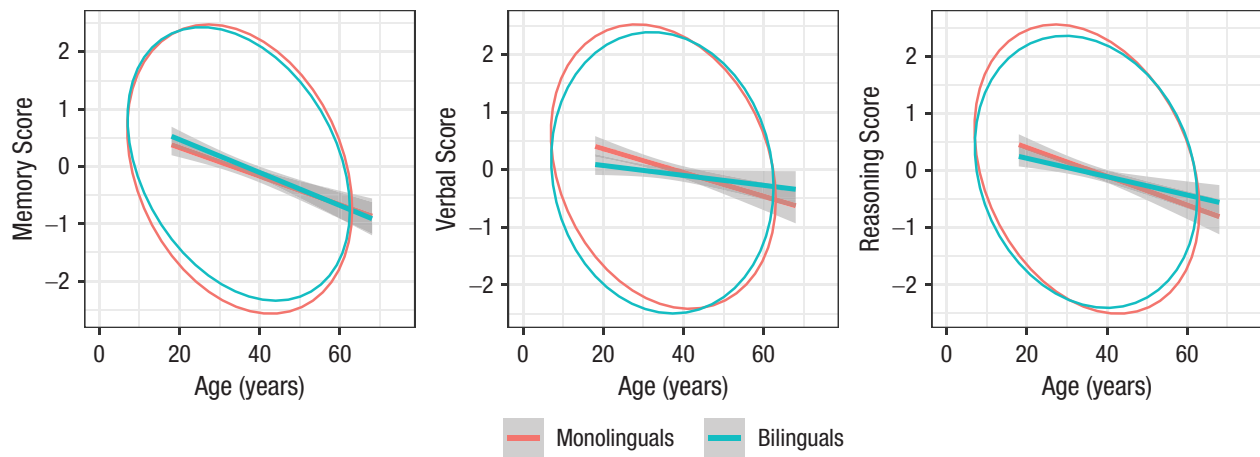
Fig. 4. Plots showing the linear relationship between age and scores for each of the three factors in the matched sample. For each regression line, a 95% confidence ellipse and 95% confidence interval is shown. Individual data points have not been included because of the large sample size.

included. The age term was again significant in 100% of models for all tests and factors except Odd One Out. Only Grammatical Reasoning showed a significant interaction in 12.5% of models (Table 5), and in all cases, monolinguals had a steeper decline with age. Distributions of group-level p values for each test and factor are shown in Figures 9 and 10. Distributions of group-level bilingualism parameter estimates for each test and factor are shown in Figures S5 and S6 in the Supplemental Material.

Although some tests showed significant effects, caution should be used when interpreting tests of significance because of the large sample size, and the focus should be placed on measures such as effect sizes and confidence intervals. The effect sizes indicate that being bilingual explains less than 1% of the variance in all significant results; for example, bilinguals outperformed monolinguals by a standard deviation of 0.05 on Digit Span. Because of the difficulty in interpreting null results, we examined the data further by estimating the BICs for both the full and reduced models, which were subsequently used to calculate the Bayes factor for the full model (Wagenmakers, 2007). We found support for a bilingual advantage only on the tests in which monolinguals showed an advantage, with the BIC for Digit Span (BIC = 168.69) strongly supporting the null hypothesis. A Bayesian analysis of the other eight tasks and factors strongly or decisively supported the null hypothesis, and the data suggest that the pattern of results was more likely to occur if there were no differences between bilinguals and monolinguals (BF_{01}s and effect sizes are reported for all 12 tasks and three factors in Table 5).

Discussion

In this study of 11,041 participants, no reliable differences in executive function were observed between monolinguals and people who reported speaking more than one language. First, when we created matched groups to eliminate confounds that may be masking an executive function advantage in bilinguals, and to ensure that our groups met the criteria for being either monolingual or bilingual, we found no significant group differences. Second, when utilizing the entire (large, though unbalanced) data set, we found that only one task, Digit Span, showed an advantage in performance in bilinguals. Although this result is statistically significant, it is important to put it in perspective: The regression coefficient was 0.05. In real terms, this means that, statistically, speaking a second language is associated with better memory for digits, but that difference is one twentieth of 1 standard deviation. To further put this into context, we note that the standardized effect size (i.e., η_p^2) was less than .01, which is well below what is considered small—confirming that this effect was trivial, even if it was statistically significant. Further, though p was below .05, the Bayes factor showed strong support for the null hypothesis, calling the statistical significance into question. In 11 other cognitive tasks and our three cognitive factors, including several that have previously suggested a bilingual advantage, there were either no differences between groups or a positive difference for monolinguals (although these differences had negligible effect sizes).

