LETTERS
https://doi.org/10.1038/s41562-019-0609-3

nature human behaviour

Corrected: Author Correction; Author Correction

No evidence for a bilingual executive function advantage in the ABCD study

Anthony Steven Dick[1*], Nelcida L. Garcia[1], Shannon M. Pruden[1], Wesley K. Thompson[2], Samuel W. Hawes[1], Matthew T. Sutherland[1], Michael C. Riedel[1], Angela R. Laird[1] and Raul Gonzalez[1]

Learning a second language in childhood is inherently advantageous for communication. However, parents, educators and scientists have been interested in determining whether there are additional cognitive advantages. One of the most exciting yet controversial[1] findings about bilinguals is a reported advantage for executive function. That is, several studies suggest that bilinguals perform better than monolinguals on tasks assessing cognitive abilities that are central to the voluntary control of thoughts and behaviours—the so-called 'executive functions' (for example, attention, inhibitory control, task switching and resolving conflict). Although a number of small-[2–4] and large-sample[5,6] studies have reported a bilingual executive function advantage (see refs. [7–9] for a review), there have been several failures to replicate these findings[10–15], and recent meta-analyses have called into question the reliability of the original empirical claims[8,9]. Here we show, in a very large sample ($n = 4{,}524$) of 9- to 10-year-olds across the United States, that there is little evidence for a bilingual advantage for inhibitory control, attention and task switching, or cognitive flexibility, which are key aspects of executive function. We also replicate previously reported disadvantages in English vocabulary in bilinguals[7,16,17]. However, these English vocabulary differences are substantially mitigated when we account for individual differences in socioeconomic status or intelligence. In summary, notwithstanding the inherently positive benefits of learning a second language in childhood[18], we found little evidence that it engenders additional benefits to executive function development.

A question commonly asked by parents, educators and scientists is whether the benefits of learning a second language outweigh the potential costs. Research on this topic is important for clarifying these issues; for example, it was once believed that exposure to two languages would confuse children, but this is now an antiquated idea that has few contemporary proponents[17]. In fact, almost two decades ago, researchers found that bilingualism might confer advantages in other cognitive domains, such as executive function[19]. This positive benefit of bilingualism was championed and replicated many times[20]. However, these effects have also been vigorously questioned because of several failures to replicate such findings, as well as claims that the findings are simply artefacts of small- and non-representative-sample studies that do not adequately control for potential confounds[1,8,9,15,21,22].

In the present paper, we conduct a large-sample study of the hypothesis that exposure to multiple languages in childhood is associated with better executive functioning. We present the results of analyses conducted on 4,524 9- to 10-year-old children from the Adolescent Brain and Cognitive Development (ABCD) study. Data from this sample were generated from 21 study sites across the United States, and approximate the demographic profile of the American Community Survey (ACS)[23] (see Supplementary Table 1). This sample also has a substantial number of bilingual children ($n = 1{,}740$) speaking more than 40 different languages other than English (although the majority speak Spanish as a second language; see Supplementary Table 2). It is an ideal sample on which to test the claim that bilingual children show an advantage over monolinguals for executive function development. The age range investigated in the sample is one that has been investigated in a number of studies of bilingual advantages for executive function[24–26], and it also meets a definition of early bilingualism (that is, second language use before 10 years of age), which has been associated with both reduced English vocabulary and better executive function[7,25]. Thus we expected that, if the effects were real and replicable in a large sample, we should find them in this age range and in this sample of children.

We began our analysis by identifying bilingual children in three different ways, based on the ABCD Youth Acculturation Survey (YAS). The first definition, 'Bilingual Status', simply established whether children spoke another language in addition to English. From this group of children, we identified which bilingual children also used the non-English language frequently. This established a group of 606 children who were consistently using the non-English language with friends and family (that is, at least equally or more often), which we defined as our 'Bilingual Degree' variable. Finally, we established a 'Bilingual Use' variable, which was a more continuous measure of how often children were using the other language with friends and family. Children who almost exclusively used the other language with friends and family scored high on this variable, which had good representation at all levels along the continuum.

Next, we conducted a number of regression analyses with the aim of replicating different approaches that have been used to address these questions in the literature. The first set of regressions employed generalized additive mixed models (GAMMs), which modelled family nested within site as random effects, but controlled for no covariates. For these analyses, we established as dependent variables well-validated measures of English vocabulary (the National Institutes of Health (NIH) Toolbox Picture Vocabulary Test[27]; that is, 'English vocabulary') and executive function. We used three executive function measures: (1) the NIH Toolbox Flanker Inhibitory Control and Attention Test[28] (that is, 'flanker'; a measure of inhibitory control and attention); (2) the NIH Toolbox Dimensional Change Card Sort (DCCS)[28] (that is, 'card sort'; a measure of task switching/cognitive flexibility); and (3) the stop-signal

[1]Florida International University, Miami, FL, USA. [2]University of California, San Diego, San Diego, CA, USA. *e-mail: adick@fiu.edu

Table 1 | Descriptive statistics for all four measures for each of the three definitions of bilingualism

Bilingual Status

Measure	Monolingual mean (s.d.)	Bilingual mean (s.d.)
Vocabulary	85.9 (7.9)	85.0 (8.1)
Flanker	94.7 (8.9)	95.1 (8.7)
Card sort	93.6 (9.1)	93.6 (9.0)
SSRT	−299.3 (78.4)	−302.7 (78.6)

Bilingual Degree

Measure	Monolingual mean (s.d.)	Bilingual mean (s.d.)
Vocabulary	85.9 (7.9)	81.3 (7.5)
Flanker	94.7 (8.9)	93.7 (9.4)
Card sort	93.6 (9.1)	92.2 (8.6)
SSRT	−299.3 (78.4)	−307.0 (78.7)

Bilingual Use

Measure	Level	Mean (s.d.)	Measure	Level	Mean (s.d.)
Vocabulary	0	87.2 (7.7)	Card sort	0	93.9 (9.4)
	1	86.6 (7.7)		1	94.6 (8.9)
	2	85.0 (8.2)		2	93.6 (8.8)
	3	82.9 (8.1)		3	93.3 (8.7)
	4	80.8 (7.1)		4	92.1 (8.5)
	5	79.0 (6.1)		5	91.5 (8.2)
	6	78.5 (6.5)		6	91.0 (7.3)
	7	78.0 (7.5)		7	91.2 (5.6)
	8	76.3 (5.5)		8	82.4 (16.6)
Flanker	0	95.5 (8.3)	SSRT	0	−298.1 (80.7)
	1	96.3 (8.1)		1	−303.6 (77.5)
	2	95.5 (8.1)		2	−307.5 (75.8)
	3	94.9 (9.0)		3	−301.4 (81.7)
	4	93.4 (10.2)		4	−304.9 (76.4)
	5	91.9 (9.1)		5	−312.7 (82.2)
	6	89.6 (11.0)		6	−307.6 (68.1)
	7	93.7 (7.7)		7	−297.7 (53.4)
	8	97.3 (8.9)		8	−352.1 (43.0)

SSRT is a response time measure which was reverse-scored such that higher values indicate better performance. NIH Toolbox measures are uncorrected scores according to the NIH Toolbox scoring for each measure (see Methods). Levels 0–8 represent the level of Bilingual Use variable (0 = speaks English all of the time with friends and family; 8 = speaks other language all of the time with friends and family).

task (SST)[29] (an additional measure of inhibitory control and attention). These three measures represent domains of executive function for which differences between monolinguals and bilinguals have been found in the literature[20], and for which theoretical explanations for the differences have been put forth[30–32]. The analyses thus examine three predictors (Bilingual Status, Bilingual Degree and Bilingual Use) against four outcome variables (English vocabulary, flanker, card sort and stop-signal reaction time (SSRT); see Table 1 for descriptive statistics).

A model accounting for no covariates does not equate the groups across a number of confounds that could possibly explain monolingual versus bilingual differences, and this has been a strong point of debate in the literature. Because the sample is large and because a number of demographic measures were collected, we were able to take advantage of a multilevel modelling statistical framework that accurately accounts for these covariates, while also modelling and controlling for individual differences across subjects that might be driven by different cultural and family environments. For the last four sets of regressions in Table 2, GAMMs were constructed to tease apart which effects could be explained by group differences in language status, and not by differences in other factors (for example, age, biological sex, race/ethnicity, highest degree of education, household income, marital status, crystallized and fluid intelligence, and English vocabulary).

