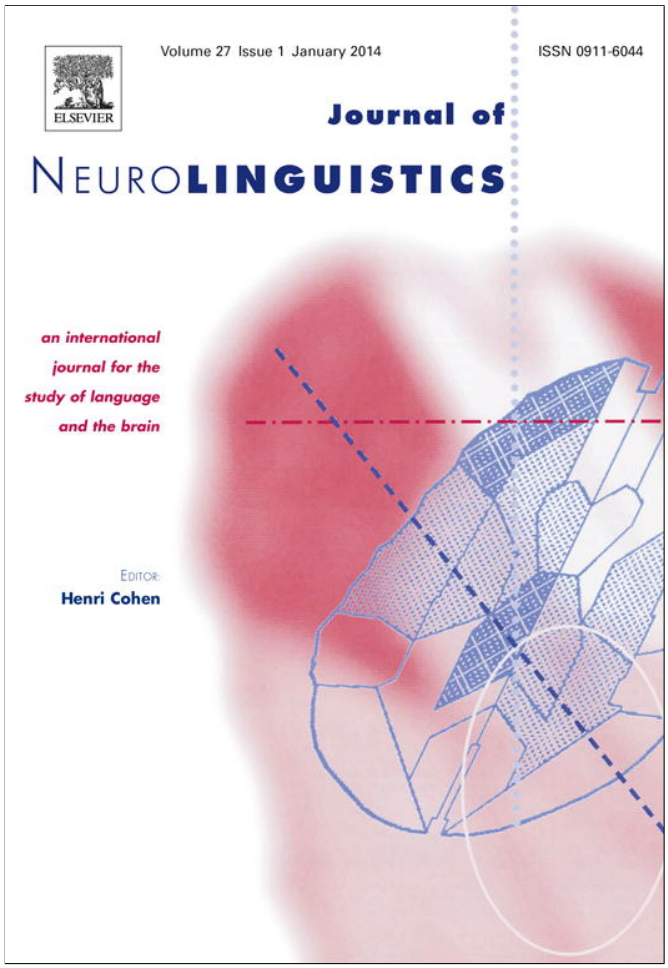
Conflict resolution in sentence processing is the same for bilinguals and monolinguals: The role of confirmation bias in testing for bilingual advantages

ARTICLE *in* JOURNAL OF NEUROLINGUISTICS · JANUARY 2014

Impact Factor: 1.49 · DOI: 10.1016/j.jneuroling.2013.09.002

2 AUTHORS, INCLUDING:

K. R. Paap
San Francisco State University
55 PUBLICATIONS **1,751** CITATIONS



Journal of Neurolinguistics 27 (2014) 50–74

Contents lists available at ScienceDirect

Journal of Neurolinguistics

journal homepage: www.elsevier.com/locate/jneuroling

Conflict resolution in sentence processing is the same for bilinguals and monolinguals: The role of confirmation bias in testing for bilingual advantages

Kenneth R. Paap*, Yunyun Liu

San Francisco State University, 1600 Holloway Avenue, EP301, Department of Psychology, San Francisco, CA 94132, USA

ARTICLE INFO

Article history:
Received 21 June 2013
Received in revised form 31 August 2013
Accepted 6 September 2013

Keywords:
Bilingual advantage
Control
Executive functions
Confirmation bias
Homograph
Conflict

ABSTRACT

The primary purpose of this study was to test for bilingual advantages in conflict resolution during sentence processing. Experiment 1 examined the time-course of a homograph-interference effect when test words were either presented immediately after the sentence-final word or after a delay. Bilinguals and monolinguals were equally adept at using the extra time to suppress the context-inappropriate meaning when the sentence-final word was a homograph. Experiment 2 tested the hypothesis that bilingual advantages in inhibitory control enable bilinguals to close the performance gap in a *sentence grammaticality* task compared to a *sentence acceptability* task. The critical group by task interactions were not significant across four different behavioral measures. Recent studies offering opposing conclusions were examined for the influence of confirmation bias.

The research question under investigation is this: Are there bilingual advantages in behavioral measures of conflict resolution that can reasonably be attributed to differences in the inhibitory control component of executive processing (EP)? Elsewhere (Paap & Greenberg, 2013; Paap & Sawi, 2013), we extend the question to the monitoring and switching components of EP, but the empirical and theoretical work reported here focus only on inhibitory control.

* Corresponding author. Tel.: +1 415 338 6840; fax: +1 415 338 2398.
E-mail address: kenp@sfsu.edu (K.R. Paap).

