Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Discussion forum

Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances

Kenneth R. Paap [a,*], *Hunter A. Johnson* [a] *and Oliver Sawi* [b]

[a] *Department of Psychology, San Francisco State University, San Francisco, CA, USA*
[b] *Department of Psychology, University of Connecticut, Storrs, CT, USA*

ARTICLE INFO

Article history:
Received 1 March 2015
Reviewed 1 April 2015
Revised 19 April 2015
Accepted 22 April 2015
Published online 7 May 2015

ABSTRACT

The hypothesis that managing two languages enhances general executive functioning is examined. More than 80% of the tests for bilingual advantages conducted after 2011 yield null results and those resulting in significant bilingual advantages tend to have small sample sizes. Some published studies reporting significant bilingual advantages arguably produce no group differences if more appropriate tests of the critical interaction or more appropriate baselines are used. Some positive findings are likely to have been caused by failures to match on demographic factors and others have yielded significant differences only with a questionable use of the analysis-of-covariance to “control” for these factors. Although direct replications are under-utilized, when they are, the results of seminal studies cannot be reproduced. Furthermore, most studies testing for bilingual advantages use measures and tasks that do not have demonstrated convergent validity and any significant differences in performance may reflect task-specific mechanism and not domain-free executive functions (EF) abilities. Brain imaging studies have made only a modest contribution to evaluating the bilingual-advantage hypothesis, principally because the neural differences do not align with the behavioral differences and also because the neural measures are often ambiguous with respect to whether greater magnitudes should cause increases or decreases in performance. The cumulative effect of confirmation biases and common research practices has either created a belief in a phenomenon that does not exist or has inflated the frequency and effect size of a genuine phenomenon that is likely to emerge only infrequently and in restricted and undetermined circumstances.

[*] *Corresponding author*. Department of Psychology, Language Attention and Cognitive Engineering Laboratory, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, USA.

E-mail addresses: kenp@sfsu.edu (K.R. Paap), haj@mail.sfsu.edu (H.A. Johnson), oliver.sawi@uconn.edu (O. Sawi).

http://dx.doi.org/10.1016/j.cortex.2015.04.014
0010-9452/© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Parents, educators, cognitive scientists, and bilinguals themselves have taken a keen interest in the consequences of bilingualism for language skills, cognitive abilities, and general quality of life. All things considered, we wish to make clear at the onset that we believe that the advantages of bilingualism across a host of personal, economic, social, and cultural dimensions overwhelmingly preponderate any disadvantages. This article examines a much narrower question: Does bilingualism enhance executive functioning as reflected in performance advantages in nonverbal tasks? Executive functions (EF) consist of a set of general-purpose control processes that are central to the self-regulation of thoughts and behaviors and that are instrumental to accomplishing goals. For purposes of exposition and organization the theoretical framework developed by Miyake and Friedman (2012) is adopted. Miyake and Friedman reported evidence for three components of EF: updating, shifting (or switching), and inhibiting with the caveat that inhibition may not be separable from updating and switching.

There is a widely held view that bilinguals enjoy an advantage over monolinguals in EF. Bialystok (2011) stated that "Studies have shown that bilingual individuals *consistently* [emphasis added] outperform their monolingual counterparts on tasks involving executive control" p. 229. In a follow-up review it was reported that "… bilinguals *at all ages* [emphasis added] demonstrate better executive control than monolinguals matched in age and other background factors" (Bialystok, Craik, & Luk, 2012, p. 212). Similarly, Kroll and Bialystok (2013) observed that "… studies of executive function demonstrate a bilingual advantage, with bilinguals outperforming their monolingual counterparts on tasks that required ignoring irrelevant information, task switching, and *resolving conflict* [emphasis added]" (p. 2). Mercier, Pivneva, and Titone (2014) state that "… bilinguals are advantaged relative to monolinguals in non-linguistic cognitive control… over the life-span during normal aging… and pathological aging" p. 90.

In contrast, based on the evidence discussed in this target article we conclude that either bilingualism does not enhance EF in any circumstance or only in very specific, but undetermined, circumstances. Some readers may find it surprising that after more than a decade of intense study the question of whether bilingualism enhances general EF is still controversial. One might hope that this forum will contribute to the attenuation of that controversy by finding more common ground with regard to a host of methodological problems that plague this research topic (and many others in psychological science).

2. The published database is biased

Biases in decision making on the part of researchers, reviewers, and editors lead to a published database that is not representative of all studies. Rosenthal (1979) coined the phrase "file drawer problem" to describe the strong tendency of researchers to set aside experiments with null results rather than submit them for publication. When researchers do resist the temptation to place their null results in a file drawer they do so with the understanding that publishing null results, particularly ones that counter earlier published findings, will be difficult. Reviewers and editors are well trained to respond favorably to results that are significant, novel, counterintuitive, and newsworthy, but not if the novelty takes the form of a failure to replicate an "established" finding. In these cases, reviewers and editors may remind themselves that null results could be Type 2 errors or the product of poor methodology. Mahoney (1977) showed that experienced reviewers for a psychology journal (who believed they were providing real reviews) were biased in favor of positive results over mixed, negative, or null results.

The field of bilingualism is not immune to these biases. de Bruin, Treccani, and Della Sala (2015) provided evidence that the combined effects of researchers deciding what to submit and editors deciding which articles to publish were leading to a bias favoring studies with bilingual advantages over those reporting null and negative results. The primary evidence stemmed from examining the fate of 104 conference abstracts presented at 52 different national and international conferences. Fifty-two were eventually published in a scientific journal. Studies with results fully supporting the hypothesis that there are bilingual advantages in EF were most likely to be published (68%), followed by studies with mixed results, and those clearly challenging the hypothesis were published the least (29%).

De Bruin et al. also report the results of a meta-analysis on the set of the published articles. The average weighted difference was $d = +.30$ and following Cohen's (1992) guidelines this is a small effect. However, any biases against null or negative effects will have inflated the true effect size and it is clearly smaller by an unknown amount. The funnel plot reported in de Bruin et al. shows several extreme scores with low precision and they are all positive effects. The asymmetry is very much expected for this meta-analysis because we already know that there were many abstracts with null or negative findings that were never published. De Bruin et al. note that the amount of bias favoring bilingual advantages in the total set of 104 abstracts is only the tip of the iceberg as it is reasonable to assume that additional researchers with null and negative results decided not to submit them for presentation at a major conference. If those triaged to the file drawer before submission to a conference could be added to the unpublished conference abstracts, they could cancel out the small effect size found in the meta-analysis of those conference abstracts that were eventually published. In the next section we present the case that the true effect size may, indeed, be zero.

3. The posited case that bilingual advantages in EF do not exist

Several lines of evidence converge on the strong possibility that managing two languages does not enhance general EF despite the many published results showing that bilinguals significantly outperform monolinguals on tasks assumed to measure EF. To logically challenge the thesis of bilingual advantages in EF one must propose that either the reported differences in performance are due to other causes or that the dependent variables are not measuring EF. Across individual studies we suggest that one, the other, or both are true. Other causes include Type 1 errors, confounds in demographic factors, and questionable statistical tests. Symptoms that these alternative causes of bilingual advantages in performance are prevalent include the inability to replicate seminal findings and the absence of significant advantages in studies using large sample sizes.

3.1. *What is the rate of false positive results in psychological science?*

Questionable research practices (QRPs) can substantially increase the rate of Type 1 errors. Bakker, van Dijk, and Wicherts (2012) speculate that publication biases favoring positive results motivate researchers to conduct multiple studies with small sample sizes rather than a single study with a large sample size. If a researcher adopts this strategy, employs directional tests at alpha = .05 for each test, and stops and reports the first study that “works” (i.e., that yields a $p < .05$ in the predicted direction), then the combined probability of a Type 1 error[1] increases with the researcher's willingness to conduct two (.097), three (.142), four (.185), or five (.226) small experiments. Although one might conjecture that the incidence of putting this strategy into practice is low, John, Loewenstein, and Prelec (2012) surveyed over 2,000 research psychologists and 48% admitted to having submitted papers that only reported the studies that “worked”. Other prevalent practices from the John et al. survey include: (1) failing to report all of a study's dependent measures, 65% admission rate; (2) deciding to collect more data when the results are not significant, 57%; (3) rerunning analyses with outliers removed, 41%; (4) failing to report all of a study's conditions, 28%; and (5) “rounding off” a p value greater than $p = .050$ to a value $\leq .05$, 22%. In their simulation demonstration Bakker et al. operationalized the first three QRPs from the list above and applied them sequentially to any simulated study that initially produced a non-significant result. Instantiating these QRPs caused the rate of false positives to jump to nearly .40 (even with a more conservative two-tailed alpha). In summary, the Bakker et al. simulations show that a strategy of conducting small-sample experiments, applying QRPs to the non-significant results, and reporting only those experiments that “work” leads to substantial rates of false positive when the true ES is zero. Although it is not possible to pinpoint the rate of false-positive bilingual advantages it seems fair to conjecture that it is nontrivially greater than .05.