As Table 2 shows, for the GAMM models with no covariates, we observed a disadvantage for bilingual children for English vocabulary, and in one case there was a significant bilingual advantage for executive function (for Bilingual Status predicting flanker). This finding was still significant when only English vocabulary was controlled. However, when more focused measures of bilingualism (that is, Bilingual Degree and Bilingual Use) were used, the significant differences in the executive function measures actually showed a bilingual disadvantage (lower scores on flanker, card sort and SSRT; note that SSRT was recoded such that higher scores reflect better performance). Moreover, after we controlled for the demographic covariates, only the English vocabulary differences, and one effect for SSRT (lower SSRT for bilinguals), remained significant. All of these effects showed a bilingual disadvantage.

Table 2 | Results of GAMM regression for the three predictors of interest

Predictor	Outcome	d.f.	*B* (s.e.)	*β*	*t*	*P*
GAMM (without covariates)						
Bilingual Status	Vocabulary	4,447	−0.23 (0.25)	−0.01	−0.94	0.347
Bilingual Degree	Vocabulary	3,331	−3.70 (0.41)	−0.18	−9.08	<0.001
Bilingual Use	Vocabulary	1,722	−1.21 (0.12)	−0.27	−10.34	<0.001
Bilingual Status	Flanker	4,444	0.71 (0.29)	0.04	2.47	0.014
Bilingual Degree	Flanker	3,329	−0.37 (0.47)	−0.02	−0.78	0.435
Bilingual Use	Flanker	1,719	−0.46 (0.13)	−0.09	−3.55	<0.001
Bilingual Status	Card sort	4,445	0.43 (0.29)	0.02	1.47	0.142
Bilingual Degree	Card sort	3,329	−0.54 (0.47)	−0.02	−1.15	0.250
Bilingual Use	Card sort	1,720	−0.29 (0.14)	−0.06	−2.14	0.032
Bilingual Status	SSRT	3,398	−3.02 (2.86)	−0.02	−1.06	0.289
Bilingual Degree	SSRT	2,532	−7.72 (4.81)	−0.04	−1.61	0.108
Bilingual Use	SSRT	1,311	−2.69 (1.37)	−0.06	−1.97	0.049
GAMM (with the English vocabulary covariate only)						
Bilingual Status	Flanker	4,443	0.76 (0.28)	0.04	2.75	0.006
Bilingual Degree	Flanker	3,328	0.57 (0.46)	0.02	1.23	0.219
Bilingual Use	Flanker	1,718	−0.17 (0.13)	−0.03	−1.36	0.174
Bilingual Status	Card sort	4,444	0.50 (0.28)	0.03	1.77	0.077
Bilingual Degree	Card sort	3,328	0.52 (0.45)	0.02	1.16	0.246
Bilingual Use	Card sort	1,719	0.09 (0.14)	0.02	0.64	0.522
Bilingual Status	SSRT	3,364	−3.63 (2.85)	−0.02	−1.27	0.204
Bilingual Degree	SSRT	2,511	−6.17 (4.90)	−0.03	−1.26	0.208
Bilingual Use	SSRT	1,294	−2.45 (1.39)	−0.05	−1.75	0.080
GAMM (with covariates 1–6)						
Bilingual Status	Vocabulary	4,434	0.16 (0.24)	0.01	0.67	0.502
Bilingual Degree	Vocabulary	3,318	−2.05 (0.43)	−0.10	−4.81	<0.001
Bilingual Use	Vocabulary	1,709	−0.65 (0.11)	−0.14	−5.71	<0.001
Bilingual Status	Flanker	4,431	0.57 (0.29)	0.03	1.94	0.053
Bilingual Degree	Flanker	3,316	−0.04 (0.54)	−0.002	−0.08	0.936
Bilingual Use	Flanker	1,706	−0.25 (0.14)	−0.05	−1.81	0.070
Bilingual Status	Card sort	4,432	0.33 (0.30)	0.02	1.10	0.271
Bilingual Degree	Card sort	3,316	−0.05 (0.53)	−0.002	−0.10	0.920
Bilingual Use	Card sort	1,707	−0.05 (0.14)	−0.01	−0.37	0.711
Bilingual Status	SSRT	3,385	−2.7 (3.0)	−0.02	−0.92	0.358
Bilingual Degree	SSRT	2,519	−8.1 (5.7)	−0.04	−1.42	0.156
Bilingual Use	SSRT	1,298	−3.0 (1.5)	−0.07	−1.99	0.047
GAMM (with covariates 1–8)						
Bilingual Status	Vocabulary	4,412	−0.64 (0.21)	−0.04	−3.08	0.002
Bilingual Degree	Vocabulary	3,299	−2.65 (0.35)	−0.13	−7.70	<0.001
Bilingual Use	Vocabulary	1,703	−0.63 (0.10)	−0.14	−6.25	<0.001
Bilingual Status	Flanker	4,411	0.48 (0.26)	0.03	1.83	0.067
Bilingual Degree	Flanker	3,299	0.26 (0.43)	0.01	0.60	0.549
Bilingual Use	Flanker	1,702	−0.12 (0.12)	−0.02	−0.97	0.332
Bilingual Status	Card sort	4,412	0.02 (0.25)	0.0008	0.06	0.952
Bilingual Degree	Card sort	3,299	−0.04 (0.39)	−0.002	−0.09	0.928
Bilingual Use	Card sort	1,703	0.06 (0.13)	0.01	0.44	0.660
Bilingual Status	SSRT	3,338	−4.4 (2.8)	−0.03	−1.54	0.124
Bilingual Degree	SSRT	2,487	−8.4 (5.0)	−0.04	−1.70	0.089
Bilingual Use	SSRT	1,277	−2.2 (1.4)	−0.05	−1.53	0.126

Continued

Table 2 | Results of GAMM regression for the three predictors of interest (continued)

Predictor	Outcome	d.f.	*B* (s.e.)	*β*	*t*	*P*
GAMM (with covariates 1–9)						
Bilingual Status	Flanker	4,410	0.50 (0.26)	0.03	1.90	0.058
Bilingual Degree	Flanker	3,298	0.37 (0.44)	0.02	0.85	0.395
Bilingual Use	Flanker	1,701	−0.11 (0.12)	−0.02	−0.86	0.390
Bilingual Status	Card sort	4,411	0.05 (0.25)	0.003	0.21	0.834
Bilingual Degree	Card sort	3,298	0.11 (0.39)	0.005	0.29	0.772
Bilingual Use	Card sort	1,702	0.11 (0.13)	0.02	0.86	0.390
Bilingual Status	SSRT	3,337	−4.6 (2.8)	−0.03	−1.63	0.103
Bilingual Degree	SSRT	2,486	−9.1 (5.0)	−0.04	−1.81	0.070
Bilingual Use	SSRT	1,276	−2.6 (1.5)	−0.06	−1.80	0.072

The GAMMs included family nested within site as random effects and the following covariates as fixed effects: (1) age; (2) biological sex; (3) race or ethnicity; (4) highest degree of education; (5) household income; (6) marital status; (7) crystallized intelligence; (8) fluid intelligence; and (9) English vocabulary. Crystallized intelligence was measured using the NIH Toolbox Oral Reading Recognition Test, and fluid intelligence was measured using the Modified NIH Toolbox Fluid Cognition Composite Test. The SSRT was reverse-scored. *B* is the unstandardized regression slope parameter estimate, s.e. is the standard error of the regression slope parameter estimate, and *β* is the standardized regression slope parameter estimate. All *P* values are two-tailed.

Thus, in our regression analyses, we replicated the disadvantage for English vocabulary in bilingual children reported in the literature[24]. However, when we controlled properly for covariates, we failed to find a bilingual advantage for executive function.

One concern is that, in the classic null hypothesis testing framework, failure to find a difference does not imply equivalence, because the negative result may simply result from a lack of power (that is, a type II error[33]). Similarly, in studies with large sample sizes, as we have here, a negligible difference may be shown to be statistically significant (known as statistical overpowering[34]). Thus, in cases where one wants to argue for the absence of a meaningful (that is, pragmatically or theoretically significant) effect, an alternative test is an equivalence test. Here, the null hypothesis is specified to say that a difference between parameter estimates is outside an a priori interval of equivalence (δ). If the observed confidence interval (CI) of the parameter estimate lies within the a priori region, the null hypothesis that the effect is large enough to be worthwhile is rejected. We chose the interval $\delta = -0.1$ to $+0.1$ to define a meaningful difference, which is consistent with the notion of a small effect size for correlation according to Cohen's standards[35]. Thus, we statistically tested for the absence of effects large enough to be deemed worthwhile. The results of the equivalence tests are shown in Fig. 1.