0911-6044/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jneuroling.2013.09.002

1. Bilingual inhibitory control advantages (BICA) in nonverbal tasks

In a seminal article on bilingual advantages in executive functioning Bialystok, Craik, Klein, and Viswanathan (2004) proposed that bilinguals are better at selecting goal-relevant information and suppressing competing and distracting information. Bilinguals exercise this type of control at two levels: (1) at a high level of goal setting when one language is selected and the other is inhibited and (2) at a lower level where the lexical forms of the goal relevant language are activated and the competing translation equivalents are inhibited (e.g., Green, 1998). Bialystok et al. further hypothesized that this ubiquitous practice hones a general ability, not specific to language, and consequently that bilinguals should be less vulnerable to interference in nonlinguistic tasks. The standard marker of inhibitory control in these tasks is the difference in mean response time between trials that require conflict resolution compared to those that do not. In the Stroop (both verbal and nonverbal versions), Simon, and Eriksen flanker tasks conflict occurs on a subset of trials because a potent but task-irrelevant stimulus is paired in an incongruent manner with the task-relevant stimulus. The effectiveness of this control can be inferred from differences in response time between congruent trials and incongruent trials with smaller interference effects implying superior ability. In a recent and comprehensive review of bilingual advantages in EP Hilchey and Klein (2011) reviewed 31 experiments and concluded that evidence for a bilingual advantage in inhibitory control is rare in both children and young adults. More emphatically they assert that the collective evidence "*…is simply inconsistent with the proposal that bilingualism has a general positive effect on inhibitory control processes*" p. 629.

Since Hilchey and Klein's insightful review there have been several additional tests for bilingual advantages in inhibitory control in these nonverbal tasks and they overwhelmingly report no group differences and on one occasion a monolingual advantage. Kousaie and Phillips (2012a) found no behavioral differences between groups of young adults in the Stroop, Simon, or flanker tasks (i.e., 0 group differences out of three tests). The Kousaie and Phillips' (2012b) study used both young adults and older adults and found no differences in the magnitude of Stroop interference (i.e. 0 group differences out of two tests). A similar study by Humphrey and Valian (2012) using the Simon and flanker tasks follows the same pattern. Four different groups of multilinguals (lifelong balanced bilinguals, late balanced bilinguals whose native language is English, late balanced bilinguals whose native language is not English, and trilinguals) show Simon and flanker effects statistically equivalent to a group of English monolinguals (i.e., 0 bilingual advantages out of eight tests). Paap and Greenberg (2013) found no bilingual advantage in three Simon experiments and one flanker experiment (i.e., 0 bilingual advantages out of four tests with one monolingual advantage). Sawi & Paap (2013) tested an additional relatively large sample (n's > 50) of bilinguals and monolinguals in both the Simon and flanker effects and neither task resulted in a bilingual advantage. In another study using young adult participants Ryskin and Brown-Schmidt (2012) report no bilingual advantages in either the Stroop or flanker task.

Turning to studies using children rather than young adults a study by Engel de Abreu, Cruz-Santos Tourinho, Martin, and Bialystok (2012) does support the hypothesis that acquiring a second language yields smaller interference effects in the flanker task. Engel de Abreu et al. characterize their results as remarkable because, at the time the enhanced inhibitory control was measured, the bilingual children have strikingly low vocabulary scores in Luxembourgish and thus are not at all proficient in their L2. In this context, these differences in the magnitude of the interference effect invite consideration of alternative or additional reasons for the group differences. The matching reported by Engel de Abreu et al. is quite thorough, but no one study can match or hold constant all factors that could contribute to group differences on tasks assumed to measure EF. In this case, in order to hold Portuguese culture constant,[1] the bilinguals were immigrants and the monolinguals were not. Also, two years of preschool are compulsory in Luxembourg, but that is not the case in Portugal. Although the Portuguese monolinguals did attend preschools the quality of those programs was not formally assessed and may have differed. More generally, it is very difficult to assure that children living in Portugal have had the same experiences as those living in Luxembourg.

[1] One could argue that the second-generation Portuguese children living their entire life in Luxembourg are no longer precisely matched in culture to the monolingual children living in Portugal.

As impressive as the Engel de Abreu et al. study is, one would have more confidence if the bilingual advantage on measures of inhibition generalized to children in the same age range, but with groups that did not differ in immigrant status and country of schooling. In a recent study Duñabeitia et al. (2013) compared Spanish monolinguals to Basque-Spanish bilinguals who were carefully matched on a large number of indices. Both groups were administered a verbal Stroop task and a non-verbal version of the same task (viz., the number size-congruency task). Results were unequivocal showing that bilingual and monolingual participants performed equally in these two tasks across all the indices or markers of inhibitory skills explored. The lack of differences between monolingual and bilingual children extended to all the age-ranges tested (six successive grades with an age range of 8–13 years). It is instructive to note that the group sizes for the study finding no group differences ($n = 252$) was far larger than the study reporting differences ($n = 40$).

One possible resolution of the conflicting results is to conclude that the large group differences obtained by Engel de Abreu et al. were due to unmeasured advantages for children living in Luxembourg compared to Portugal. A contrasting resolution might attribute the conflicting results to the type of bilinguals. But, the bilingual children in Luxembourg had low proficiency in their L2 whereas the bilingual children in the Basque Country of Spain were far more balanced and fluent bilinguals. Thus, why should the bilingual experience of the beginning bilingual lead to large bilingual advantages and those of the far more balanced bilinguals to no advantages at all?