Another issue that we examined in this study was whether being bilingual protects against age-related cognitive decline (Bialystok, 2017; Perani et al., 2017).
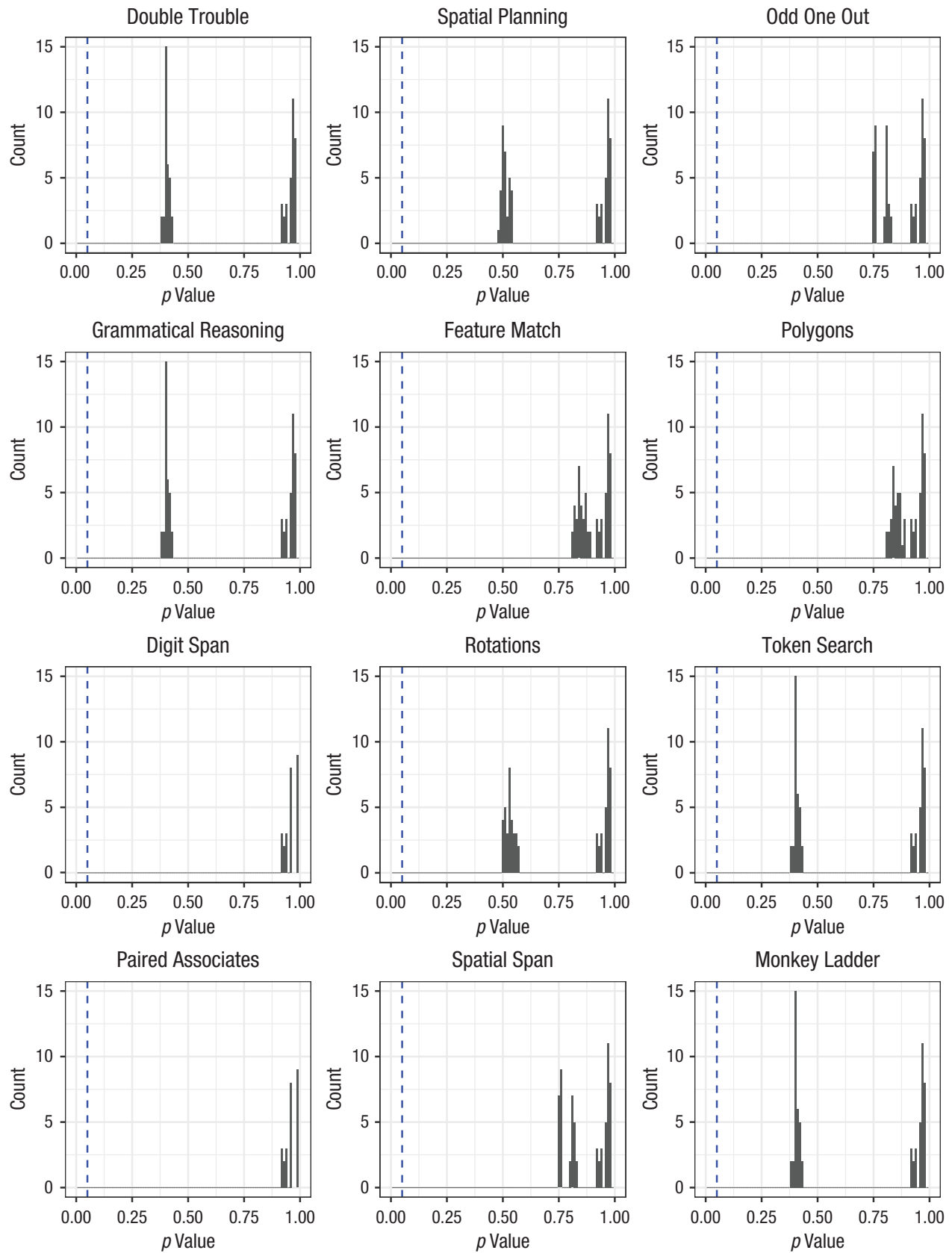
Fig. 5. Distributions of *p* values for each test over 64 models in the matched sample. The dashed blue line indicates a *p* value of .05.
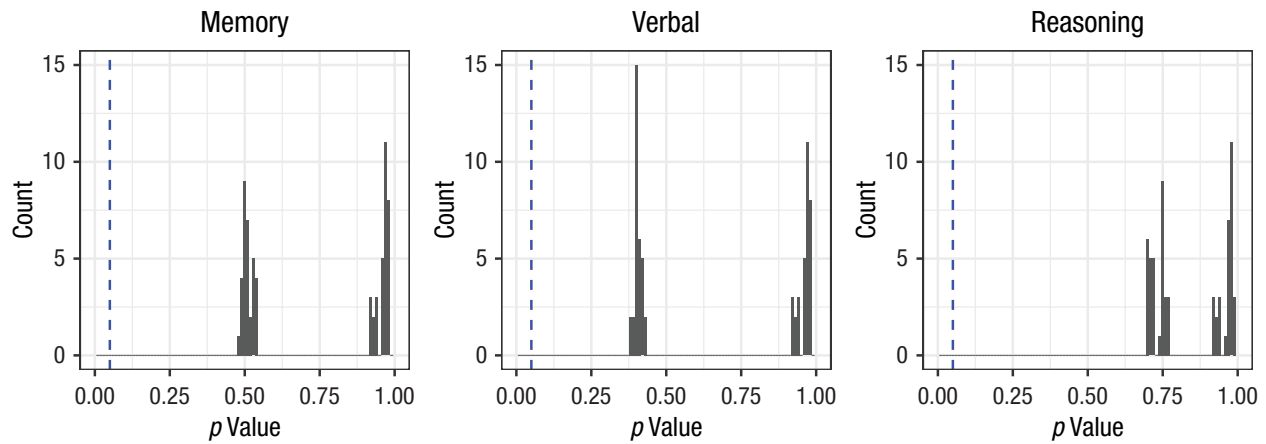
Fig. 6. Distributions of *p* values for each factor over 64 models in the matched sample. The dashed blue line indicates a *p* value of .05.

The interaction between group (bilinguals vs. monolinguals) and age showed no relationship in both our age-, education-, SES-, and language-matched subgroup and in our full, unmatched sample. Therefore, this study provides no support for such protective effects, even in tests that are sensitive to age-related decline (e.g., Paired Associates and Polygons). Indeed, Bayes factors for all tests showed substantial or strong support for the null hypothesis.