Figure 1 shows that we failed to evidence equivalence in 13 cases. For the executive function variables, non-equivalence was found for all three outcome measures, but in all cases bilinguals actually showed a disadvantage. When covariates were entered in the model, we failed to find evidence for equivalence only for the SSRT (again, bilinguals showed a disadvantage), but these effects were still quite small (the largest standardized β for the executive function measures, after controlling for covariates, was −0.07, indicating a bilingual disadvantage). A notable finding was that the English vocabulary disadvantage was still evident after controlling for a number of demographic covariates. The results of the equivalence analyses thus mirror those reported for the regression analyses. We found no meaningful evidence for an executive function bilingual advantage, but we did find evidence for a disadvantage for bilinguals for English vocabulary, indicating a small effect size ($\beta = -0.14$ when controlling for demographic and intelligence covariates).

To put these findings in context, a further discussion of the effect sizes is warranted. First, in the case of the first two predictors (Bilingual Status and Bilingual Degree), examination of the unstandardized regression coefficient tells us how much the dependent measure changes given a change in status from monolingual to bilingual. For the English vocabulary measure, the effects ranged from a reduction of 0.23 (Cohen's $d = 0.03$) to 3.7 points (Cohen's $d = 0.41$), but when covariates were controlled, the reduction was a 2.7-point difference between monolinguals and bilinguals ($t(3,299) = -7.70$; $P < 0.001$; Cohen's $d = 0.34$; 95% CI of $d = 0.26$–0.44). This was a statistically significant difference, but the Cohen's d for this difference was rather small (Cohen provides a guideline for a small effect as $d = 0.20$; ref. [35]). For comparison, in the norming study of the NIH Toolbox Picture Vocabulary Test, the effect size difference between college and high school-educated adults was large (Cohen's $d = 0.98$; ref. [27]).

Effect size differences for the executive function measures were even smaller. For example, when examining data with no covariates, the only significant effect showing a bilingual advantage was of Bilingual Status predicting flanker ($t(4,444) = 2.47$; $P = 0.014$; Cohen's $d = 0.08$; 95% CI of $d = 0.02$–0.14)—an effect that remained when English vocabulary was controlled ($t(4,443) = 2.75$; $P = 0.006$; Cohen's $d = 0.08$; 95% CI of $d = 0.02$–0.14) but not when other demographic covariates were controlled ($t(4,431) = 1.94$; $P = 0.053$; Cohen's $d = 0.06$; 95% CI of $d = -0.0006$–0.12). These effect sizes are very small by any reasonable standard, and mirror the small effects reported in other large-sample studies of bilingual executive function advantages (for example, Hartanto and colleagues[5] report a standardized β of 0.04 for the card-sort task in their large-sample study, which translates to Cohen's $d = 0.08$). To put these effects in context, in a norming study of the NIH Toolbox card-sort and flanker tests, the effect size difference between college- and high-school-educated adults was much larger (Cohen's $d = 0.39$ and 0.44, respectively[36]). Our reported effect sizes are even smaller than the sex differences reported in the same norming study, which are themselves small and non-significant (in the norming study, Cohen's $d = 0.10$ and 0.13 for card sort and flanker, respectively). In the present study, the small effect sizes for bilingual variables predicting executive function are very small and are unlikely to reflect a meaningful difference in the population.

To summarize the findings thus far, the results from our regression and equivalency analyses suggest that any bilingual executive function advantage is too small to consistently detect, even with our large sample. This raises the question of why the bilingual executive function advantage has been replicated so often in the literature. It has been suggested that other factors—namely, publication bias[37]—may have favoured the publication of positive effects in studies with small samples. Indeed, there is evidence from recent meta-analyses that this may be the case in the bilingual literature[8,9]. The

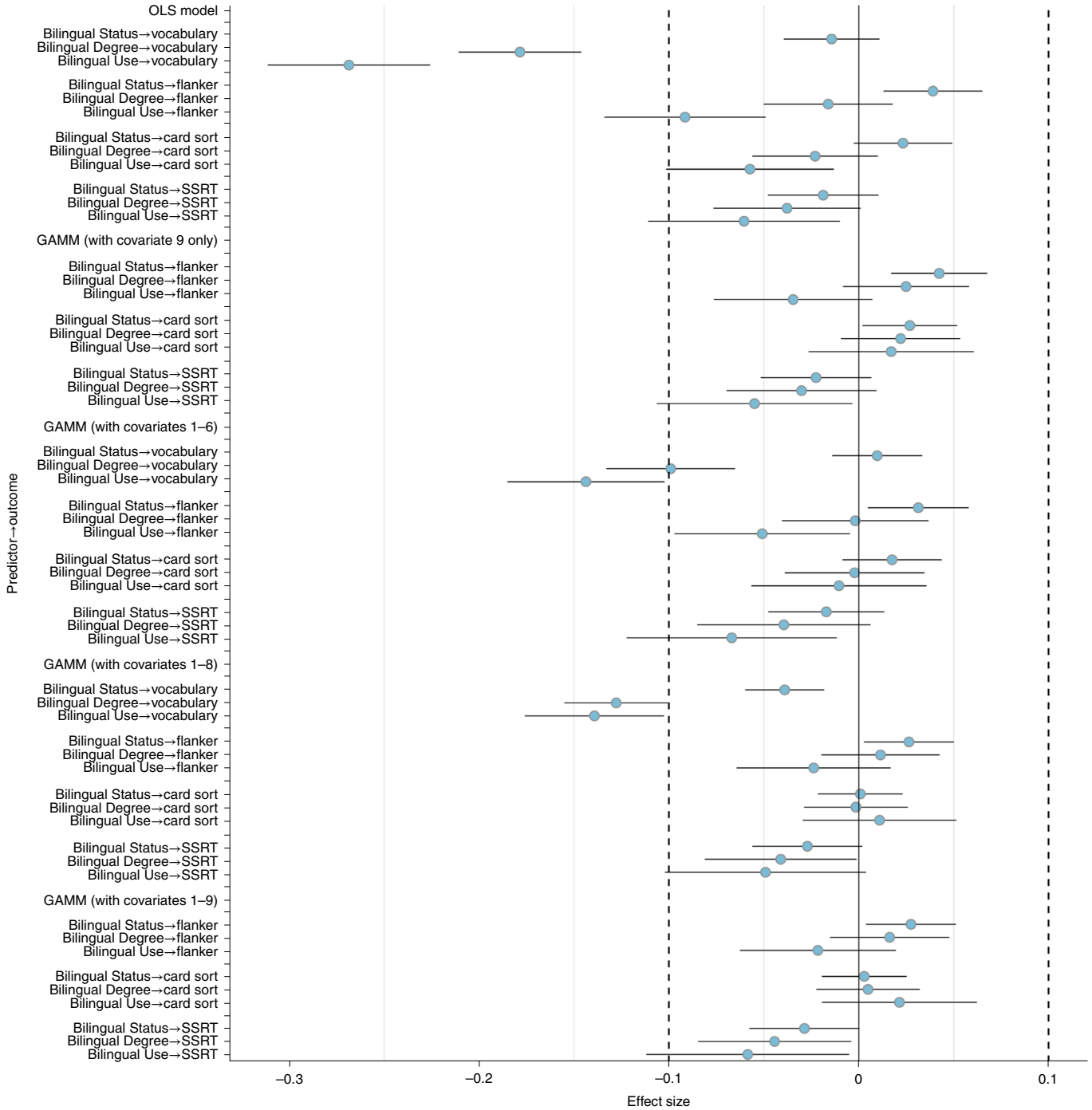

Fig. 1 | Results of the tests of equivalence for the standardized regression slope β. The figure shows the effect sizes (β) and CIs plotted against the interval of equivalence. Data points represent the parameter estimates (β) from Table 2, along with the calculated CIs for the slope (C_{β}^{-} and C_{β}^{+}). To evidence statistical equivalence, CIs should be contained within the a priori defined interval of equivalence (I_{β}^{-} to I_{β}^{+}), which was set to (−0.1 to 0.1). Slope estimates at 0 would be exactly equivalent. A bilingual advantage would show to the right of 0 on the *x* axis, and any disadvantage to the left. The β values reported from the GAMMs included family nested within site as random effects and the covariates listed in Table 2.

large sample of the present study allows us to assess this possibility using a bootstrap approach. In the bootstrap, we can repeatedly take smaller samples from our larger sample and plot a distribution of the effects that turn out to be statistically significant, which will show us the frequency of significant effects gleaned from small samples taken from our data.