3.2. *How often do significant bilingual advantage occur?*

Is the proportion of significant bilingual advantages sufficiently small to suggest that a lion's share are false positives caused by chance and augmented by biases and QRPs? Paap, Johnson, and Sawi (2014) examined the tests for bilingual advantages in nonverbal switching tasks and interference tasks that were not included in Hilchey and Klein’s (2011) review. The analysis did not consider tests on preschool children as these generally use different tasks and tend to rely on accuracy rather than latency as the primary dependent variable. We have updated the Paap, Johnson, and Sawi (2014) analysis to include reports appearing since 2014. The selection criteria, scoring rubric, and a list of the specific studies and comparisons appear in the supplementary materials. In nonverbal interference tasks the interference score (incongruent RT minus congruent RT) is usually assumed to measure inhibitory control and 13 of the 64 tests (.203) reported significant bilingual advantages. Global RT (mean of both congruent and incongruent RTs) is often taken as a measure of monitoring ability and yielded 6 bilingual advantages out of 46 tests (.130). Switching costs (switch RT minus repeat RT) are usually measured as the difference between switch trials and repeat trials in a block where task is randomly varied and cued on each trial. Significant bilingual advantages in switch costs were reported in 4 of 32 tests (.125). Many switching tasks also include pure blocks where only a single task is performed and this enables one to compute mixing costs (repeat RT from the mixed block minus single-task RT). Mixing costs are usually assumed to measure the cost of monitoring and maintaining a task set when a switch is not required. Significant bilingual advantages in mixing costs were reported in 5 of 23 tests (.217). When all four measures are combined, .170 of all tests for bilingual advantages yielded significant advantages.

3.3. *Confounds can produce significant bilingual advantages*

Another inherent difficulty in testing for bilingual advantages in EF is that factors ranging from genetics (Friedman, et al., 2008) to acquiring specialized skills (see Valian, 2014 for a review) influence the development and maintenance of EF. Because random assignment is not possible in quasi-experimental studies comparing monolinguals to bilinguals there is always the unfortunately all too likely possibility that bilingualism covaries with other factors that affect EF. Hilchey, Saint-Aubin, & Klein (in press) speculated that confounding variables may have caused many of the reported bilingual advantages and urged the field to be more systematic and quantitative in matching demographic variables.

3.3.1. *Socioeconomic status (SES)*

In reviewing the research with children Hilchey and Klein (2011) concluded that there was no evidence for bilingual advantages in inhibitory control obtained with the Simon task: across six different comparisons there were no significance differences in the magnitude of the interference effect, but five of the six did show significant differences in global RT. The exception was the study by Morton and Harper (2007) that in terms of task, methods, and design was a direct replication of Bialystok, Craik, Klein, and Viswanathan (2004). The principle difference between the studies was that Morton and Harper used measures of parent's educational level and family income to ensure that the bilinguals and monolinguals were matched on SES. They were also matched in terms of ethnicity and immigrant status.

One of the most rigorous efforts to control for SES was that conducted by Engel de Abreu, Cruz-Santos, Tourinho, Martin,

[1] The combined probability of Type 1 error across k experiments is: $p = 1 - (1 - .95)^k$.

and Bialystok (2012) in a study that produced bilingual advantages in both monitoring and inhibiting. However, the bilingual children in Engel de Abreu et al.'s investigation resided in Luxembourg whereas the monolingual children resided in Portugal and, as Hilchey, et al. put it: this difference allows for the possibility that "unaccounted-for social experiences unique to each country" were responsible for the unusual finding of a bilingual advantage on inhibitory control and global RT. Paap and Liu (2014) made much the same point and also noted that the authors themselves characterize their results as remarkable because the bilingual children have strikingly low vocabulary scores in Luxembourgish and thus were not at all proficient in their L2. The bilingual advantages in inhibitory control and monitoring obtained by Engle de Abreu et al. were not replicated in two very large studies using Basque–Spanish bilinguals that included children the same age and older (Anton et al., 2014; Duñabeitia et al., 2014). The participants were all native residents of Spain and the groups were carefully matched on SES. These studies are noteworthy because the bilinguals acquired both languages early, were highly proficient, and were immersed in a bilingual region. Thus, in all respects the Basque–Spanish children were better candidates to show bilingual advantages in comparison to the Portuguese–Luxembourgish bilinguals, yet they did not. In this context, it seems more likely that a factor other than bilingualism caused the "bilingual advantages" in the Engel de Abreu et al. study.

3.3.2. *Immigrant status in EF and dementia onset*

Many reports of bilingual advantages have confounded immigrant status with bilingualism, especially in research testing older adults. In a study of older adults Bialystok, Craik, and Luk (2008) 20 of the 24 bilinguals were immigrants. In a direct replication of Bialystok et al. (2004); Kirk, Fiala, Scott-Brown, and Kempe (2014) found no language-group differences in either the magnitude of the Simon interference scores or in global RT in five groups of older adults (mean age = 70.8 years). In the research using older adults language-group differences often occur when immigrant status is not matched (Bialystok et al., 2008; Gold, Kim, Johnson, Kryscio, & Smith, 2013; Salvatierra & Rosselli, 2011; Schroeder & Marian, 2012 [2]) and do not occur when it is (Billig & Scholl, 2011; Kirk et al., 2014; Kousaie & Phillips, 2012).

The importance of controlling for immigrant status is underscored by investigations of the role of bilingualism on the onset of mild cognitive decline or dementia. The first studies used retrospective reports of patients at memory clinics and showed that bilingualism delayed the onset of symptoms or diagnosis by several years. Some of these studies confounded bilingualism with immigrant status (Bialystok, Craik, & Freedman, 2007) while another found bilingual benefits within immigrant samples, but not between native samples (Chertkow et al., 2010). Immigrant status is important because it is associated with higher intelligence that, in turn, is associated with delays in dementia onset (Fuller-Thomson, & Kuh 2014). Furthermore, the four studies that have used a prospective cohort design following individuals without dementia at baseline have all found no significant effects of bilingualism and the trend in three of those favors the monolinguals (Crane et al., 2009; Lawton, Gasquoine, & Weimer, 2015; Sanders, Hall, Katz, & Lipton, 2012). Fuller-Thomson (2015) suggest that the longitudinal design is less open to biases in sampling, measurement, and publication. Weighing the prospective studies more heavily, there is little evidence that bilingualism protects against cognitive decline.

3.3.3. *Cultural differences*

When immigrant status is confounded it is also likely that there are differences in culture across the language groups. Even when studies match the groups with respect to measured SES and immigrant status there may be cultural differences. Morton and Carlson (in press) present evidence for cultural differences in the development of EF and show how these differences can confound tests of bilingual advantages. A study by Carlson and Choi (2009) provides a dramatic demonstration of the entanglement between culture and bilingualism. Using six different measures of EP they found significant bilingual advantages comparing a group of Korean–English bilinguals living in the United States to a "matched" sample of American monolinguals. However, the performance of the Korean–American bilinguals was indistinguishable from a third group of matched Korean monolinguals. This clearly questions an interpretation that the obtained group differences were due to bilingualism and strongly supports the view that cultural differences play an influential role in the development of EP.

Positive associations between measures of EF and degree of bilingualism face the usual problem with correlation studies in that degree of bilingualism may enhance general abilities in EF, but superior EF may make it easier to achieve higher levels of L2 proficiency, or some third factor such as cultural differences in parenting practices may be driving both measures. A study by Chen, Zhou, Uchikoshi, and Bunge (2014) provides an excellent example of these complexities. A large ($n = 223$) sample of Chinese American immigrant children between ages 7 and 10 of varying levels of Chinese and English proficiencies were tested on laboratory tasks of EF and measures of self regulation at home and school. Some measures of control were predicted only by proficiency in Chinese (e.g. higher behavioral persistence, fewer commission errors on the response inhibition task) and others only by proficiency in English (e.g., fewer omission errors in the response inhibition task). As Chen et al. speculate higher fluency in Chinese may reflect "greater adherence to traditional Chinese values of behavioral control and self-restraint…" p. 9. Intriguingly performance on a laboratory measure of cognitive flexibility involving both incongruent trials and task switching was highest for those children highly proficient in both Chinese and English. Although Chen et al. conservatively conclude that their "study provides only limited evidence" for a bilingual advantage in EF, the result with the cognitive flexibility measure merits further investigation.

[2] Gold et al. interpreted their behavioral results with older participants as a bilingual advantage in global switch costs, but the differences were not significant at a conventional alpha of $p < .05$.