Through model selection, we were able to identify which regressors needed to be included to provide the best fit to the data. This also showed that the set of regressors included in the model can sometimes lead to a significant result; when groups were not well matched, there were a number of combinations of regressors that led to significant bilingual advantages. We highlight that Double Trouble, a test of inhibition that is a variant of the Stroop task (Stroop, 1935) and one of the tasks most frequently used in bilingual-advantage research, showed a significant result in our unmatched sample 50% of the time, depending on the set of regressors included. This suggests that extreme

Table 4. Parameters for the Best-Fitting Model in the Unmatched Sample and the Percentage of Significant Results in 64 Model Iterations

Variable	Regressors	BIC	Significant group *p* values (%)	Significant age *p* values (%)	Significant interaction *p* values (%)
Tests					
Double Trouble	Education + Gender + Age × Group	30,190.2	50	100	0
Spatial Planning	Gender + Age × Group	30,027.5	0	100	0
Odd One Out	Education + Age × Group	31,272.2	0	0	0
Grammatical Reasoning	SES + Education + Gender + Age × Group	30,776.1	62.5	100	12.5
Feature Match	SES + Age × Group	31,285.2	12.5	100	0
Polygons	SES + Age × Group	30,762.0	0	100	0
Digit Span	Education + Gender + Age × Group	30,720.1	100	100	0
Rotations	SES + Age × Group	30,001.9	25	100	0
Token Search	Education + Gender + Age × Group	30,006.8	25	100	0
Paired Associates	Gender + Age × Group	30,690.1	18.75	100	0
Spatial Span	Gender + Age × Group	30,935.4	0	100	0
Monkey Ladder	Gender + Age × Group	30,548.7	25	100	0
Factors					
Memory	Gender + Age × Group	30,116.8	0	100	0
Verbal	Education + Age × Group	30,972.9	87.5	100	0
Reasoning	SES + Age × Group	30,303.5	25	100	0

Note: BIC = Bayesian information criterion; SES = socioeconomic status.

Table 5. Group Regression Parameters for the Best-Fitting Model Following Model Selection in the Unmatched Sample

Task	ΔR^2	Group β	t	p	99% CI β	BF_{01}	η_p^2
Tests							
Double Trouble	< .01	−0.04	$t(11031) = -0.60$	.111	[−0.08, 0.01]	637.05	< .01
Spatial Planning	< .01	−0.01	$t(11035) = 1.25$	.585	[−0.06, 0.04]	1,839.80	< .01
Odd One Out	< .01	−0.02	$t(11033) = -0.83$	.508	[−0.24, 0.14]	1,741.56	< .01
Grammatical Reasoning	< .01	−0.19	$t(11030) = -9.98$	< .001	[−0.24, −0.14]	$3.54e^{-20}$	.01
Feature Match	< .01	−0.04	$t(11036) = -2.30$	.046	[−0.09, < 0.01]	738.73	< .01
Polygons	< .01	< −0.01	$t(11036) < 0.01$	.994	[−0.05, 0.05]	9,948.14	< .01
Digit Span	< .01	0.05	$t(11031) = 2.52$	.029	[< −0.01, 0.10]	168.69	< .01
Rotations	< .01	−0.11	$t(11036) = -5.86$	< .001	[−0.16, −0.06]	$1.15e^{-5}$	< .01
Token Search	< .01	−0.11	$t(11031) = -5.83$	< .001	[−0.16, −0.06]	$1.11e^{-4}$	< .01
Paired Associates	< .01	0.01	$t(11035) = 0.61$	.585	[−0.18, 0.19]	8,020.24	< .01
Spatial Span	< .01	−0.02	$t(11035) = -1.13$	.351	[−0.07, 0.03]	5,428.21	< .01
Monkey Ladder	< .01	0.03	$t(11035) = 1.41$	.240	[−0.02, 0.07]	2,109.20	< .01
Factors							
Memory	< .01	0.03	$t(11035) = 1.41$	.240	[−0.02, 0.07]	3,052.83	< .01
Verbal	< .01	−0.06	$t(11033) = -3.33$	.003	[−0.11, −0.01]	4.48	< .01
Reasoning	< .01	−0.10	$t(11036) = -5.57$	< .001	[−0.15, −0.06]	$1.20e^{-3}$	< .01

Note: BF_{01} is the Bayes factor showing the likelihood of the null over the alternative hypothesis. CI = confidence interval.

caution in regressor selection must be taken when testing whether bilinguals show cognitive benefits over monolinguals, as spurious results can occur, potentially explaining the large discrepancy in the literature over whether such effects exist.

Despite these results, several potential caveats need to be considered. First, is it possible that the 12 tasks did not assess aspects of cognition that are relevant to a potential bilingual advantage? This is very unlikely, as versions of most of the tests have been used in previous work demonstrating the cognitive benefits of bilingualism. For example, Double Trouble is a version of the Stroop task and a measure of inhibition that has been used extensively in this research area (Bialystok

Table 6. Interaction Regression Parameters for the Best-Fitting Model Following Model Selection in the Unmatched Sample