We can also explore another question about the expected distribution of significant results from a population in which the effects are real and replicable, as opposed to one in which the effects are probably due to chance. On the assumption that the error distribution is normal, it is expected that effects would be statistically significant about 5% of the time (that is, this is expected based on the type I error rate set by our cut-off of $\alpha = 0.05$, and should occur in about 250 out of 5,000 bootstrap samples). If the null hypothesis is true (that is, there is no effect), this distribution should be uniform[38]. If the null hypothesis is false (that is, there is an effect), this distribution should be non-uniform[39,40] and indeed should be right-skewed (that is, more low *P* values (for example, 0.01 s) than

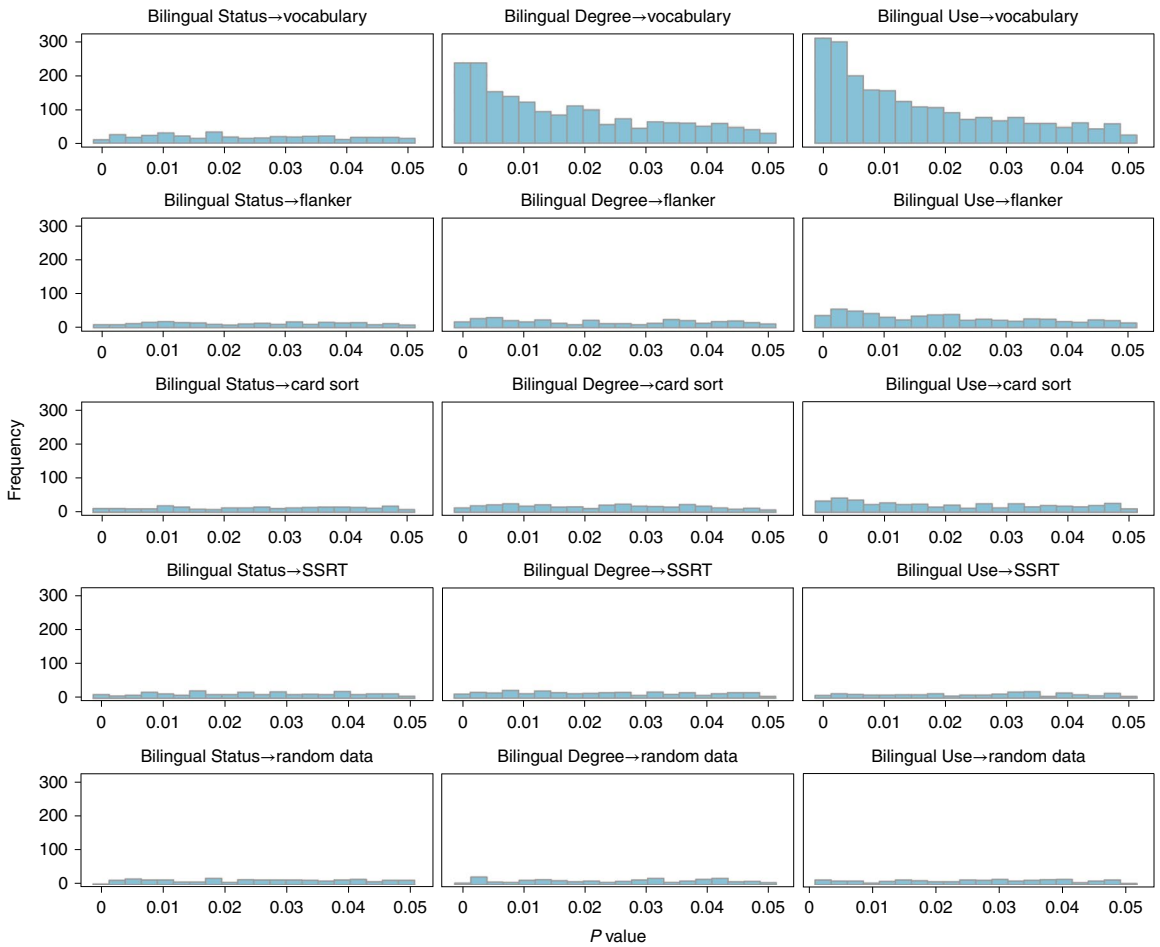

Fig. 2 | Histograms representing the frequency of *P* values for $n = 30$ of 5,000 bootstrap replicates, for OLS regressions with no covariates. Predictors are listed with outcomes. Significant results (that is, $P < 0.05$) were relatively frequent for Bilingual Degree and Bilingual Use predicting English vocabulary. Significant results were less frequent for the flanker, card sort and SSRT outcome measures, although they still occurred. Notably, they also occurred for data randomly generated from a normal distribution (bottom row). These random data also show the expected uniform distribution under the assumption that the null hypothesis is true. The executive function measures, unlike the English vocabulary measure, also show a uniform distribution, suggesting that the results derived from the executive function measures are not different from those derived from random data.

high *P* values (for example, 0.04 s))[38]. We can show this empirically as well by bootstrapping random data using our bilingual predictors. For random data, we would expect a uniform distribution because no real effects should be detected in random data. Here, we are essentially plotting a *P* curve—the distribution of statistically significant *P* values—from the available data[38].

Figure 2 shows, as expected, that the distribution for random data was uniform (row five), and with small sample sizes ($n = 30$) even findings from random data could manifest as significant effects. This was expected based on the type I error rate set by $\alpha = 0.05$. In contrast, when effect sizes were larger, as was the case for English vocabulary predicted by Bilingual Use and Bilingual Degree, the distribution became non-uniform (row 1 in Fig. 2), suggestive of a real effect. However, when effect sizes were small, as was the case for all of the executive function measures (rows 2–4), the distribution was uniform, and resembled that of randomly generated data. This suggests that no real effect is driving significance for these comparisons, which complements the conclusions reached in the regression and equivalency analyses. It also suggests that previously reported significant effects in the literature for executive function may reflect type I error, but effects for English vocabulary probably reflect a true difference between bilinguals and monolinguals.

In summary, in one of the largest studies to date addressing this question, we failed to find consistent evidence for a bilingual advantage for executive function. Although the size and demographic profile of the sample suggests that our findings have a high likelihood of replicability, it is limited by the measures we used and the population we studied. Thus, it is accurate within this context to note that the previously reported bilingual advantage for executive function is: (1) either very small or non-existent; (2) not present at 9–10 years of age and only either earlier or later; (3) not measurable using the operationalizations of 'cool' executive function on offer from the NIH Toolbox and SST; and (4) not present in the specific bilingual sample, which might be advantaged or disadvantaged in a number of ways specific to the United States cultural, linguistic or educational context. Thus, with these caveats as a context, we must entertain the possibility that bilingual executive function advantages might be revealed at different developmental time points[6], using different tasks, or in children raised in a different cultural context.

Despite the non-replication of previously reported findings of a bilingual advantage for executive function, we did replicate the disadvantage for bilinguals in terms of English vocabulary. Our replication of this finding is important because it lends support to our method for operationalization of bilingualism. However, these results should not be taken to endorse the idea that learning a second language is disadvantageous—it is in fact advantageous in a number of domains[18]. It is also important to note that while English vocabulary may be reduced for bilinguals compared with

monolinguals, the effects are small to modest, explaining around 1–5% of additional variance in English vocabulary, depending on the statistical model, and when proper controls are considered. To put this in context, in a norming study of the NIH Toolbox Picture Vocabulary Test, including socioeconomic status in the statistical model is a more robust predictor, accounting for an additional 6.3% of the variance in scores[41]. Perhaps more importantly, there is little evidence to suggest that bilinguals have lower total vocabulary than monolinguals[16,17]. With these caveats in mind, we think the difference in English vocabulary is worth consideration in future studies of bilingualism, as there may be ways to mitigate such effects for dual-language learners. More broadly, the present study contributes to the discussion of important issues surrounding the education of bilingual children, such as whether and to what extent there should be concern about language and literacy development, and whether and to what extent the cognitive benefits of bilingual education extend outside the domain of second-language proficiency.