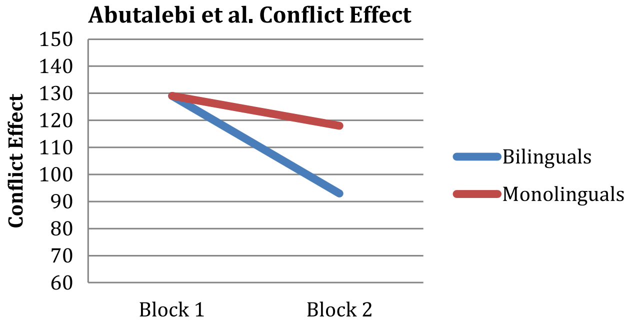

Fig. 1 – The magnitude of the flanker interference effect for bilinguals and monolinguals across the two blocks of the Abutalebi et al. (2012) study.

3.4. *"Bilingual advantages" are sometimes supported by questionable statistics*

3.4.1. *Two-way interactions should test if one difference significantly differs from another*

Interactions have played a key role in evaluating the presence of bilingual advantages, but the interactions must be statistically real and have the right pattern. The study by Abutalebi, et al. (2012) using a flanker task provides an example. As shown in Fig. 1 in a first block of trials monolinguals and bilinguals both show similar interference effects of about 130 msec. However, in Block 2 the bilinguals reduced their flanker effect by 36 msec (CI = 21–53) compared to only 11 msec (CI = −12 to 41) for the monolinguals. The 36 msec reduction was significant, but the 11 msec reduction was not. This statistical evidence led Abutalebi et al. to conclude that bilinguals "… are better able to adjust to conflict, hence, to adapt to conflicting situations" (p. 2085). However, a more direct test of the hypothesis that bilinguals are "better able to adjust to conflict" would directly test if the 36 msec improvement for the bilinguals was significantly greater than the 11 msec improvement for the monolinguals. These mean reductions in the conflict effect are shown in Fig. 2 together with the reported 95% confidence intervals. The high degree of overlap between the two CIs is caused by the lack of precision in the parameter estimates (especially about the mean for the monolinguals) and this suggests that these means would not differ in a *t*-test.

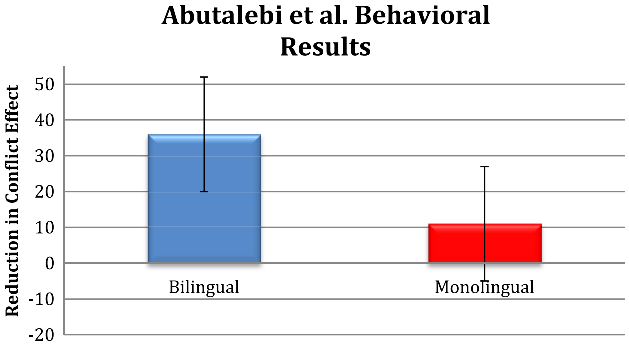

Fig. 2 – Mean reduction in the flanker interference scores with 95% confidence intervals for the Abutalebi et al. (2012) experiment.

Indeed, given the specification of the number of participants (17 bilinguals, 14 monolinguals) and the two CI's, one can recover the standard deviations of each group, compute the pooled-variance estimate, and then use the difference in means and the pooled-variance estimate to compute the independent groups *t*-value. It is $t(29) = 1.48$, $p = .15$. By the conventions of null hypothesis statistical testing the two groups do not statistically differ in the degree of improvement from Block 1 to Block 2.

The group sample sizes and *t* value can also be used to compute a Bayes Factor that shows the ratio of the probability of the data given the null hypothesis to the probability of the data given the alternative. Using Rouder's web site (Rouder, Speckman, Sun, Morey, & Iverson, 2009) the BF is equal to 1.52 indicating that given the actual data obtained by Abutalebi et al. the null is 1.5 times more likely to be true than the alternative. The quality of the statistical evidence supporting a bilingual advantage in the behavioral data is very weak at best. Nonetheless, the interaction was interpreted as showing that bilinguals "… are better able to adjust to conflict, hence, to adapt to conflicting situations" (Abutalebi et al., 2012, p. 2085). However, as described in Section 3.5 studies with much larger samples sizes either show no advantages in any block or the reverse pattern, that is, an advantage in the first block.

3.4.2. *The pattern of interaction matters*

Many tests for bilingual advantages in EF employ mixed-factorial designs with two (or more) language groups and two types of trials. In the Simon, flanker, and Stroop tasks interference scores are the difference between the congruent and incongruent trials. If bilinguals have better inhibitory control, then the interference scores should be smaller for bilinguals compared to monolinguals. Thus, the specific pattern of a significant interaction matters. As Hilchey et al. (in press) point out a bilingual advantage in inhibitory control is not consistent with an interaction that is caused by a monolingual advantage on the conflict-free congruent trials and comparable performance on the incongruent (conflict) trials. They note that this is precisely the pattern of interaction obtained in three studies of older adults. In each case the smaller Simon interference effects for the bilinguals were interpreted as "bilingual advantages" in inhibitory control (Bialystok et al., 2008; Salvatierra & Rosselli, 2011; Schroeder & Marian, 2012). To illustrate, the mean RTs for the congruent and incongruent trials of the Schroeder and Marian study are shown in Fig. 3. The Language Group × Trial Type interaction is significant, the Simon interference effect is smaller for bilinguals, but this difference is caused by monolinguals responding faster on conflict-free trials– not by bilinguals responding faster on the conflict trials. In fact, the monolinguals are a bit faster on the incongruent trials as well. Thus, the results cannot feasibly support the hypothesis that bilinguals have superior inhibitory control.

Based on their analysis of these studies and all other published articles testing elderly participants with nonverbal interference tasks Hilchey et al. concluded that there was no "compelling" support for the notion of a bilingual executive-processing advantage for the simple reason that across the studies in this age group the monolinguals either outperformed the bilinguals or there was no difference at all.

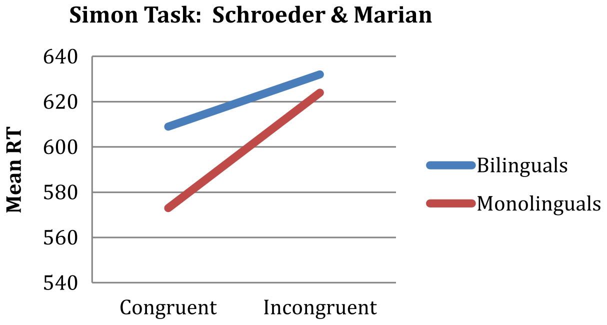

Fig. 3 – Mean RT on congruent and incongruent trials for bilinguals and monolinguals in Schroeder and Marian's Simon task.

3.4.3. *Missing interactions*

Calvo and Bialystok (2014) explored the joint effects of bilingualism and SES on several tasks including a flanker task and concluded that the effects of bilingualism and SES were independent. Relevant to present purposes we will re-examine the strength of their evidence for bilingual advantages in EF. In all three tasks chevrons were used as targets. In the control task, a single chevron was shown in the center of the screen. In the flanker task a centered target was flanked by chevrons pointing in the same (congruent) or opposite (incongruent) direction. There was also a visual-search task that required a response to a red chevron in a row that included four black diamonds. Response times and accuracy were analysed in separate three-way ANOVAs that included the three tasks, language, and SES group as factors. The only evidence for a bilingual advantage in EF was a significant main effect showing that bilinguals were more accurate than monolinguals. This main effect collapsed across all three tasks (control, visual search, flanker) and across both the congruent and incongruent trials within the flanker task. Critically, language group did not interact with task in the analysis of accuracy data. In the analysis of RTs the main effect of group was not significant and there were no significant interactions.

The main effect on the accuracy measure is highlighted in the authors' discussion, but there is no mention of other comparisons that, if significant, would provide far more compelling evidence that the language group differences were due to differences in EF. For example, did the two groups differ with respect to the magnitude of the flanker effect? The answer is unknown: neither the means and standard deviations for the congruent and incongruent trials nor a separate analysis of the conflict condition (standard flanker task) were reported.

Perhaps more important is that there is no mention that the absence of a Language Group × Task interaction implies that the bilingual advantages are not significantly greater in the flanker task than in the control task. Furthermore, inspection of the results appearing in Calvo and Bialytok's Table 2 shows that the bilingual advantage in overall accuracy is driven more by the control and search tasks than the flanker task. A conflict effect measured as the difference in accuracy between the control task and the flanker task is the same for both monolinguals and bilinguals. Likewise, *t*-tests (derived from the means and standard deviations provided) that compare bilinguals to monolinguals in just the flanker task show no significant differences for either the working-class groups, $t(62) = 1.36$, $p > .05$ (one-tailed) or for the middle-class groups, $t(109) = .96$, $p > .05$. In summary, there is very little support in the accuracy data and no support in the RT data for bilingual advantages that one could attribute to advantages in conflict monitoring or inhibitory control. Furthermore, the interpreted bilingual advantages in EF in the Calvo and Bialystok study do not cohere with the results of the study by Anton et al. (2014) who tested 180 Spanish–Basque bilinguals and 180 Spanish monolinguals in a child-friendly version of the ANT task and found no significant differences in either the magnitude of the flanker effect or in global performance.