Variable	ΔR^2	Interaction β	t	p	99% CI β	BF_{01}	η_p^2
Tests							
Double Trouble	< .01	< −0.01	$t(11031) = -1.54$	.273	[< −0.01, < 0.01]	32.27	< .01
Spatial Planning	< .01	< 0.01	$t(11035) = 1.78$	.234	[< −0.01, < 0.01]	21.72	< .01
Odd One Out	< .01	< −0.01	$t(11033) = -1.76$	.234	[< −0.01, < 0.01]	22.18	< .01
Grammatical Reasoning	< .01	< 0.01	$t(11030) = 2.61$	.096	[< −0.01, < 0.01]	3.46	< .01
Feature Match	< .01	< 0.01	$t(11036) = 0.28$	.781	[< −0.01, < 0.01]	101.10	< .01
Polygons	< .01	< 0.01	$t(11036) = 0.46$	.738	[< −0.01, < 0.01]	94.70	< .01
Digit Span	< .01	< 0.01	$t(11031) = 1.51$	.273	[< −0.01, < 0.01]	33.36	< .01
Rotations	< .01	< 0.01	$t(11036) = 2.49$	.096	[< −0.01, < 0.01]	4.74	< .01
Token Search	< .01	< 0.01	$t(11031) = 1.46$	.273	[< −0.01, < 0.01]	36.38	< .01
Paired Associates	< .01	< −0.01	$t(11035) = -0.50$	.738	[−0.01, 0.02]	92.52	< .01
Spatial Span	< .01	< −0.01	$t(11035) = -0.40$	.738	[< −0.01, < 0.01]	96.96	< .01
Monkey Ladder	< .01	< −0.01	$t(11035) = -1.11$	.443	[< −0.01, < 0.01]	56.59	< .01
Factors							
Memory	< .01	< −0.01	$t(11035) = -0.72$	.640	[< −0.01, < 0.01]	80.89	< .01
Verbal	< .01	< 0.01	$t(11033) = 2.00$	.229	[< −0.01, 0.01]	14.30	< .01
Reasoning	< .01	< 0.01	$t(11036) = -0.75$	.640	[< −0.01, < 0.01]	78.55	< .01

Note: BF_{01} is the Bayes factor showing the likelihood of the null over the alternative hypothesis. CI = confidence interval.
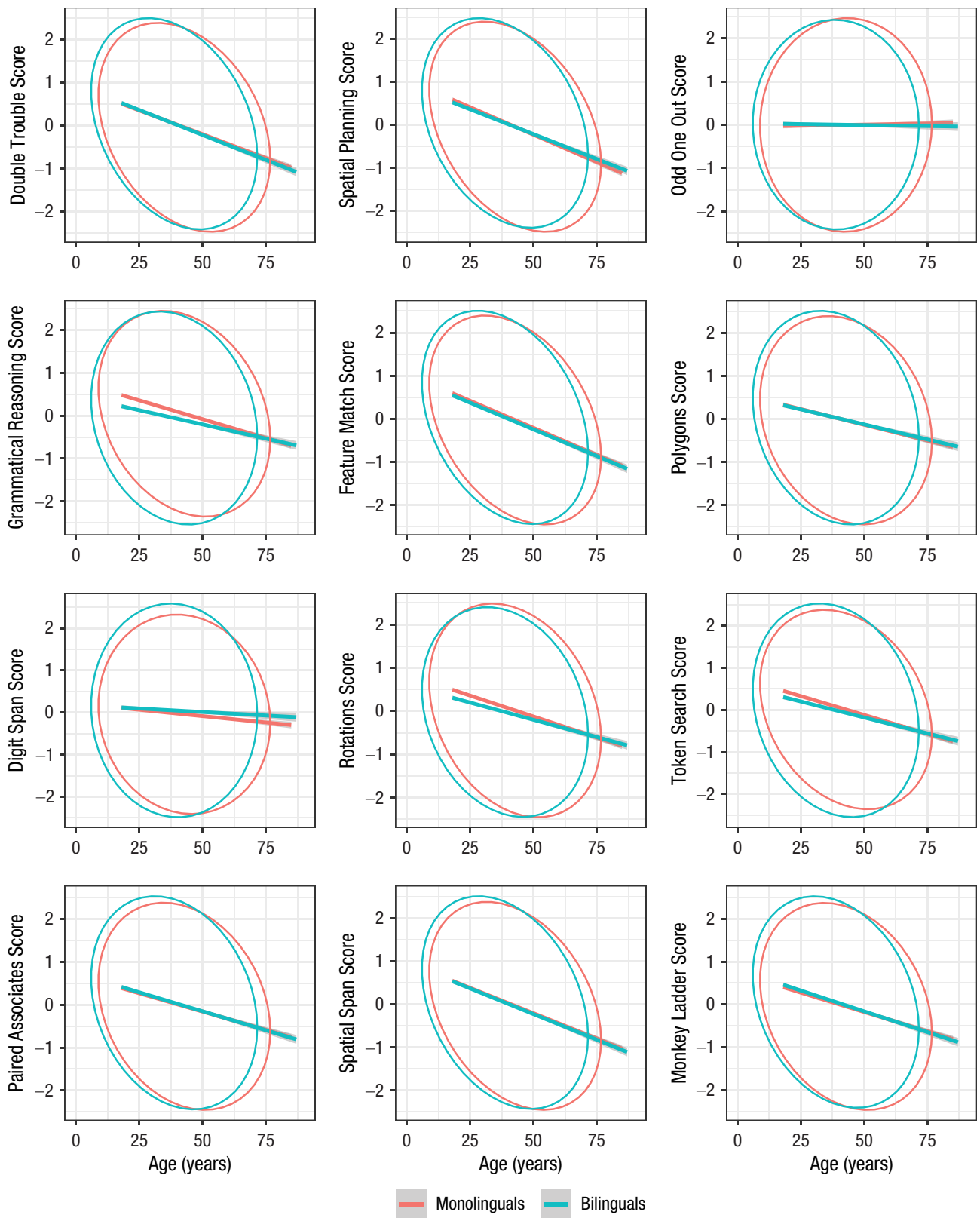
Fig. 7. Plots showing the linear relationship between age and scores for each of the tests in the unmatched sample. For each regression line, a 95% confidence ellipse and a 95% confidence interval is shown. Because the groups were not age matched, the monolingual ellipse begins and extends farther right than the bilingual ellipse in each of the plots. Individual data points have not been included because of the large sample size.
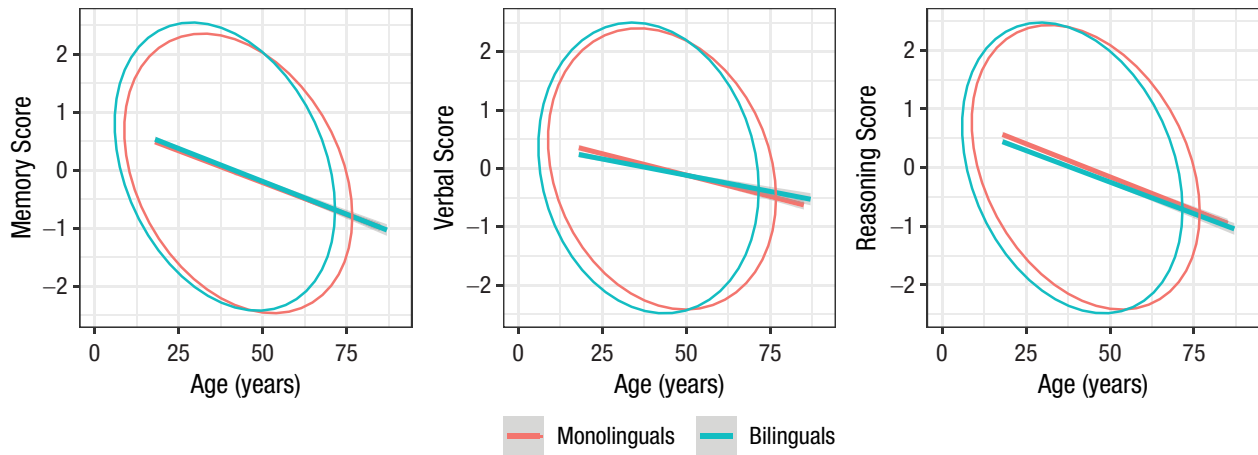
Fig. 8. Plots showing the linear relationship between age and each of the three factors in the unmatched sample. For each regression line, a 95% confidence ellipse and a 95% confidence interval is shown. Individual data points have not been included because of the large sample size.