Methods

Data analysis was conducted on the ABCD Study Curated Annual Release 1.0. Comprehensive details about the ABCD study are published elsewhere (ref. [42] and other articles in the same focus issue). Data collection and analysis were performed blind to the conditions of the experiments. The study was reviewed and approved by the University of California, San Diego's Institutional Review Board.

Participants. The sample comprised data on 4,524 9- to 10-year-old children, collected from 21 study sites across the United States. Demographically, the ABCD study used a multistage probability sample of eligible children by probability sampling schools within the catchment area of each site. The goal of this sampling strategy was to match the demographic profile of two national surveys—the ACS (a large-scale survey of approximately 3.5 million households conducted annually by the US Census Bureau) and annual third- and fourth-grade school enrolment data maintained by the National Center for Education Statistics. The sampling strategy was additionally constrained by the requirement that study sites had available magnetic resonance imaging (MRI) scanners. Because these are typically available at research universities in urban areas, the sampling tended to oversample urban as opposed to rural students and families.

Despite this caveat, the ABCD study sample was largely successful at approximating the ACS survey demographic profiles[23]. That said, although it approximates the demographic profile of the ACS survey, because the sampling strategy heavily relied on schools in urban areas, it is more accurate to describe the sample as having a population-based, demographically diverse sample that is not necessarily representative of the U.S. national population[43]. Demographic assessments of the sample are summarized in Barch et al.[44]. The demographic profile of the sample is presented in Supplementary Table 1.

Measures. For the present study, we used measures of Bilingual Status and other language use, attention and executive function, English vocabulary, and fluid and crystallized intelligence.

Measurement of bilingualism. Measurement of bilingualism is challenging and multifaceted, and there is no established or consistent measure[45]. Previous studies have variously used self-[46] and parent/caregiver-[2] reports, more-detailed language background questionnaires[3,11,12,46–48] or a ratio of vocabulary scores in the two languages[49]. In the present study, bilingualism was measured using the self-report ABCD YAS (a modified version of the PhenX Acculturation Measure; https://www.phenxtoolkit.org)[50], which provided a measure of whether the child spoke more than one language, as well as how often this language was spoken with friends and family. The bilingual variables were calculated and used as predictors of executive function and English vocabulary.

Three variables were calculated based on the ABCD YAS. The first was Bilingual Status (a categorical variable), and consisted of a categorical answer to the question 'Besides English, do you speak or understand another language or dialect?' Participants also answered the question 'What other language or dialect do you speak or understand (besides English)?'. A dropdown menu was available, and participants were allowed to choose 'other' if their language was not represented. If they spoke more than two languages, they were instructed to answer the language (other than English) that they spoke the most. Supplementary Table 2 provides a breakdown of the responses to these questions. Note that a small number of children answered 'pig Latin' or 'English' ($n=8$). Because these are either dialects of English or 'pretend languages', the answers were recoded such that participants were counted as monolingual. After recoding, 2,761 participants were identified as monolingual and 1,740 as bilingual (23 provided no response).

This Bilingual Status variable is useful and replicates the measurement from previous studies, but it does not provide any detail about the familiarity with the other language. Thus, we calculated a second categorical variable, Bilingual Degree, to assess the degree to which the child used the other language. This was based on the answers to two questions. The first question was 'What language do you speak with most of your friends?', with answers applied on a Likert scale consisting of 'Other language all the time', 'Other language most of the time', 'Other language and English equally', 'English most of the time' and 'English all the time'. The second question was 'What language do you speak with most of your family?', with the same answer choices. Bilingual Degree was calculated as a categorical variable dummy-coded to include participants as bilingual if they endorsed that they spoke another language other than English, and if they endorsed that they spoke this other language with friends all of the time, most of the time or equally, or they spoke the other language with family all of the time, most of the time or equally. After this coding, 2,761 participants were identified as monolingual, and 606 participants were identified as having a high degree of exposure to and use of the second language.

Finally, to obtain a more continuous measure of Bilingual Use, we summed the answers to the 'What language do you speak with most of your friends?' and 'What language do you speak with most of your family?' questions, and reverse scored the answers (for a range of 0–8). Thus, a child who endorsed 'English all of the time' for both friends and family would score low on this measure (0), while a child who endorsed 'Other language all of the time' (4) for both friends and family would receive a high score (8).

Children were also asked 'How well do you speak English?'. The majority of children (98%) endorsed 'good' or 'excellent' for this question. When looking within the monolingual and bilingual groups, both groups reported mostly good or excellent (monolingual: excellent = 77.7%; good = 20.5%; fair = 1.2%; poor = 0.5%; bilingual: excellent = 71.2%; good = 25.7%; fair = 2.9%; poor = 0.1%). However, although the effect was small, the rates were significantly different, ($\chi^2(3)=39.8$; $P<0.001$; Cramer's $\phi_c=0.05$; 95% CI of $\chi^2=17.3$–66.1). Across both groups, there was also a significant association between self-reporting of English proficiency and English vocabulary ($F(3, 4{,}466)=25.34$; $P<0.001$; $\eta_p^2=0.02$; 90% CI of $\eta_p^2=0.011$–0.023). For children who reported being bilingual, there was an additional association between self-reporting of English proficiency and the Bilingual Use measure ($F(3, 1{,}743)=54.06$; $P<0.001$; $\eta_p^2=0.09$; 90% CI of $\eta_p^2=0.06$–0.11). These results suggest that while most children were proficient in English (measured via self-reports), it is important to control for English language proficiency when examining any group differences. We did this using the standardized English vocabulary measure (see below).

NIH Toolbox measures and SST. Measures of English vocabulary, executive function, and fluid and crystallized intelligence were administered from the Cognition Battery of the NIH Toolbox[51], in addition to the SST measuring inhibitory control. The NIH Toolbox measures were administered on an iPad with a touchscreen. The SST was administered in the MRI scanner. Because demographic variables were entered in the regressions, the uncorrected score was used for all NIH Toolbox measures. The SSRT was the outcome measure for the SST.

Measurement of English vocabulary. English vocabulary was measured using the NIH Toolbox Picture Vocabulary test[27]. In this test, single words are presented auditorily, paired simultaneously with four images of objects, actions and/or depictions of concepts. The participant must select the picture with the meaning that most closely matches that of the spoken word. Items are scored as correct or incorrect.

Measurement of executive function. Because many of the previously reported differences between bilinguals and monolinguals have been on tasks assessing attention, task switching/cognitive flexibility and inhibitory control, we analysed the NIH Toolbox Flanker Inhibitory Control and Attention Test and ABCD SST[29] (to measure inhibitory control and attention), as well as the NIH Toolbox Dimensional Change Card Sort (DCCS) Test to measure task switching/cognitive flexibility[52].

The NIH Toolbox card-sort task is based on the Dimensional Change Card Sort (DCCS) task developed by Zelazo and colleagues[28]. In the standard version of the DCCS, children are shown two target cards (for example, red rabbits and blue boats) and asked to sort the test cards (for example, blue rabbits and red boats) first according to one dimension (for example, colour) and then according to the other (for example, shape). Because of the conflict between the target and test cards, switching between the first rule and the second incurs a switch cost, both in terms of accuracy and response time. In the NIH Toolbox version, there are four card-sorting blocks: practice, preswitch, postswitch and mixed. In the practice block, participants were instructed to match centrally presented stimuli to one of two lateralized target stimuli. The preswitch and postswitch trials were similar to the practice trials, but there was a conflict between the test and target cards. A rule switch was employed between the pre- and postswitch phase, with the sorting dimension (shape or colour) counterbalanced across participants. Children who succeeded on at least four trials of the postswitch received the mixed block, which consisted of 50 trials of 40 frequent and 10 infrequent trials. The frequent trials corresponded to the dimension that had been presented in the postswitch phase.

The standard toolbox scoring was used, in which both accuracy and response time were included in the score[52].

The NIH Toolbox Flanker Inhibitory Control and Attention Test was also administered in the standard fashion. In this task, participants were required to indicate the left versus right orientation of a centrally presented stimulus while inhibiting attention to the flankers that surrounded it on each side. On some trials, the orientation of the flankers was congruent with the orientation of the central stimulus, while on others it was incongruent. The NIH Toolbox version consists of a practice block, a block using child-friendly fish stimuli and a block using more difficult arrow stimuli. Scoring was similar to the card-sort scoring, and incorporated both accuracy and response time.