3.4.4. *ANCOVA cannot statistically "control" for failures to match*

As Paap, Johnson, and Sawi (2014) discussed a critical assumption of ANCOVA is that the covariate and groups are independent (Miller & Chapman, 2001). When the independence assumption is violated the regression adjustment may either obscure part of the grouping effect (e.g., language effect) or produce spurious effects; thus the ANCOVA results are uninterpretable when there are systematic differences in the covariate across groups. As is true in almost every area of psychological science, the questionable use of ANCOVA is common and this holds true for the literature on the bilingual advantage (Blom, Küntay, Messer, Verhagen, & Leseman, 2014; Marzecová, Asanowicz, Krivá, & Wodniecka, 2013; Prior & Gollan, 2011; Tao, Marzecova, Taft, Asanowicz, & Wodniecka, 2011). The problem is not using covariates; it is using covariates that are confounded with the monolingual versus bilingual grouping.

3.5. *The distribution of significant results across sample size is consistent with a null effect*

In building the case that bilingual advantages in EF do not exist we observed that the proportion of positive results is relatively small and that the significant differences in performance may have been due to other causes such as Type 1 errors, confounds, and questionable statistics. In this section we examine the distribution of significant results across sample size and conclude that the pattern is far more consistent with the assumption of a null effect than the assumption of a small effect size.

The meta-analysis by de Bruin et al. (2015) of conference abstracts that were eventually published yielded a mean effect size of only .30 and one can safely assume that this overestimates the size of any genuine bilingual advantage. In order to consistently detect a small effect one needs powerful designs based on large sample sizes. Thus, a survey of the research on the bilingual advantage should show that significant bilingual advantages are usually associated with large n studies.

Fig. 4a and b shows the number of null reports (left side in blue) and the number of significant bilingual advantages (right side in green) for different sample sizes (mean number of participants per group) from our updated meta-analysis. The top figure combines tests of inhibition and switching costs.

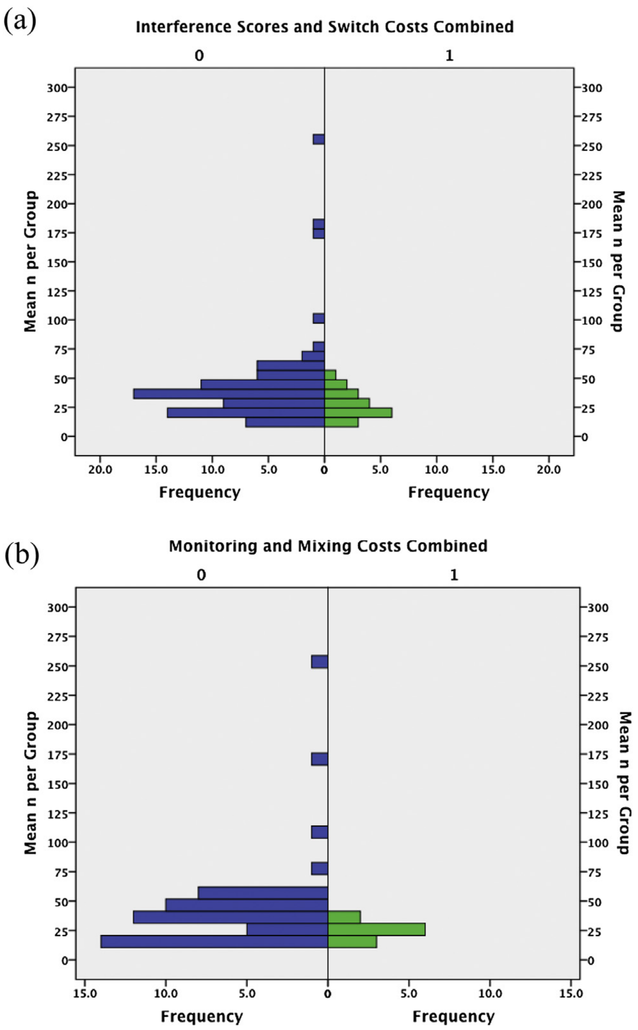

Fig. 4 – Frequency of nonsignificant ($p > .05$, left side) and significant ($p < .05$, right side) bilingual advantages as samples size grows. The top figure combines all 96 tests using either interference scores or switching costs. The bottom figure combines all 69 tests of monitoring in either nonverbal interference tasks or switching tasks.

The bottom figure combines tests of monitoring from nonverbal interference tasks with mixing costs (another assumed measure of monitoring).

Inspection of the left side of the histograms shows that null results occur in small, medium, and large-n studies. The striking effect on the right side shows that significant bilingual advantages did not occur with large-n studies and appear principally when sample size is small ($n < 30$). This is not the expected pattern if bilingualism truly does enhance EF. If the null is false, then as the sample size becomes arbitrarily large—the t value grows without bound, and the p value converges to zero. This is a good property because it implies that the null will always be rejected in the large-sample limit if there is a real difference to detect. As Rouder et al. (2009) put it: "Researchers, therefore, can rest assured that increasing sample size will, on average, result in a gain of evidence against the null when the null is, indeed, false" p. 226. Thus, all others things being equal, one would expect significant effects (especially for a small effect size) to cluster at the higher end of sample sizes. A simple explanation for the pattern observed in Fig. 4 is the combined assumption that the null is true and that there is a bias on the part of researchers and reviewers to prefer positive results to null results.

Had the analysis reported in Fig. 4 included the earlier studies reviewed by Hilchey and Klein a clear exception to the general pattern would be the results of Costa, Hernández, and Sebastián-Gallés (2008) who reported significant bilingual advantages in flanker interference and global RT with 100 participants in each language group. The advantage in the flanker effect appeared in the early blocks but disappeared in the third block leading Hilchey and Klein to suggest that the advantage reflects "... a reconfiguration of cognitive processes, rather than any enduring bilingual advantage on inhibitory control..." p. 654. Furthermore, in "high monitoring" conditions where the proportion of incongruent trials is .5, three additional experiments have failed to produce any bilingual advantages in the flanker effect in any of the three blocks (Costa, Hernández, Costa-Faidella, & Sebastián-Gallés, 2009; Paap & Greenberg, 2013; Paap & Sawi, 2014).

3.6. *Seminal studies do not replicate*

Another symptom consistent with the hypothesis that bilingualism does not enhance EF is that seminal studies do not replicate. Following the lead of Pashler and Harris (2012) it is important to distinguish between direct (or exact) replications and conceptual replications. A direct replication uses the same task, method, procedure and analysis to determine if the original results can be reproduced with a new sample of participants. It is the best tool available in science to correct mistakes. Successful replications of bilingual advantages are unlikely because the typical designs are severely underpowered. If a true ES of $d = .3$ is assumed (the mean from de Bruin's meta-analysis) and typical samples of 24 per group are used, the power associated with a one-tail test at an alpha of .05 is only .27 for a two independent groups t-test. Thus, the probability of obtaining significant advantages in two consecutive studies is only .06.

The replicability of the bilingual advantages in switching cost reported by Prior and MacWhinney (2010) offers a good case study for the value of exact replication. Given the means and standard errors of the switching cost measure reported for each group the estimated effect size was $d = .52$. There were 44 participants in each group and thus the estimated power for a one-tailed test and alpha equal to .05 was .78. By itself this looks like a compelling demonstration that bilinguals have better switching ability.

Prior and Gollan (2011) conducted a direct replication. The color-shape switching task was the same in all respects, but the young adults were recruited from a different university and included three groups: Spanish–English, Mandarin–English, and English monolinguals. The bilingual advantage failed to replicate for the Mandarin–English Bilinguals and did not replicate when the original ANOVA was run on the Spanish–English bilinguals. However, if the dependent measure was changed from absolute RT to a measure of relative speed and if parent's educational level (PED) was included as a covariate, then the difference between the Spanish–English and monolingual groups was significant. Prior and Gollan made clear that a significant bilingual advantage for the Spanish–English bilinguals only occurred when both the dependent variable was transformed and when PED was used as a covariate. Running additional analyses until one finds a significant result is not an uncommon practice, but even when it is fairly reported it can inflate the rate of false positives. Adding to the ambiguity of the Prior and Gollan (2011) results, many experts would consider their use of the PED covariate as inappropriate because there were significant differences in PED across the three groups, that is, the covariate was confounded across the three language groups with the parents of the Spanish–English bilinguals having significantly lower educational levels. See Section 3.4.4 for a discussion of the ANCOVA problem. Perhaps more important, in another replication, Prior and Gollan (2013) did not find any significant bilingual advantages, even for the Spanish–English group, when they used the same color-shape switching task but replaced keyboard responses with spoken responses.