et al., 2008; Blumenfeld & Marian, 2014). Similarly, spatial tasks have been used to show a bilingual advantage (Morales et al., 2013), but none of our spatial tasks had a significant group effect. The battery of tests employed was also cognitively broad; many of the tasks required executive function (Bor et al., 2006; Owen et al., 1990, 1995), and all required aspects of attention and working memory. If there were benefits to any of these processes afforded by bilingualism, it is reasonable to expect that they would be expressed through differences in performance on some, or all, of these tasks. It is of course possible that differences between monolinguals and bilinguals would have been observed if we had used a different set of cognitive tasks entirely, although in the context of the available literature on bilingualism and executive function, it is not at all clear what those tasks would have been, nor what executive processes they would have tapped.

Second, is it possible that the 12 tasks included in the battery are simply not sensitive to the subtle effects of bilingualism? This is also extremely unlikely, because the tasks have previously been shown to be highly sensitive to subtler cognitive differences related to disease or pharmacological intervention. For example, the test of planning (the Hampshire Tree Task) is sensitive to performance differences between specific genotypes in early Parkinson's disease (Williams-Gray et al., 2007); tests of paired-associates learning, such as the one employed in this study, are able to distinguish between first-episode schizophreniform psychosis and established schizophrenia (Wood et al., 2002); and the Token Search task used here has been used to detect increases in spatial working memory in children with attention-deficit/hyperactivity disorder following a low dose of methylphenidate (Mehta et al., 2000). More importantly,

however, the sheer sample size of more than 11,000 participants makes it extremely unlikely that a genuine effect of bilingualism on executive function would have been missed if it were there.

Third, is it possible that the observed results occurred because the two samples were not perfectly matched with respect to age, SES, and education? This is not the case, as the effects of all three factors were controlled by including them as variables of no interest. However, even if this statistical procedure did not adequately control their effects, separate analyses run on an age-, SES-, and education-matched subsample again provided absolutely no evidence for a bilingual advantage, although age effects remained.

Finally, whereas previous studies have shown that online testing produces results that are comparable with those acquired in more traditional lab-based settings (Hampshire et al., 2012), it is possible that inaccurate reporting of demographic information and test scores led to data that were too noisy for differences to emerge. However, when we imposed strict cleaning procedures in the matched subsample, ensuring that our bilingual sample met several criteria for bilingualism, the effects seen in the unmatched sample disappeared completely.

These results demonstrate that, across a broad battery of cognitive tasks of executive function, no systematic differences exist between monolinguals and bilinguals. When groups were poorly matched, a difference on Digit Span was detected, although given the modest size of this effect in terms of the performance advantage it affords and the weak support for the difference, its real world relevance is questionable.