Administration of the ABCD SST is described in detail elsewhere[29]. Briefly, the SST requires that participants withhold a motor response to a 'go' stimulus when it is followed unpredictably by a signal to stop. For the ABCD study, the SST was administered inside an MRI scanner over 2 experimental runs of 180 trials each. On each trial, a leftward- or rightward-pointing arrow was presented (that is, the go stimulus). Participants were instructed to press a button corresponding to the direction of the arrow. A proportion (16.67%) of the trials were stop trials, on which the arrow was followed by an upward-facing arrow indicating that the participant should withhold responding. These trials were unpredictable, requiring participants to inhibit the prepotent response to go. The stop-signal delay (that is, the time between the onset of the go trial and the stop trial) began at 50 s, but was adaptively modified in response to participant performance. This was designed to equate task difficulty across individuals. The primary measure of interest was SSRT, which was proposed to index inhibitory control, and was computed here by taking the mean go response time and subtracting the mean stop-signal delay[53]. To be consistent with the other measurements, we reverse scored the SSRT (by multiplying the values by −1) so that higher SSRT indicated better inhibitory control. Notably, because some children were fatigued by the length of the MRI scanner protocol, attrition data on the SST were only available for 75% of the sample. Thus, the results are reported for the sample of children who completed the task.

Measurement of fluid and crystallized intelligence. Measurements of fluid (adaptive problem-solving skills) and crystallized intelligence (accumulated knowledge through experience) were accomplished as part of the NIH Toolbox Cognition Battery[54]. Specifically, we used modified versions of the Toolbox Crystallized Cognition Composite and Toolbox Fluid Cognition Composite scores. The Toolbox Crystallized Cognition Composite is typically derived from two subtests of the cognition battery: the Picture Vocabulary Test and Oral Reading Recognition Test. The Toolbox Fluid Cognition Composite is typically derived from five subtests: the Card Sort, Flanker Inhibitory Control and Attention Test, Picture Sequence Memory Test, List Sorting Working Memory Test and Pattern Comparison Processing Speed Test. However, because some of our outcome measures were included in the calculation of the standard composite scores, we computed new composites by removing those variables (that is, vocabulary was removed from the Toolbox Crystallized Cognition Composite, and card sort and flanker were removed from the Toolbox Fluid Cognition Composite). For crystallized intelligence, we simply used the remaining subtest—oral reading recognition—which is already a standardized score. For fluid intelligence, we averaged the Picture Sequence Memory Test, List Sorting Working Memory Test and Pattern Comparison Processing Speed Test scores, and standardized those average scores.

Missing data. With the exception of the SST measure, missing data as a percentage of the sample were minimal (see Supplementary Table 3). As missing data imputation for dependent measures is not recommended[55], we focused on dealing with missing data for the demographic measures. For the three missing demographic variables (highest household income, highest household education and race/ethnicity), using logistic regression, we checked for the association between missingness and the outcome measures of interest. Two significant associations were revealed: missingness on the household income measure was significantly predicted by English vocabulary and card sort (respectively, $z(4,477) = -6.23$; $P < 0.001$; odds ratio = 1.43; and $z(4,475) = 2.80$; $P = 0.005$; odds ratio = 1.16). This analysis rules out 'missing completely at random' for this particular predictor, which is the standard assumption underlying modern data imputation methods. On this assumption, we proceeded to missing data imputation for demographic measures using the Multivariate Imputation via Chained Equations (MICE) package in R (version 3.5). For the regression analyses to follow, these additional missing data were dealt with using casewise deletion. Degrees of freedom are reported for each comparison, which account for the missing data. We additionally repeated the analysis without imputation, using casewise deletion, to check for any introduction of potential bias. These results (reported in Supplementary Table 4) are not materially different from those reported in the main article.

GAMMs. Simple and multiple regressions were conducted using GAMMs[56]. Generalized additive models replace the linear form from ordinary least squares (OLS) models with a sum of smooth functions incorporating an iterative scatterplot smoother, and are able to better identify and model nonlinear covariate effects. In the present study, we incorporated a mixed-effects model to model the correlated observations within families and at sites. Thus, for all GAMM models, the random effect was specified to model family nested within site.

We conducted five different analyses to attempt to replicate the approaches used in the existing bilingual literature, and to be comprehensive in our control of potentially confounding covariates which have, in the past, been associated with the outcome measures. For the first set of analyses, we used a simple GAMM without covariates (while incorporating the random effects) to predict the four outcomes of interest (English vocabulary, flanker, card sort and SSRT). The three language measures—Bilingual Status, Bilingual Degree and Bilingual Use—were entered as predictors. This analysis covers the approaches used in many studies in the available literature.

We then conducted additional regressions incorporating covariates. For the second set of models, we only controlled for English vocabulary, and investigated only the executive function measures. In previous studies[2,3,25], controlling for vocabulary has been proposed because of its known association with executive function during development[57–60]. In addition, it is possible that controlling for vocabulary can actually help uncover associations between bilingualism and executive function that might be missed. This is predicated on the assumption that bilinguals who have lower verbal ability on average might perform on par with their monolingual peers, indicating that bilingualism compensates to some degree for what would otherwise be poorer executive function. Controlling for English vocabulary without additional covariates allows for investigation of both possibilities.

For the third set of models, English vocabulary was removed as a covariate, and the demographic covariates of age, biological sex, race/ethnicity, highest household education, household marriage status and highest household income were entered (for a total of six covariates). These demographic measures are also known to be associated with the outcome measures of interest[41,61], particularly in bilingual samples[6,10].

For the fourth set of models, we added the fluid and crystallized composites (entered separately due to differential associations with executive function[62]) for a total of eight covariates. For the fifth set of models, we added the English vocabulary covariate back in for a total of nine covariates. It is important to note that while there is a high correlation between intelligence and vocabulary[63], vocabulary is also a useful proxy (albeit with limitations) of the level of acculturation in minority groups[64], which is itself a potentially important confounding factor in studies of the bilingual advantage for executive function[8].

Equivalence testing. In the null hypothesis testing framework, failure to find a difference does not imply equivalence, because the negative result may simply result from a lack of power (that is, a type II error[33]). Similarly, in studies with large sample sizes, a negligible difference may be shown to be statistically significant (known as statistical overpowering[34]). Thus, in cases where one wants to evidence the absence of an effect that is large enough to be deemed worthwhile, one must use an alternative test. One option is an equivalence test. Here, the null hypothesis is specified such that the effect size is outside an a priori-specified interval of equivalence. This flips the null hypothesis testing framework on its head. In other words, in classic null hypothesis testing, the null hypothesis is that there is no difference, but in equivalence testing, the null hypothesis is that there is a certain difference. Rejecting the null hypothesis, in this case, implies that the true effect is close enough to zero for practical purposes (that is, any difference is too small to be meaningful as defined by the a priori effect size interval).

This can be tested using a variety of methods, but one that has been proposed for regression models is the method by Anderson and Hauck[65], validated using Monte Carlo estimation by Counsell and Cribbie[66]. For this procedure, one must first choose an effect size. We chose the standardized regression coefficient, β, because it has desirable estimation properties[67] and accompanying CIs. Second, one must choose a value of δ. We chose the interval −0.1–0.1, which is consistent with the notion of a small effect size for correlation according to Cohen's standards[35]. It is also the effect size used in the validation study by Counsell and Cribbie[66]. Finally, the equivalence statistic was calculated.

The proposed statistic is given by:

$$P = \phi\left[\frac{|B_1 - B_2| - \delta}{s_{B_1 - B_2}}\right] - \phi\left[\frac{-|B_1 - B_2| - \delta}{s_{B_1 - B_2}}\right] \quad (1)$$

where ϕ represents the standard normal probability function, δ represents the interval of equivalence (for example, 0.1), B represents the regression coefficient, and s represents the standard error. When $P \leq \alpha$, the null hypothesis of a difference is rejected and the conclusion is that the coefficients are equivalent (or, for tests against a slope of zero, not different from zero). Similarly, if the parameter estimate of the difference, along with its CI, fall within the interval of equivalence, the parameter estimates are determined to be equivalent.

Bootstrapping analysis of small samples. A number of previous studies that have reported executive function advantages for bilingual children used small samples, making them prone to statistical error, including type I error[68]. We took advantage of our large sample to determine the probability that such effects are driven

by sampling error in small samples. To do this, we conducted a bootstrapping exercise to establish probability distributions of *P* values for each regression slope, with a sample size of $n = 30$, at 5,000 bootstrap replicates. The sample size was chosen based on the average sample size showing significant language group differences as reported in meta-analyses of the literature[8,69]. For this analysis, we used OLS models with no covariates, as the bootstrap sample size was too small to appropriately model nested effects.