Although Prior and Gollan (2011) concluded that bilingual advantages may be restricted to bilinguals who switch languages frequently, it may be the case that the original advantage in switching costs reported by Prior and MacWhinney is simply an outcome that does not replicate. Paap and Greenberg (Study 1, 2013) reported a direct replication of the Prior and MacWhinney study using undergraduate students and bilinguals who speak a variety of languages in addition to English. Furthermore Paap and Greenberg (Study 2 and Study 3) and Paap and Sawi (2014) conducted three additional exact replications. The four studies averaged 65 participants per language group and there was no hint of a bilingual advantage in switching costs in any of them. Concurrently, Hernández, Martin, Barceló, and Costa (2013) were conducting a series of switching experiments comparing Catalan–Spanish bilinguals to Spanish monolinguals. Study 3 was a direct replication of Prior and MacWhinney's original color-shape switching task, but there were no language-group differences in switching costs. Although the first two studies used different switching tasks and are better considered conceptual replications when Hernández et al. performed an omnibus analysis combining the standardized switching costs there were no bilingual advantages for switching costs in a very powerful test based on 292 participants. Mor, Yitzhaki-Amsalem, and Prior (2014) also reported null results for switching costs in a direct replication that compared 20 Russian–Hebrew bilinguals to 20 Hebrew monolinguals. Finally, Moradzadeh, Blumenthal, and Wiseheart (2014) in an interesting study examining both musical training and bilingualism found that musical training yielded both smaller mixing costs and switching costs in a digit-number switching experiment (and also on an updating measure), but there was no language differences on any of the measures.

The purpose of this section was to show that a single study (viz., Prior and MacWhinney) that appears to yield a moderately large and significant bilingual advantage will not necessarily replicate and that only when other researchers are willing to do direct replications of their own and others work can science be self-correcting. The null results obtained in our lab with various types of bilinguals (like Prior & MacWhinney) and in Barcelona with Catalan–Spanish bilinguals who are highly proficient and frequent switchers (like Prior & Gollan's Spanish–English bilinguals) should cast doubt on the initial conclusion that there is a bilingual advantage in switching costs for young adults.

3.7. *The search for a specific bilingual experience that hones a specific component of EF has yet to succeed*

If EF in general consists of somewhat independent abilities to monitor, update, switch, and inhibit then some aspects of coordinating two languages (e.g., switching between languages) may exercise some components (e.g., switching) more than others. Within this framework many investigators have anticipated that it will only be a matter of time until we discover the specific type of bilingual experiences that lead to enhancements of the corresponding components of EF. As Pashler and Harris (2012) describe in their essay on replication a steady stream of direct replications of a true effect should lead to cumulative knowledge regarding the necessary and sufficient conditions for producing it, but to chaos when the true effect size is nil and researchers enthusiastically pursue conceptual replications. Based on a review of the literature and the analysis of a large composite database Paap, Johnson, and Sawi (2014) concluded that there is no compelling evidence that either the L2/L1 proficiency ratio or the number of languages spoken predicts performance on any component of EF. Similarly, Paap, Darrow, Dalibar, & Johnson (2015) closely examined Coderre and van Heuven's (2014) claim that the magnitude of bilingual advantages in the Simon and Stroop effect depends on the script similarity of a bilingual's L1 and L2 and concluded that the evidence was very weak. Furthermore, in an analysis of a composite database with much larger samples sizes (Paap, Darrow, et al., 2015) script similarity did not moderate the relationship between bilingualism and various measures of EF.

The trajectory of the investigation of the role of age-of-acquisition provides a powerful demonstration of how an initial positive finding is rarely questioned when conceptual replications produce a variety of outcomes. Luk, De Sa, and Bialystok (2011) reported the following pattern of differences across the three language groups: no language-group differences in monitoring and an advantage in inhibitory control for early bilinguals over both late bilinguals and monolinguals. Among studies including all three groups the overall pattern of results has never been replicated (Humphrey and Valian, 2012; Kalia, Wilbourn, & Ghio, 2014; Kapa & Colombo, 2013; Paap, Johnson, & Sawi, 2014; Pelham & Abrams, 2014; Tao et al., 2011). Not one of these subsequent studies resulted in a unique advantage of early bilinguals over both late bilinguals and monolinguals in inhibitory control.

Von Bastian, Souza, and Gade (2015) have conducted a study that provides a large and important piece of the puzzle of how specific bilingual experience might influence specific components of EF. The study is unusual in its scope. Three aspects of bilingualism were used as continuous predictors of EF: age-of-acquisition, usage, and proficiency. A parallel analysis used k-means clustering to create three groups that differed along a composite indicator formed from the three dimensions of bilingualism. The bilingualism measures were used to predict nine different components of EF and each of the nine components was measured with multiple tasks. Furthermore objective measures of SES and special activities (video-gaming, musical training, etc.) were matched across the three clusters. All of these measures were obtained from each of the 118 participants (students at Swiss universities) – a process requiring up to 4.5 h per participant. The results are completely consistent: no aspect of bilingualism or interaction of aspects significantly predicts measures of inhibitory control, monitoring, switching, or a composite reflecting a generalized cognitive advantage. This further confirms the conclusions that age-of-acquisition, L2/L1 proficiency, proportion of language use, or combinations of these specific experiences are not related to EF. In fact, the von Bastian et al. study shows no relationship at all on any component of EF. Furthermore, near significant findings trend in the direction of advantages for the least bilingual group. In summary, there does not appear to be any cumulative progress in identifying aspects of bilingualism that are necessary or sufficient for potentiating bilingual advantages in EF. A lack of cumulative progress is what one expects if bilingualism does not enhance EF.

3.8. Lack of convergent validity: bilingual advantages in what?

In earlier sections we discussed a variety of alternative explanations for why significant bilingual advantages in performance sometimes occur. This section raises a different concern, namely, that many of the standard measures of inhibition (or monitoring) obtained with nonverbal interference tasks lack convergent validity that is, they do not correlate with one another. If a measure lacks convergent validity, then individual or group differences are more likely to reflect a combination of chance factors and task-specific factors than differences in a domain-general component of EF.

3.8.1. *Convergent validity in inhibitory control*

Based on studies conducted in our laboratory and a very large number of others Paap and Sawi (2014) concluded that there is little or no convergence between measures of inhibitory control. This may seem surprising given that Miyake and Friedman's latent variable analyses showed significant correlations between the task variables assumed to require inhibition. One piece of the puzzle is that Miyake and Friedman never included the Simon interference effect as one of their measures. But even different variants of the Stroop task do not correlate with one another (Shilling, Chetwynd, & Rabbitt, 2002) nor do different versions of the flanker task correlate with one another (Salthouse, 2010). These nonverbal interference tasks are sensitive to the conflict present on incongruent trials, but they do not appear to measure individual differences in domain-general inhibitory control.

In their seminal study of inhibition Friedman and Miyake (2004) considered three categories of interference tasks, but the best model did not empirically separate response inhibition (viz., stop signal, antisaccade, Stroop) from interference control (e.g., resistance to distraction as in the flanker task). Unsworth and colleagues (Unsworth, Fukuda, Awh, & Vogel, 2014; Unsworth, McMillan, Brewer, & Spillers, 2012), also treat both types of inhibition as a single latent variable. Thus, the outcome of these latent variable analyses do not mesh with studies assuming that bilinguals are better at interference control, but not response inhibition (e.g., Blumenfeld & Marian, 2014).

3.8.2. *Convergent validity in monitoring*

The current trend in the literature on bilingual advantages is to appeal to monitoring (e.g., Costa, et al., 2008) or less formally defined constructs such as *coordination* or *mental flexibility* (e.g., Kroll & Bialystok, 2013) as the essence of the bilingual advantage in EF. The construct of monitoring is chameleon-like with respect to the target of the monitor. In different contexts it may refer to the monitoring of incoming information from the environment in terms of potential affordances for present goals, it may refer to monitoring the contents of working memory for goal relevant information, or it may refer to a special "conflict monitor" that provides imperative information for the up-regulation of goal relevant neural pathways.

One strategy for isolating monitoring from inhibitory control and switching is to derive a measure not involving trials where conflict is present (or where a switch is required), thus removing the effects of having to actually resolve a conflict (or having to actually switch tasks). The most common measures are global RT (the average across both congruent and incongruent trials) or simply the mean RT on congruent trials. Global RT is clearly inferior to RT on the congruent trials as conflict must be resolved on half the trials. Even when one includes only congruent trials considerations of experimental control would seem to require a baseline condition that experiences no conflict at all throughout an extended block of trials: otherwise any group differences could be attributed to differences in perceptual processing, motor processing, or general fluid intelligence rather than monitoring. Thus, a purer test would be a difference score between the congruent trials from a mixed block and the same type of trials in a pure block that includes no conflict trials. This difference score is sometimes referred as "mixing costs". Studies that include this control frequently do not show a bilingual advantage (e.g., Luk, et al., 2011; Paap & Greenberg, 2013; Paap & Sawi, 2014).