We conclude by emphasizing, however, that despite the fact that no meaningful relationship was found
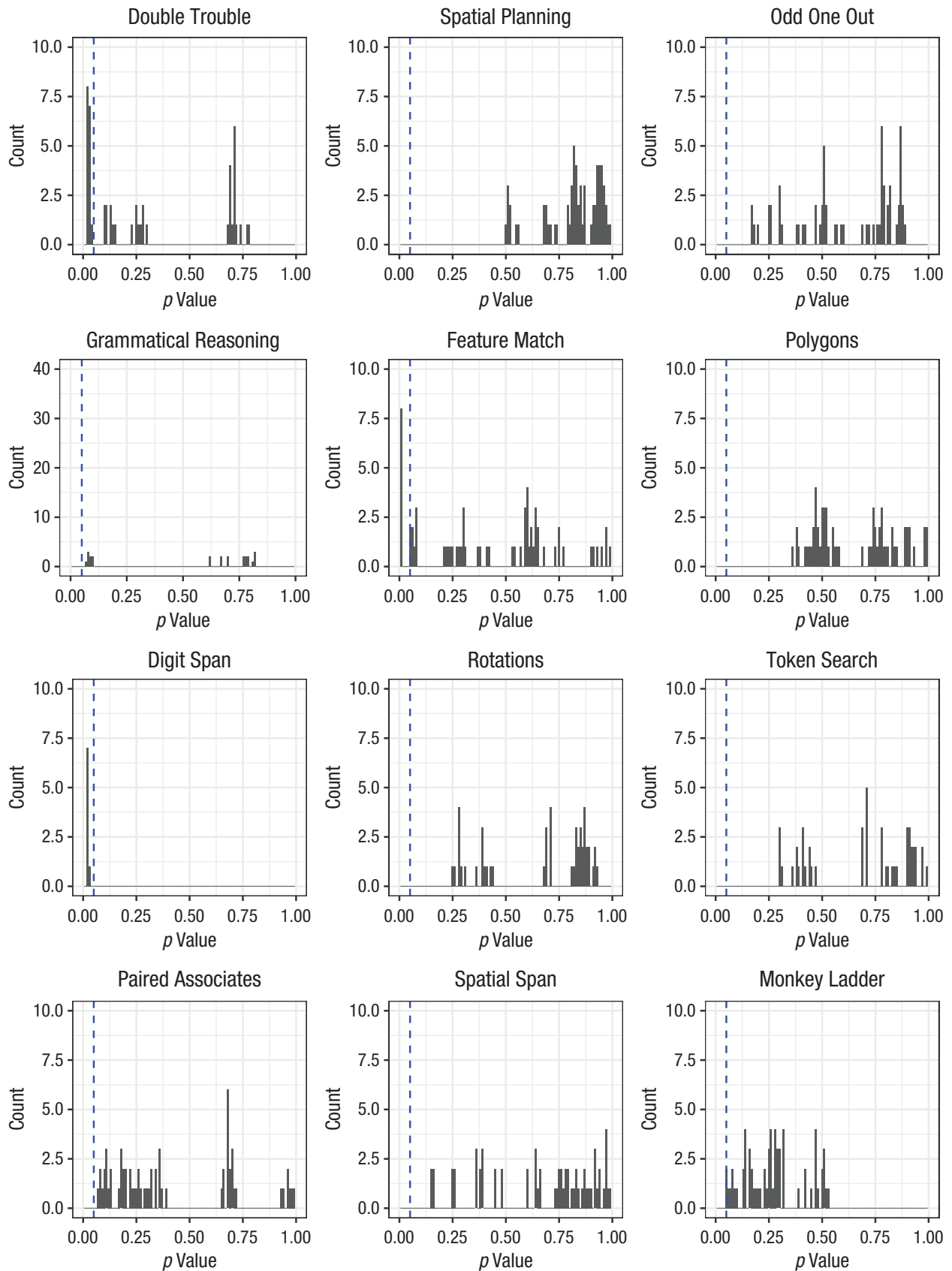
Fig. 9. Distributions of *p* values for each test over 64 models in the unmatched sample. The dashed blue line indicates a *p* value of .05.
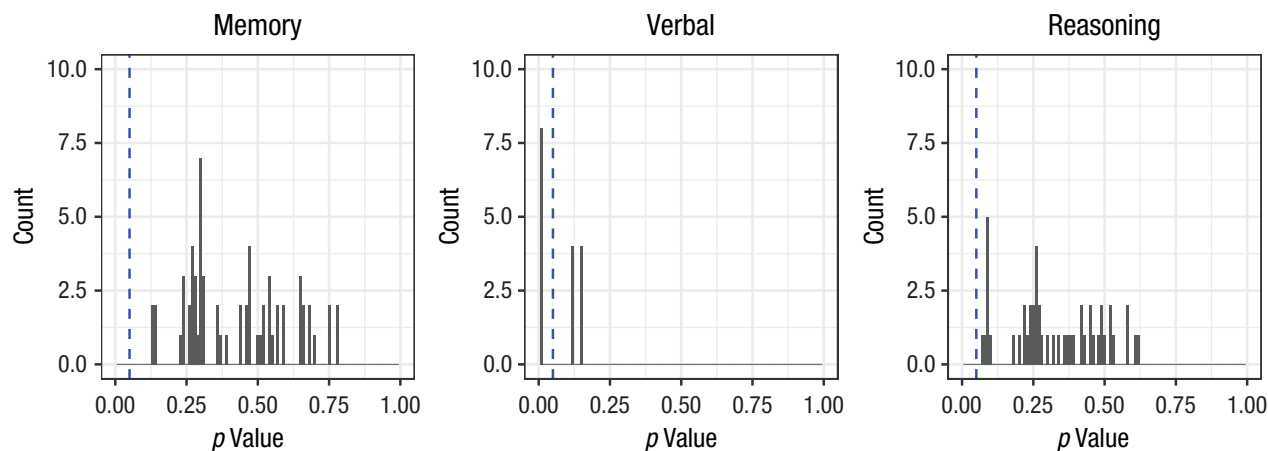
Fig. 10. Distributions of *p* values for each factor over 64 models in the unmatched sample. The dashed blue line indicates a *p* value of .05.