We also assessed the probability distribution of the three predictors when random data were generated for the dependent measures. The expectation is that, because random data should not differ as a function of bilingualism, the distribution should be uniform under the null[39,40]. This was verified in our simulation. This simulation thus also allows the qualitative comparison of data from our dependent measures of interest against randomly generated data, using the same independent measures.

Power estimates. In cases where null effects are reported, care must be taken to avoid type II error. Fortunately, because of the large sample sizes employed in the present study, power was universally high. Thus, a posteriori power analysis for the lowest sample size reported in the paper (d.f. = 1,276) showed that the power was 0.95 to detect a small effect of $r = 0.1$ at $\alpha = 0.05$, based on Cohen[35].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data are from the ABCD Study Curated Annual Release 1.0 and are available on request from the NIMH Data Archive (https://data-archive.nimh.nih.gov/abcd).

Code availability

All software used in the present analysis is open source from the Comprehensive R Archive Network (version 3.5.0; ref. [70]). The R code to replicate the analysis is available at https://github.com/anthonystevendick/bilingual_abcd.

Received: 11 December 2018; Accepted: 12 April 2019;
Published online: 20 May 2019

References

1. Morton, J. B. Still waiting for real answers. *Cortex* **73**, 352–353 (2015).
2. Bialystok, E. Cognitive complexity and attentional control in the bilingual mind. *Child Dev.* **70**, 636–644 (1999).
3. Carlson, S. M. & Meltzoff, A. N. Bilingual experience and executive functioning in young children. *Dev. Sci.* **11**, 282–298 (2008).
4. Prior, A. & MacWhinney, B. A bilingual advantage in task switching. *Biling. Lang. Cogn.* **13**, 253–262 (2010).
5. Hartanto, A., Toh, W. X. & Yang, H. Bilingualism narrows socioeconomic disparities in executive functions and self-regulatory behaviors during early childhood: evidence from the early childhood longitudinal study. *Child Dev.* https://doi.org/10.1111/cdev.13032 (2018).
6. Santillan, J. & Khurana, A. Developmental associations between bilingual experience and inhibitory control trajectories in head start children. *Dev. Sci.* **21**, e12624 (2018).
7. Luk, G., De, Sa,E. & Bialystok, E. Is there a relation between onset age of bilingualism and enhancement of cognitive control. *Biling. Lang. Cogn.* **14**, 588–595 (2011).
8. Paap, K. R., Johnson, H. A. & Sawi, O. Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex* **69**, 265–278 (2015).
9. Lehtonen, M. et al. Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychol. Bull.* **144**, 394–425 (2018).
10. Brito, N. H., Noble, K. G. & Pediatric Imaging, Neurocognition, and Genetics Study. The independent and interacting effects of socioeconomic status and dual-language use on brain structure and cognition. *Dev. Sci.* **21**, e12688 (2018).
11. Morton, J. B. & Harper, S. N. What did Simon say? Revisiting the bilingual advantage. *Dev. Sci.* **10**, 719–726 (2007).
12. Anton, E. et al. Is there a bilingual advantage in the ANT task? Evidence from children. *Front. Psychol.* **5**, 398 (2014).
13. Von Bastian, C. C., Souza, A. S. & Gade, M. No evidence for bilingual cognitive advantages: a test of four hypotheses. *J. Exp. Psychol. Gen.* **145**, 246–258 (2016).
14. Gathercole, V. C. et al. Does language dominance affect cognitive performance in bilinguals? Lifespan evidence from preschoolers through older adults on card sorting, Simon, and metalinguistic tasks. *Front. Psychol.* **5**, 11 (2014).
15. Paap, K. R. & Greenberg, Z. I. There is no coherent evidence for a bilingual advantage in executive processing. *Cogn. Psychol.* **66**, 232–258 (2013).
16. Hoff, E. et al. Dual language exposure and early bilingual development. *J. Child Lang.* **39**, 1–27 (2012).
17. Hoff, E. & Core, C. What clinicians need to know about bilingual development. *Semin. Speech Lang.* **36**, 89–99 (2015).
18. Callahan, R. M. & Gåndara, P. C. *The Bilingual Advantage: Language, Literacy and the US Labor Market* (Short Run Press, 2014).
19. Bialystok, E. *Bilingualism in Development: Language, Literacy, and Cognition* (Cambridge Univ. Press, 2001).
20. Barac, R., Bialystok, E., Castro, D. C. & Sanchez, M. The cognitive development of young dual language learners: a critical review. *Early Child Res. Q.* **29**, 699–714 (2014).
21. Crago, M. & Dussias, G. Introduction. *Appl. Psycholinguist.* **35**, 855 (2014).
22. Issue, S. Bilingualism forum. *Cortex* **73**, 330–377 (2015).
23. Garavan, H. et al. Recruiting the ABCD sample: design considerations and procedures. *Dev. Cogn. Neurosci.* **32**, 16–22 (2018).
24. Martin-Rhee, M. M. & Bialystok, E. The development of two types of inhibitory control in monolingual and bilingual children. *Biling. Lang. Cogn.* **11**, 81–93 (2008).
25. Kapa, L. L. & Colombo, J. Attentional control in early and later bilingual children. *Cogn. Dev.* **28**, 233–246 (2013).
26. Bialystok, E. & Viswanathan, M. Components of executive control with advantages for bilingual children in two cultures. *Cognition* **112**, 494–500 (2009).
27. Gershon, R. C. et al. IV. NIH Toolbox Cognition Battery (CB): measuring language (vocabulary comprehension and reading decoding). *Monogr. Soc. Res. Child Dev.* **78**, 49–69 (2013).
28. Zelazo, P. D. et al. The development of executive function in early childhood. *Monogr. Soc. Res. Child Dev.* **68**, vii–137 (2003).
29. Casey, B. J. et al. The Adolescent Brain Cognitive Development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
30. Hilchey, M. D. & Klein, R. M. Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychon. Bull. Rev.* **18**, 625–658 (2011).
31. Green, D. W. Mental control of the bilingual lexico-semantic system. *Biling. Lang. Cogn.* **1**, 67–81 (1998).
32. Costa, A., Hernandez, M., Costa-Faidella, J. & Sebastian-Galles, N. On the bilingual advantage in conflict processing: now you see it, now you don't. *Cognition* **113**, 135–149 (2009).
33. Altman, D. G. & Bland, J. M. Absence of evidence is not evidence of absence. *Br. Med. J.* **311**, 485 (1995).
34. Ialongo, C. The logic of equivalence testing and its use in laboratory medicine. *Biochem. Med. (Zagreb)* **27**, 5–13 (2017).
35. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Erlbaum, 1988).
36. Zelazo, P. D. et al. NIH Toolbox Cognition Battery (CB): validation of executive function measures in adults. *J. Int. Neuropsychol. Soc.* **20**, 620–629 (2014).
37. De Bruin, A., Treccani, B. & Della Sala, S. Cognitive advantage in bilingualism: an example of publication bias? *Psychol. Sci.* **26**, 99–107 (2015).
38. Simonsohn, U., Nelson, L. D. & Simmons, J. P. *P*-curve: a key to the file-drawer. *J. Exp. Psychol. Gen.* **143**, 534–547 (2014).
39. Bland, M. Do baseline P-values follow a uniform distribution in randomised trials? *PLoS One* **8**, e76010 (2013).
40. Besag, J. & Clifford, P. Sequential Monte Carlo *p*-values. *Biometrika* **78**, 301–304 (1991).
41. Akshoomoff, N. et al. The NIH Toolbox Cognition Battery: results from a large normative developmental sample (PING). *Neuropsychology* **28**, 1–10 (2014).
42. Jernigan, T. L. & Brown, S. A. Introduction. *Dev. Cogn. Neurosci.* **32**, 1–3 (2018).
43. Compton, W. M., Dowling, G., & Garavan, H. Ensuring the best use of data: the Adolescent Brain Cognitive Development Study. *JAMA Pediatrics* (in the press).
44. Barch, D. M. et al. Demographic, physical and mental health assessments in the Adolescent Brain and Cognitive Development study: rationale and description. *Dev. Cogn. Neurosci.* **32**, 55–66 (2018).
45. Place, S. & Hoff, E. Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child Dev.* **82**, 1834–1849 (2011).
46. Blumenfeld, H. K. & Marian, V. Cognitive control in bilinguals: advantages in stimulus–stimulus inhibition. *Biling. (Camb. Engl.)* **17**, 610–629 (2014).
47. Bialystok, E., Craik, F. I., Klein, R. & Viswanathan, M. Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychol. Aging* **19**, 290–303 (2004).
48. Ansaldo, A. I., Ghazi-Saidi, L. & Adrover-Roig, D. Interference control in elderly bilinguals: appearances can be misleading. *J. Clin. Exp. Neuropsychol.* **37**, 455–470 (2015).
49. Bialystok, E. & Barac, R. Emerging bilingualism: dissociating advantages for metalinguistic awareness and executive control. *Cognition* **122**, 67–73 (2012).