The convergent validity across these measures of monitoring has been examined in the three studies conducted by Paap and Greenberg (2013) and a fourth study reported in Paap and Sawi (2014). It is the case that global RT in the flanker task is highly correlated with global RT in the Simon task ($r = +.60$), but as cautioned above a host of nonexecutive processes may be contributing to the association. This possibility is reinforced by observing that the mixing costs for the Simon and flanker tasks yield correlations near zero. Because mixing costs in the color-shape switching task are also assumed to reflect the monitoring component of EF, this measure can also be correlated with the mixing costs obtained in the Simon and flanker tasks: the correlation between mixing in the switching task and mixing in the flanker or Simon task is usually near zero and in one case was slightly negative. These results reinforce the need for additional work in operationally defining constructs like monitoring, coordination, or mental flexibility and establishing their validity as measures of EF. Significant bilingual advantages obtained with only a single measure of the EF component are likely to be caused by task-specific factors and unlikely to replicate with a conceptually similar measure derived from a different task. Only studies using two or more measures that converge with each other offer compelling evidence that differences in performance reflect differences in a specific component of EF.

4. Does neuroscience data provide compelling evidence for a bilingual advantage?

Many researchers believe that the inconsistent evidence for the bilingual advantage in behavioral data has been strengthened by the studies that have jointly explored both behavioral and neuroscience measures. Although we agree that this could be so in theory we have argued that this has yet to happen (Paap, 2014; Paap, Sawi, Dalibar, Darrow, & Johnson, 2014, 2015).

4.1. *The alignment problem*

In general, cortical areas shown to be involved in managing two languages overlap with those shown to be involved with inhibitory control and switching (Bialystok et al., 2012). Furthermore, it is clear from the neuroimaging results that the neural processing of bilinguals and monolinguals differs during the performance of EF tasks. This is consistent with the view that coordinating two languages leads to a reorganization of neural networks in cortical areas involved in EF. However, reorganization to accommodate bilingualism does not logically need to result in more efficient performance. Alternatively, it could lead to comparable performance or even to a compromise that results in inferior performance. Thus, it is imperative that the observed neural differences be aligned with the behavioral differences so that bilingual advantages in actual performance can be confirmed. We propose that the existence of a behavioral phenomenon can only be adjudicated at the behavioral level, an argument first advanced by Hilchey and Klein (2011).

4.1.1. *Alignment problems in the Simon task*

Bialystok, Craik, Grady, et al. (2005) reported that two groups of bilinguals showed substantial overlap in terms of the loci associated with fast responding in a Simon task and that these markedly differed from the specific areas associated with fast responding for monolinguals. However, these neural differences did not align with the differences observed in behavior. There were no group differences at all in the size of the Simon interference effect, but there was a global RT advantage for the Chinese–English bilinguals compared to both the French–English bilinguals and the monolingual group. Bialystok et al. offer this summary statement: “The present study used magneto-encephalography (MEG) to determine the neural correlates of the bilingual advantage previously reported for behavioral measures in conflict tasks” p. 40. But the pattern of global RT does not match the pattern of MEG differences: only the Chinese–English bilinguals showed a global RT advantage. Furthermore given that the RT advantage of the Chinese–English bilinguals was also present in the control condition (a pure block of no conflict trials) there is simply no evidence at all in this conflict task for a bilingual advantage in either monitoring or inhibitory control.

A similar misalignment between neural and behavioral differences in the Simon task was reported by Ansaldo, Ghazi-Saidi, and Adrover-Roig (2015) using fMRI. There were no differences in either accuracy or RT between a group of French–English bilinguals and a group of French monolinguals. This null result was true for both the magnitude of the Simon interference effect and global RT. The neuroimaging results showed that on incongruent trials monolinguals activated the prefrontal cortex network, whereas bilinguals activated the left inferior parietal lobe. The authors interpretation was that “… bilinguals did not need to resort to a cognitive control circuit to resolve a visuospatial conflict whereas monolinguals did” p. 10. This recruitment of different neural circuits is intriguing, but it is curious that the bilingual group with supposedly superior EF recruits an alternative mechanism for conflict resolution and that mechanism provides no advantages at all in performance. Ansaldo et al. conclude that “bilingual advantages can be detected by neuroimaging techniques even when it is not evident behaviourally” p. 11. If these differences in neural circuits have not caused behavioral advantages in this study or in the sizeable majority of other studies using the Simon task, then they should probably not be called “advantages”.

4.1.2. *Alignment problems in the flanker task*

A study by Luk, Anderson, Craik, Grady, and Bialystok (2010) claims to have shown neural correlates of a bilingual advantage in inhibitory control. This conclusion runs counter to the actual behavioral results. With respect to the RT data there was neither a main effect of group nor a significant Group × Trial Type interaction. Thus, there was no behavioral evidence for either a monitoring or inhibitory control advantage. With respect to neural differences Luk et al. reported that the regions associated with faster responding were the same on both congruent and incongruent trials for monolinguals, but with different regions for bilinguals. This additional and different pathway employed by bilinguals on incongruent flanker trials led Luk et al. to conclude that bilinguals have superior inhibitory control. But the flanker effects for the two groups were nearly identical.

4.1.3. *Misalignment in a switching task*

Rodriguez-Pujada, et al. (2013) compared 18 Catalan–Spanish bilinguals to 18 Spanish monolinguals in a color-shape switching task where cues to either switch or stay appeared relatively infrequently. As expected there were no differences in behavioral switch costs, but fMRI measures showed that bilinguals use language-control areas more than monolinguals. Rodriguez-Pujada et al. conclude that bilingualism exerts an effect on the neural circuitry employed during nonverbal task switching. Conservatively, they do not jump to the conclusion that the neural differences are indicative of a bilingual advantage. To the contrary, “we may even argue that the brain control exerted by monolinguals in the present task is even more efficient than for bilinguals” p. 7.

4.2. *The valence-ambiguity problem*

The interpretation of neural differences is very risky in the absence of behavioral data that show the same pattern of significant differences. When alignment problems occur the *ambiguity* of many neural differences is exposed. One type of ambiguity, referred to as *valence ambiguity* (Paap, Sawi, Dalibar, et al., 2014), occurs if increasing neural scores are interpreted as having a positive effect on performance by some researchers and as having a negative effect by others. In considering behavioral measures like speed or accuracy in choice RT tasks it seems to be inherently the case that individuals or groups who are faster and/or more accurate enjoy a performance advantage over those who are slower and/or less accurate. Speed-accuracy tradeoffs can complicate the interpretation, but the point remains that within each measure faster is better and more accurate is better. Differences in neural measures have no such *acta est fabula plaudit*: they must be interpreted in the context of their behavioral consequences.

4.2.1. *What does a larger N400 amplitude mean?*

To take one example Paap and Liu (2014) challenged the interpretation of the language-group differences observed by Moreno, Bialystok, Wodniecka, and Alain (2010) in the N400 component of the ERP during sentence grammaticality judgments for sentences that were syntactically correct, but semantically anomalous. The N400 is generally assumed to index difficulty in semantic integration during sentence processing. Moreno et al. interpreted the larger N400 components in bilinguals as a bilingual *advantage* in conflict resolution while Paap and Liu argued that larger N400s are indicative of a bilingual *disadvantage* because the larger N400s on the semantically anomalous sentences indicated that the bilinguals were less able to filter out the task-irrelevant semantics.

4.2.2. *What does a larger N2 amplitude mean?*

Fernandez, Acosta, Douglass, Doshi, and Tartar (2014) compared English monolinguals to Spanish–English bilinguals on both auditory and visual Go/NoGo tasks while recording ERPs. As predicted, bilinguals showed larger N2 amplitudes during the auditory NoGo trials, which require inhibitory control, but no differences during the Go trials. In contrast there were no group differences in N2 amplitude during the visual NoGo trials. A disadvantage of the Go/NoGo task is that one cannot check to see if the behavioral differences align with the N2 differences because there are no behavioral responses on a NoGo trial unless the participant false alarms and, in this study, false alarms were too rare to analyse. Fernandez et al. concluded that bilinguals have “enhanced inhibitory control”.

In a commentary on this study Paap, Sawi, Dalibar, et al. (2015) ask if larger N2 amplitudes are really indicative of better inhibitory control. Some of the most compelling evidence comes from developmental studies showing that N2 amplitude in the Go/NoGo task declines over the span of 7–16 years even when potential physical artifacts are taken into account (Lamm, Zelazo, & Lewis, 2006). Similarly, Espinet, Anderson, and Zelazo (2012) found that 3 to 4.5 year-old-children who can pass the dimensional change card sort (DCCS) task have significantly smaller N2 amplitudes during the post-switch phase of the task compared to those who perseverate and fail the test. The last results are compelling because the neural results align with the behavioral performance results (viz., success *vs* failure on the DCCS). These studies strongly suggest that a smaller N2 amplitude reflects superior performance.