between bilingualism and executive function, the broader social, employment, and lifestyle benefits that are available to speakers of a second language are clearly numerous.

Transparency

Action Editor: Charles Hulme
Editor: D. Stephen Lindsay
Author Contributions
E. S. Nichols codesigned the study, analyzed and interpreted the data, and took overall responsibility for writing each draft of the manuscript. C. J. Wild codesigned the study, designed the data checking and cleaning protocols, was responsible for converting data into a format for analysis, and contributed to each draft of the manuscript. B. Stojanoski and M. E. Battista codesigned the study and contributed to each draft of the manuscript. A. M. Owen codesigned the study and contributed to each draft of the manuscript. All authors approved the final version of the manuscript for submission.
Declaration of Conflicting Interests
The cognitive tests used in this study are marketed by Cambridge Brain Sciences, of which A. M. Owen is the Chief Scientific Officer. Under the terms of the existing licensing agreement, A. M. Owen and his collaborators are free to use the platform at no cost for their scientific studies, and such research projects neither contribute to, nor are influenced by, the activities of the company. Consequently, there is no overlap between the current study and the activities of Cambridge Brain Sciences, nor was there any cost to the authors, funding bodies, or participants who were involved in the study. The authors declared that there were no other potential conflicts of interest with respect to the authorship or the publication of this article.
Open Practices
Data for this study have not been made publicly available, and the design and analysis plans were not preregistered.

Materials can be obtained from Cambridge Brain Sciences (www.cambridgebrainsciences.com).

ORCID iD

Emily S. Nichols [ID] https://orcid.org/0000-0003-0541-9233

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797620903113

References

Baddeley, A. D. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science*, *10*, 341–342. doi:10.3758/BF03331551

Barton, K. (2019). MuMIn: Multi-Model Inference (Version 1.43.6). Retrieved from https://CRAN.R-project.org/package=MuMIn

Bialystok, E. (2015). Bilingualism and the development of executive function: The role of attention. *Child Development Perspectives*, *9*, 117–121. doi:10.1111/cdep.12116

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin*, *143*, 233–262. doi:10.1037/bul0000099

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 859–873. doi:10.1037/0278-7393.34.4.859

Blumenfeld, H. K., & Marian, V. (2014). Cognitive control in bilinguals: Advantages in stimulus–stimulus inhibition. *Bilingualism: Language and Cognition*, *17*, 610–629. doi:10.1017/S1366728913000564

Bor, D., Duncan, J., Lee, A. C. H., Parr, A., & Owen, A. M. (2006). Frontal lobe involvement in spatial span:

Converging studies of normal and impaired function. *Neuropsychologia*, *44*, 229–237. doi:10.1016/j.neuropsychologia.2005.05.010

Brito, N. H., Murphy, E. R., Vaidya, C., & Barr, R. (2016). Do bilingual advantages in attentional control influence memory encoding during a divided attention task? *Bilingualism: Language and Cognition*, *19*, 621–629. doi:10.1017/S1366728915000851

Cattell, R. B. (1949). The dimensions of culture patterns by factorization of national characters. *The Journal of Abnormal and Social Psychology*, *44*, 443–469. doi:10.1037/h0054760

Collins, P., Roberts, A. C., Dias, R., Everitt, B. J., & Robbins, T. W. (1998). Perseveration and strategy in a novel spatial self-ordered sequencing task for nonhuman primates: Effects of excitotoxic lesions and dopamine depletions of the prefrontal cortex. *Journal of Cognitive Neuroscience*, *10*, 332–354.

Curtin, J. (2018). lmSupport: Support for linear models (R package Version 2.9.13). Retrieved from https://CRAN.R-project.org/package=lmSupport

de Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science*, *26*, 99–107. doi:10.1177/0956797614557866

Festman, J., Rodriguez-Fornells, A., & Münte, T. F. (2010). Individual differences in control of language interference in late bilinguals are mainly related to general executive abilities. *Behavioral and Brain Functions*, *6*, Article 5. doi:10.1186/1744-9081-6-5

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198. doi:10.1016/0022-3956(75)90026-6

Gould, R. L., Brown, R. G., Owen, A. M., Bullmore, E. T., Williams, S. C. R., & Howard, R. J. (2005). Functional neuroanatomy of successful paired associate learning in Alzheimer's disease. *American Journal of Psychiatry*, *162*, 2049–2060.

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

Gunzenhauser, C., Karbach, J., & Saalbach, H. (2019). Function of verbal strategies in monolingual vs. bilingual students' planning performance: An experimental approach. *Cognitive Development*, *50*, 1–12. doi:10.1016/j.cogdev.2019.01.003

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

Hernández, M., Costa, A., Fuentes, L. J., Vivas, A. B., & Sebastián-Gallés, N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Bilingualism: Language and Cognition*, *13*, 315–325. doi:10.1017/S1366728909990010

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8). doi:10.18637/jss.v042.i08

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, *17*, 1004–1005.