50. Zucker, R. A. et al. Assessment of culture and environment in the Adolescent Brain and Cognitive Development study: rationale, description of measures, and early data. *Dev. Cogn. Neurosci.* **32**, 107–120 (2018).
51. Weintraub, S. et al. I. NIH Toolbox Cognition Battery (CB): introduction and pediatric data. *Monogr. Soc. Res. Child Dev.* **78**, 1–15 (2013).
52. Zelazo, P. D. et al. II. NIH Toolbox Cognition Battery (CB): measuring executive function and attention. *Monogr. Soc. Res. Child Dev.* **78**, 16–33 (2013).
53. Logan, G. D. *On the Ability to Inhibit Thought and Action: a Users' Guide to the Stop Signal Paradigm* (Academic Press, 1994).
54. Akshoomoff, N. et al. VIII. NIH Toolbox Cognition Battery (CB): composite scores of crystallized, fluid, and overall cognition. *Monogr. Soc. Res. Child Dev.* **78**, 119–132 (2013).
55. Von Hippel, P. T. Regression with missing *Y*s: an improved strategy for analyzing multiply imputed data. *Sociol. Methodol.* **37**, 83–117 (2007).
56. Wood, S. *Generlized Additive Models: An Introduction with R* (Chapman and Hall/CRC, 2006).
57. Mezzacappa, E. Alerting, orienting, and executive attention: developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child Dev.* **75**, 1373–1386 (2004).
58. Noble, K. G., Norman, M. F. & Farah, M. J. Neurocognitive correlates of socioeconomic status in kindergarten children. *Dev. Sci.* **8**, 74–87 (2005).
59. Carlson, S. M. & Moses, L. J. Individual differences in inhibitory control and children's theory of mind. *Child Dev.* **72**, 1032–1053 (2001).
60. Hughes, C. Finding your marbles: does preschoolers' strategic behavior predict later understanding of mind? *Dev. Psychol.* **34**, 1326–1339 (1998).
61. Engel de Abreu, P. M., Cruz-Santos, A., Tourinho, C. J., Martin, R. & Bialystok, E. Bilingualism enriches the poor: enhanced cognitive control in low-income minority children. *Psychol. Sci.* **23**, 1364–1371 (2012).
62. Friedman, N. P. et al. Not all executive functions are related to intelligence. *Psychol. Sci.* **17**, 172–179 (2006).
63. Smith, B. L., Smith, T. D., Taylor, L. & Hobby, M. Relationship between intelligence and vocabulary. *Percept. Mot. Skills* **100**, 101–108 (2005).
64. Deyo, R. A., Diehl, A. K., Hazuda, H. & Stern, M. P. A simple language-based acculturation scale for Mexican Americans: validation and application to health care research. *Am. J. Public Health* **75**, 51–55 (1985).
65. Anderson, S. & Hauck, W. W. A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Commun. Stat. Theory Methods* **12**, 2663–2692 (1983).
66. Counsell, A. & Cribbie, R. A. Equivalence tests for comparing correlation and regression coefficients. *Br. J. Math. Stat. Psychol.* **68**, 292–309 (2015).
67. Kelley, K. & Preacher, K. J. On effect size. *Psychol. Methods* **17**, 137–152 (2012).
68. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
69. Paap, K. R., Johnson, H. A. & Sawi, O. Are bilingual advantages dependent upon specific tasks or specific bilingual experiences. *J. Cogn. Psychol.* **26**, 615–639 (2014).
70. R Core Development Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2018).

Acknowledgements

We thank the families and children who participated, and continue to participate, in the ABCD study, as well as staff at the study sites, Data Analysis and Informatics Core (DAIC), and site personnel involved in data collection and curating the data release. We also thank A. Counsell for discussion on the equivalence testing approach and for sharing R code. This study was supported by an NIH/NIDA U01DA041156 ABCD study grant. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.S.D. originally conceived the study, analysed the data and wrote the draft manuscript. A.S.D. and W.K.T. designed the analysis. N.L.G., S.M.P., S.W.H., M.T.S., M.C.R., A.R.L. and R.G. contributed to the conception, discussion, data collection, curation of the data and write-up of the study. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0609-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.S.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Anthony Steven Dick

Last updated by author(s): 7/16/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The NIH toolbox measures, administered on an iPAD, were used.
Data analysis	All software used in the present analysis is open source from the Comprehensive R Archive Network (CRAN; v. 3.5.0;). The R code to replicate the analysis is available at https://github.com/anthonystevendick/bilingual_abcd.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data are from the ABCD Study Curated Annual Release 1.0 and are available by request from the NIMH Data Archive (https://data-archive.nimh.nih.gov/abcd).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We examined data from the Adolescent Brain and Cognitive Development study and measured their performance on intelligence, vocabulary, and executive function tests, and their degree of bilingualism. It is a quantitative quasi-experimental design study.
Research sample	The sample is comprised of 4524 9-10-year-old children collected from 21 study sites across the United States.
Sampling strategy	Demographically, the ABCD Study used a multi-stage probability sample of eligible children by probability sampling of schools within the catchment area of each site. The goal of this sampling strategy was to approximate the demographic profile of two national surveys, the American Community Survey (ACS; a large-scale survey of approximately 3.5 million households conducted annually by the U.S. Census Bureau) and annual 3rd and 4th grade school enrollment data maintained by the National Center for Education Statistics. The sampling strategy was additionally constrained by the requirement that study sites had available MRI scanners. Because these are typically available at research universities in urban areas, the sampling tends to oversample urban as opposed to rural students and families. Despite this caveat, the ABCD Study sample was largely successful at matching the ACS survey demographic profiles. Demographic assessments of the sample are summarized here in Barch et al. The demographic profile of the sample is presented in Table 1 in the paper.
Data collection	Researchers were blind to the study hypotheses of the present paper. During data collection, the research assistant, the parent, and the children in the study were present. For the present study, the NIH toolbox, administered on an iPAD, and various demographic questionnaires, administered on a computer, were analyzed.
Timing	Data collection began in the Fall of 2016 and continued until December of 2017.
Data exclusions	Missing data analysis is described in the paper and was applied to demographic variables that were missing. All data that were available were analyzed.
Non-participation	Some participants declined participation, but this depended on the task. For one task, the stop-signal task, because it was administered in a MRI scanner, a number of participants declined to continue in data collection. That task had 25% attrition. Available data for all of the other tasks approached or matched the original targeted sample (n = 4524).
Randomization	Random assignment was not used. Demographic and cognitive covariates were entered into the generalized linear model analyses. These covariates were age, sex, race/ethnicity, highest degree of education, household income, marital status, IQ, and English vocabulary. We also modeled family within study site as random factors.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	The sample is comprised of 4524 9-10-year-old children collected from 21 study sites across the United States. Demographically, the ABCD Study used a multi-stage probability sample of eligible children by probability sampling of schools within the catchment area of each site. The goal of this sampling strategy was to match the demographic profile of two national surveys, the American Community Survey (ACS; a large-scale survey of approximately 3.5 million households conducted annually by the U.S. Census Bureau) and annual 3rd and 4th grade school enrollment data maintained by the National Center for Education Statistics. The sampling strategy was additionally constrained by the requirement that study sites had available MRI scanners. Because these are typically available at

	research universities in urban areas, the sampling tends to oversample urban as opposed to rural students and families. Despite this caveat, the ABCD Study sample was largely successful at approximating the ACS survey demographic profiles. That said, although it closely approximates the demographic profile of the ACS survey, because the sampling strategy heavily relied on schools in urban areas, it is more accurate to describe the sample as having a population-based, demographically diverse sample that is not necessarily representative of the U.S. national population.
Ethics oversight	University of California at San Diego Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.