4.2.3. *What does greater ACC activity mean?*

The study by Abutalebi, et al. (2012) study presents another case of valence ambiguity in its conclusion that bilinguals adapt better to conflicting situations because “… they seem to require less ACC activity to outperform monolinguals” (p. 2084), and “… require fewer neural resources to monitor cognitive conflict” (p. 2085). Paap, Sawi, Dalibar et al. (2014) discuss the myriad of interpretations of the function of the ACC and the interested reader is encouraged to consult the extended discussion.

4.3. *“Conflict monitoring” in neuroimaging studies*

It is noteworthy that neuroimaging studies of “conflict monitoring” actually study “inhibitory control”. If the goal is to determine the neural circuitry that is responsible for a monitoring advantage that applies equally to both congruent and incongruent trials, then the key contrast should be the congruent trials from a mixed block and identical congruent trials in a pure block that includes no incongruent trials. This is not the contrast studied and the article by Abutalebi, et al. provides an example of this discrepancy. The paper is titled *“Bilingualism Tunes the Anterior Cingulate Cortex for Conflict Monitoring*”, but the neural contrast referred to as “conflict monitoring” (incongruent BOLD – congruent BOLD) actually corresponds to the behavioral contrasts typically referred to as “inhibitory control” (incongruent RT – congruent RT). Abutalebi et al. report no behavioral differences in global RT and do not even run a baseline condition that would enable them to discover cortical regions involved in monitoring and preparing for conflict that would facilitate both congruent and incongruent trials. In short, there is currently a disconnect between the favored EF component in the behavioral literature (monitoring/mental flexibility) and the contrasts used most frequently in the neuroimaging work that typically compare statistical maps between congruent and incongruent trials.

5. Pursuing and characterizing bilingual advantages; if they exist

Based on an array of evidence we have made a case that it is likely that bilingual advantages in EF do not exist. In this final section we would like to consider the implications of our review if our tentative conclusion is wrong and bilingual advantages are real. One possibility is that most bilingual experiences enhance EF, but that the effect is small. In broad stroke this fits the fact that significant group differences occur infrequently and are small in average magnitude. More specifically it does not fit the fact that the studies with the largest sample sizes yield null results. A more likely scenario is that bilingual advantages accrue only in very specific circumstances that pair the right set of bilingual experiences with the resonating set of EF measures. However, based on studies that have investigated age-of-acquisition, L2/L1 ratio, multilingualism, language similarity, and other aspects of bilingualism it appears that the results are inconsistent. The most comprehensive effort to match specific aspects of bilingualism with specific measures of EF (von Bastian et al., 2015) yielded consistent, but null results. Thus, those hypothesizing that bilingual advantages occur in only specific circumstances seem obligated to admit that, at this point in time, the sufficient circumstances for producing bilingual advantages are undetermined.

One possible roadmap for pursuing the specific circumstances for producing bilingual advantages was provided by Paap and Greenberg (2013) and could stand to be updated. Studies should start with a theory of how two or more languages are co-ordinated and specify which (if any) of the control mechanisms employ domain-general EF. This should lead to the identification of the particular type (or combination of types) of bilingual experience that is most important to enhancing that component of EF. That critical experience should play the lead role in predicting which bilingual groups should have better EF and, in turn, be better than monolinguals. The groups should be matched on SES, immigrant status, culture and so forth using operationally defined measures. Tasks and measures should be selected that have demonstrated convergent validity and, even then, it is best to have two measures, derived from two different tasks, included in the study so that one can rule out the concern that any obtained bilingual advantage in performance is task specific. The number of participants in each group should be determined in advance and based on a reasonable level of desired power. In estimating power a small effect size should be assumed given the meta-analysis reported by de Bruin et al. and the effect sizes reported in Hilchey and Klein (2011) and Hilchey et al. (in press).

6. Conclusion

It is likely that bilingual advantages in EF do not exist. If they do exist they are restricted to specific aspects of bilingual experience that enhance only specific components of EF. Such constraints, if they exist, have yet to be determined.

Acknowledgments

We thank Jack Darrow and Chirag Dalibar for their comments on the manuscript and Morgan Bockelman, Sarah Wagner, Oscar Gonzalez-Munoz, Angel Avalos, and Eugene Eusebio, for their contributions to the project.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cortex.2015.04.014.

REFERENCES

Abutalebi, J., Della Rosa, P. A., Green, D. W., Hernandez, M., Scifo, P., Keim, R., et al. (2012). Bilingualism tunes the anterior cortex for conflict monitoring. *Cerebral Cortex, 22*(9), 2076–2086.

Ansaldo, A. I., Ghazi-Saidi, L., & Adrover-Roig, D. (2015). Interference control in elderly bilinguals: appearances can be misleading. *Journal of Clinical and Experimental Neuropsychology*, 1–16.

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L. J., et al. (2014). Is there a bilingual advantage in the ANT task? Evidence from children. *Frontiers in Psychology: Language Sciences, 5*, 1–12.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives in Psychological Science, 7*, 534–554.

von Bastian, C. C., Sousa, A. S., & Gade, M. (2015). *No evidence for bilingual cognitive advantages: A test of four hypotheses.* Manuscript submitted for publication.

Bialystok, E. (2011). Reshaping the mind: the benefits of bilingualism. *Canadian Journal of Experimental Psychology, 65*(4), 229–235.

Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of dementia. *Neuropsychologia, 45*(2), 459–464.

Bialystok, E., Craik, F. I. M., Grady, C., Chau, W., Ishii, Y., Gunji, A., et al. (2005). Effects of bilingualism on cognitive control in the Simon task: evidence from MEG. *NeuroImage, 24*, 40–49.

Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology of Aging, 19*(2), 290–303.

Bialystok, E., Craik, F. I. M., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology Learning Memory and Cognition, 34*(4), 859–873.

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Science, 16*(4), 240–250.

Billig, J. D., & Scholl, A. P. (2011). The impact of bilingualism and aging on inhibitory control and working memory. *Organon*, 39–52.

Blom, E., Küntay, A. C., Messer, M., Verhagen, J., & Leseman, P. (2014). The benefits of being bilingual: working memory in bilingual Turkish-Dutch children. *Journal of Experimental Child Psychology, 128*, 105–119.

Blumenfeld, H. K., & Marian, V. (2014). Cognitive control in bilinguals: advantages in stimulus–stimulus inhibition. *Bilingualism: Language and Cognition, 17*(3), 610–629.

de Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: an example of publication bias. *Psychological Science, 26*(1), 99–107.

Calvo, A., & Bialystok, E. (2014). Independent effects of bilingualism and socioeconomic status on language ability and executive functioning. *Cognition, 130*(3), 278–288.

Carlson, S. M., & Choi, H. P. (2009, April). Bilingual and bicultural: Executive function in Korean and American children. In *Paper presented at the 2009 biennial meeting of the society for research in child development. Denver, Colorado*.

Chen, S. H., Zhou, Q., Uchikoshi, Y., & Bunge, S. A. (2014). Variations on the bilingual advantage? Links of Chinese and English proficiency to Chinese American children's self regulation. *Frontiers in Psychology, 5*, 1069.

Chertkow, H., Whitehead, V., Phillips, N., Wolfson, C., Atherton, J., & Bergman, H. (2010). Multilingualism (but not always bilingualism) delays the onset of Alzheimer disease: evidence from a bilingual community. *Alzheimer Disease and Associated Disorders, 24*(2), 118–125.

Coderre, E. L., & van Heuven, W. J. B. (2014). The effect of script similarity on executive control in bilinguals. *Frontiers in Psychology.*, 5, 1070.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Costa, A., Hernández, M., Costa-Faidella, J., & Sebastián-Gallés, N. (2009). On the bilingual advantage in conflict processing: now you see it, now you don't. *Cognition, 113*, 135–149.

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: evidence from the ANT task. *Cognition, 106*, 59–86.

Crane, P. K., Gibbons, L. E., Arani, K., Nguyen, V., Rhoads, K., McCurry, S. M., et al. (2009). Midlife use of written Japanese and protection from late life dementia. *Epidemiology, 20*(5), 766–774.

Duñabeitia, J. A., Hernández, J. A., Antón, E., Macizo, P., Estévez, A., Fuentes, L. J., et al. (2014). The inhibitory advantage in bilingual children revisited. *Experimental Physiology, 61*(3), 234–251.

Engel de Abreu, P. M. J., Cruz-Santos, A., Tourinho, C. J., Martin, R., & Bialystok, E. (2012). Bilingualism enriches the poor: enhanced cognitive control in low-income minority children. *Psychological Science, 23*, 1364–1371.

Espinet, S. D., Anderson, J. E., & Zelazo, P. D. (2012). N2 amplitude as a neural marker of executive function in young children: an ERP study of children who switch versus perseverate on the dimensional change card sort. *Developmental Cognitive Neuroscience*, (1), S49–S58.