Kempe, V., Kirk, N. W., & Brooks, P. J. (2015). Revisiting theoretical and causal explanations for the bilingual advantage in executive functioning. *Cortex*, *73*, 342–344. doi:10.1016/j.cortex.2015.07.021

Kerrigan, L., Thomas, M. S. C., Bright, P., & Filippi, R. (2017). Evidence of an advantage in visuo-spatial memory for bilingual compared to monolingual speakers. *Bilingualism: Language and Cognition*, *20*, 602–612. doi:10.1017/S1366728915000917

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*, 394–425.

Macnamara, B. N., & Conway, A. R. A. (2014). Novel evidence in support of the bilingual advantage: Influences of task demands and experience on cognitive control and working memory. *Psychonomic Bulletin and Review*, *21*, 520–525. doi:10.3758/s13423-013-0524-y

Mehta, M. A., Owen, A. M., Sahakian, B. J., Mavaddat, N., Pickard, J. D., & Robbins, T. W. (2000). Methylphenidate enhances working memory by modulating discrete frontal and parietal lobe regions in the human brain. *The Journal of Neuroscience*, *20*(6), Article RC65.

Morales, J., Calvo, A., & Bialystok, E. (2013). Working memory development in monolingual and bilingual children. *Journal of Experimental Child Psychology*, *114*, 187–202. doi:10.1016/j.jecp.2012.09.002

Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science*, *10*, 719–726. doi:10.1111/j.1467-7687.2007.00623.x

Noble, K. G., Norman, F. M., & Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental Science*, *8*(1), 74–87.

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., & Robbins, T. W. (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, *28*, 1021–1034. doi:10.1016/0028-3932(90)90137-D

Owen, A. M., Sahakian, B. J., Semple, J., Polkey, C. E., & Robbins, T. W. (1995). Visuo-spatial short-term recognition memory and learning after temporal lobe excisions, frontal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia*, *33*(1), 1–24.

Paap, K. R., Johnson, H. A., & Sawi, O. (2016). Should the search for bilingual advantages in executive functioning continue? *Cortex*, *74*, 305–314. doi:10.1016/j.cortex.2015.09.010

Perani, D., Farsad, M., Ballarini, T., Lubian, F., Malpetti, M., Fracchetti, A., . . . Abutalebi, J. (2017). The impact of bilingualism on brain reserve and metabolic connectivity in Alzheimer's dementia. *Proceedings of the National Academy of Sciences, USA*, *114*, 1690–1695. doi:10.1073/pnas.1610909114

Ratiu, I., & Azuma, T. (2015). Working memory capacity: Is there a bilingual advantage? *Journal of Cognitive Psychology*, *27*(1), 1–11. doi:10.1080/20445911.2014.976226

R Core Team. (2019). *R: A language and environment for statistical computing.* Retrieved from https://www.R-project.org/

Rosselli, M., Ardila, A., Lalwani, L. N., & Vélez-Uribe, I. (2016). The effect of language proficiency on executive functions in balanced and unbalanced Spanish-English bilinguals. *Bilingualism: Language and Cognition*, *19*, 489–503. doi:10.1017/S1366728915000309

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *298*, 199–209. doi:10.1098/rstb.1982.0082

Silverman, I. I., Choi, J., Mackewn, A., Fisher, M., & Olshansky, E. (2000). Evolved mechanisms underlying wayfinding: Further studies on the hunter-gatherer theory of spatial sex differences. *Evolution and Human Behavior*, *21*, 201–213.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Treisman, A. M., & Gelade, G. (1980). Engineering education—status quo in Austria in comparison with the academic field of business education. *Cognitive Psychology*, *12*, 97–136.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*, 779–804. doi:10.3758/BF03194105

Wechsler, D. (1981). The psychometric tradition: Developing the Wechsler Adult Intelligence Scale. *Contemporary Educational Psychology*, *6*, 82–85. doi:10.1016/0361-476X(81)90035-7

Wesnes, K. A., Brooker, H., Ballard, C., Mccambridge, L., Stenton, R., & Corbett, A. (2017). Utility, reliability, sensitivity and validity of an online test system designed to monitor changes in cognitive function in clinical trials. *International Journal of Geriatric Psychiatry*, *32*, e83–e92. doi:10.1002/gps.4659

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* New York, NY: Springer-Verlag.

Wild, C. J., Nichols, E. S., Battista, M. E., Stojanoski, B., & Owen, A. M. (2018). Dissociable effect of self-reported daily sleep duration on high-level cognitive abilities. *Sleep*, *41*, Article zsy182. doi:10.1093/sleep/zsy182

Williams-Gray, C. H., Hampshire, A., Robbins, T. W., Owen, A. M., & Barker, R. A. (2007). Catechol O-methyltransferase val-158met genotype influences frontoparietal activity during planning in patients with Parkinson's disease. *The Journal of Neuroscience*, *27*, 4832–4838. doi:10.1523/JNEUROSCI.0774-07.2007

Wood, S. J., Proffitt, T., Mahony, K., Smith, D. J., Buchanan, J. A., Brewer, W., . . . Pantelis, C. (2002). Visuospatial memory and learning in first-episode schizophreniform psychosis and established schizophrenia: A functional correlate of hippocampal pathology? *Psychological Medicine*, *32*, 429–438. doi:10.1017/S0033291702005275