Fernandez, M., Acosta, J., Douglass, K., Doshi, N., & Tartar, J. L. (2014). Speaking two languages enhances an auditory but not a visual neural marker of cognitive inhibition. *AIMS Neuroscience, 1*(2), 145–157.

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

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

Fuller-Thomson, E. (2015). Emerging evidence contradicts the hypothesis that bilingualism delays dementia onset. *Cortex*, 1–3.

Fuller-Thomson, E., & Kuh, D. (2014). The healthy migrant effect may confound the link between bilingualism and delayed onset of Alzheimer's disease. *Cortex, 52*, 128–130.

Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J., & Smith, C. D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *The Journal of Neuroscience, 33*(2), 387–439.

Hernández, M., Martin, C. D., Barceló, F., & Costa, A. (2013). Where is the bilingual advantage in task-switching? *Journal of Memory and Language, 69*, 257–276.

Hilchey, M. D., & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for plasticity of executive control processes. *Psychonomic Bulletin & Review, 18*, 625–658.

Hilchey, M. D., Saint-Aubin, J., & Klein, R. M. (2015). Does bilingual exercise enhance cognitive fitness in non-linguistic executive processing tasks. In J. W. Schwieter (Ed.), *Cambridge handbook of bilingual processing*. Cambridge University Press. in press.

Humphrey, A. D., & Valian, V. V. (2012, November). Multilingualism and cognitive control: Simon and flanker task performance in monolingual and multilingual young adults. In *Presentation at the 53rd annual meeting of the Psychonomic Society. Minneapolis, MN*.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23*(5), 524–532.

Kalia, V., Wilbourn, M. P., & Ghio, K. (2014). Better early or late? Examining the influence of age of exposure and language proficiency on executive function in early and late bilinguals. *Journal of Cognitive Psychology, 26*(7), 6990713.

Kapa, L. L., & Colombo, J. (2013). Attentional control in early and later bilingual children. *Cognitive Development, 28*, 233–246.

Kirk, N. W., Fiala, L., Scott-Brown, K., & Kempe, V. (2014). No evidence for reduced Simon cost in elderly bilinguals and bidialectals. *Journal of Cognitive Psychology*, 1–9.

Kousaie, S., & Phillips, N. A. (2012). Ageing and bilingualism: absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. *The Quarterly Journal of Experimental Psychology, 65*(2), 356–369.

Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology*, 497–514.

Lamm, C., Zelazo, P. D., & Lewis, M. D. (2006). Neural correlates of cognitive control in childhood and adolescence: disentangling the contributions of age and executive function. *Neuropsychologia, 44*, 2139–2148.

Lawton, D. M., Gasquoine, P. G., & Weimer, A. A. (2015). Age of dementia diagnosis in community dwelling bilingual and monolingual Hispanic Americans. *Cortex, 66*, 141–145.

Luk, G., Anderson, J. A. E., Craik, F. I. M., Grady, C., & Bialystok, E. (2010). Distinct neural correlates for two types of inhibition in bilinguals: response inhibition versus interference suppression. *Brain and Cognition, 74*, 347–357.

Luk, G., De Sa, E., & Bialystok, E. (2011). Is there a relation between onset age of bilingualism and enhancement of cognitive control? *Bilingualism: Language and Cognition, 14*, 588–595.

Mahoney, M. J. (1977). Publication prejudices: an experimental study of confirmation bias in the peer review system. *Cognitive Therapy and Research, 1*, 161–175.

Marzecová, A., Asanowicz, D., Krivá, L., & Wodniecka, Z. (2013). The effects of bilingualism on efficiency and lateralization of attentional networks. *Bilingualism: Language and Cognition, 16*(3), 608–623.

Mercier, J., Pivneva, I., & Titone, D. (2014). Individual differences in inhibitory control relate to bilingual spoken word processing. *Bilingualism: Language and Cognition, 17*(1), 89–117.

Miller, G. A., & Chapman, J. P. (2001). Misundertanding analysis of covariance. *Journal of Abnormal Psychology, 110*(1), 40–48.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: four general conclusions. *Current Directions in Psychology, 21*(1), 8–14.

Moradzadeh, L., Blumenthal, G., & Wiseheart, M. (2014). Musical training, bilingualism, and executive function: a closer look at task switching and dual-task performance. *Cognitive Science, 39*, 1–29.

Moreno, S., Bialystok, E., Wodniecka, Z., & Alain, C. (2010). Conflict resolution in sentence processing by bilinguals. *Journal of Neurolinguistics, 23*, 564–579.

Morton, J. B., & Carlson, S. M. (2015). In M. Hoskyn, G. Iarocci, & A. Young (Eds.), *The bilingual advantage: Evidence and alternative views*. Oxford University Press. in press.

Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science, 10*, 719–726.

Mor, B., Yitzhaki-Amsalem, S., & Prior, A. (2014). The joint effect of bilingualism and ADHD on executive functions. *Journal of Attention Disorders, 19*(6), 527–541.

Paap, K. R. (2014). The role of componential analysis, categorical hypothesizing, replicability and confirmation bias in testing for bilingual advantages in executive functioning. *Journal of Cognitive Psychology, 26*(3), 242–255.

Paap, K. R., Darrow, J., Dalibar, C., & Johnson, H. A. (2015). Effects of script similarity on bilingual advantages in executive control are likely to be negligible or null. *Frontiers in Psychology, 5*, 1539.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology, 66*, 232–258.

Paap, K. R., Johnson, H. A., & Sawi, O. (2014). Are bilingual advantages dependent upon specific tasks or specific bilingual experiences? *Journal of Cognitive Psychology, 26*(6), 615–639.

Paap, K. R., & Liu, Y. (2014). Conflict resolution in sentence processing is the same for bilinguals and monolinguals: the role of confirmation bias in testing for bilingual advantages. *Journal of Neurolinguistics, 27*(1), 50–74.

Paap, K. R., & Sawi, O. (2014). Bilingual advantages in executive functioning: problems in convergent validity, divergent validity, and the identification of the theoretical constructs. *Frontiers in Psychology, 5*(962), 1–15.

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

Paap, K. R., Sawi, O., Dalibar, C., Darrow, J., & Johnson, H. A. (2015). Beyond panglossian optimism: larger N2 amplitudes probably signal a bilingual disadvantage in conflict monitoring. *AIMS Neuroscience, 2*(1), 12–17.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531–536.

Pelham, S. D., & Abrams, L. (2014). Cognitive advantages and disadvantages in early and late bilinguals. *Journal of Experimental Psychology Learning Memory and Cognition, 40*(2), 313–325.

Prior, A., & Gollan, T. (2011). Good language-switchers are good task-switchers: evidence from Spanish–English and Mandarin–English bilinguals. *Journal of International Neuropsychological Society, 17*, 1–10.

Prior, A., & Gollan, T. H. (2013). The elusive link between language control and executive control: a case of limited transfer. *Journal of Cognitive Psychology, 25*(5), 622–645.

Prior, A., & MacWhinney, B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition, 13*, 253–262.

Rodriguez-Pujada, A., Sanjuan, A., Ventura-Campos, N., Román, P., Martin, C., Barceló, F., et al. (2013). Bilinguals use language-control brain areas more than monolinguals to perform non-linguistic switching tasks. *PLoS One, 8*(9), 73028.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Salthouse, T. A. (2010). Is flanker-based inhibition related to age? Identifying specific influences of individual differences on neurocognitive variables. *Brain and Cognition, 73*, 51–61.

Salvatierra, J. L., & Rosselli, M. (2011). The effect of bilingualism and age in inhibitory control. *International Journal of Bilingualism, 15*, 26–37.

Sanders, A. E., Hall, C. B., Katz, M. J., & Lipton, R. B. (2012). Non-native language use and risk of incident dementia in the elderly. *Journal of Alzheimer's Disease*, 99–108.

Schroeder, S. R., & Marian, V. (2012). A bilingual advantage for episodic memory in older adults. *Journal of Cognitive Psychology, 24*(5), 591–601.

Shilling, V. M., Chetwynd, A., & Rabbitt, P. M. A. (2002). Individual inconsistency across measures of inhibition: an investigation of the construct validity of inhibition in older adults. *Neuropsychologia, 40*, 605–619.

Tao, L., Marzecova, A., Taft, M., Asanowicz, D., & Wodniecka, Z. (2011). The efficiency of attentional networks in early and late bilinguals: the role of age of acquisition. *Frontiers in Psychology, 2*, 123.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. *Cognitive Psychology, 71*, 1–26.

Unsworth, N., McMillan, B. D., Brewer, G. A., & Spillers, G. J. (2012). Everyday attention failures: an individual differences investigation. *Journal of Experimental Psychology Learning Memory and Cognition, 38*(6), 1765–1772.

Valian, V. (2014). Bilingualism and cognition. *Bilingualism: Language and Cognition, 18*(1), 3–24.