Recall that Kousaie and Phillips (2012b) study used both young adults and older adults and found no differences in the magnitude of Stroop interference in these non-immigrant samples. Likewise, Kirk, Scott-Brown, and Kempe (2013) report that when immigrant status and culture are matched, there are no differences in inhibitory control (in the Simon task) between older Gaelic-English bilinguals and three monolingual control groups.

For those readers keeping score, since Hilchey and Klein's review, there has been one (Engel de Abreu et al.) significant bilingual advantage in Simon, Stroop, or flanker effects out of 23 tests. Any neutral and objective umpire must surely side with Hilchey and Klein that there is no compelling evidence that bilingualism enhances inhibitory control, at least so far as it would reduce the magnitude of the interference effect in these nonverbal tasks.

2. Neuroimaging and the bilingual advantage in inhibitory control

Although the empirical contribution of this study will focus on conflict resolution in language processing a reviewer encouraged discussion of the neuroimaging research that has linked bilingualism to differences in neural processing during the performance of nonverbal tasks like the Simon and flanker task because these linkages "*reinforce drastically*" the evidence supporting bilingual advantages in inhibitory control. We disagree, but understand that the associative trail can be very seductive. In general, cortical areas shown to be involved in managing two languages overlap with those shown to be involved with inhibitory control and other components of EP (see section on neural correlates of cognitive reorganization in Bialystok, Craik, & Luk, 2012). Furthermore, it is clear from the neuroimaging results that the neural processing of bilinguals and monolinguals differs during the performance of the Simon and flanker tasks, in part, because some of the cortical areas recruited by bilinguals are not employed by monolinguals. All of this is consistent with the view that managing two languages leads to an organization (or reorganization) of neural networks in cortical areas (broadly considered) involved in EP. However, a 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.

We contend that the existence of a behavioral phenomenon (e.g., a bilingual advantage in conflict resolution) can only be adjudicated at the behavioral level. This is not to say that the neuroscience isn't valuable. If there was clear, coherent, and compelling behavioral evidence for a bilingual advantage in inhibitory control, then it would be both pragmatically and theoretically beneficial to understand how (not just where) the neural implementation of conflict resolution differs for bilinguals compared to monolinguals. However, close examination of the two studies that reveal differences in neural processing between bilinguals and monolinguals also shows that the brain–behavior relationships are inconsistent with an interpretation that the identified neural differences are causing bilingual advantages in conflict resolution (viz. smaller interference effects).

Bialystok et al. (2005) used magnetoencephalography imaging (MEG) to investigate the relationship between brain and behavior when participants were engaged in a Simon task. There were two groups of bilinguals, French–English and Cantonese–English, and one group of English monolinguals. The Simon task included experimental blocks where congruent and incongruent trials were randomly presented and control blocks where the target was presented at fixation and there was never any conflict between the physical location of the target and the response required by the task rule. With respect to RTs there was a main effect of group (the Cantonese–English bilinguals were faster than the other two groups), but no Group × Trial Type interaction. With respect to the neuroimaging results all groups recruited similar areas, but there were group differences with respect to the specific areas associated with faster responding. Both bilingual groups showed substantial overlap in terms of the loci associated with fast responding (ACC, superior frontal, and inferior frontal regions) and these differ from the specific areas associated with fast responding for the monolinguals (middle frontal region).

Bialystok, Craik, et al. arrive at a cautious and fair conclusion: "*The evidence supports the interpretation that bilinguals perform the Simon task differently from monolinguals, even when they respond at the same speed, and that the group differences are evident for both congruent and incongruent trials*" (p. 48). Note that they do not claim that their neuroimaging results support a bilingual advantage in inhibitory control. They avoid an unjustifiable leap from differences between bilingual and monolinguals in neural processing to claims about differences in inhibitory control for the simple reason that no such behavioral differences in the magnitude of the interference effect were observed in their data. Furthermore, Bialystok, Craik, et al. offer no explanation for the global RT differences that were observed: "*We have no explanation for the faster reaction times of the Cantonese–English bilinguals, but due to the relatively small numbers of participants in each group this result could be due to sampling variability*", p. 46. This is a very prudent position given that the speed advantage was also apparent in the blocks of neutral (no conflict) trials, that the *n*'s were small (9 or 10 per group), and that if one of the Cantonese–English bilinguals had not been removed from the analyses because she was unusually slow the main effect may have reversed.

In contrast to the study by Bialystok, Craik, et al. reviewed above a more recent study by Luk, Anderson, Craik, Grady, & Bialystok (2010) does use neuroimaging data to bolster claims about bilingual advantages in conflict resolution; claims that are unjustified given the general arguments laid out in the beginning of this section. Luk et al. obtained fMRI images when participants were engaged in a flanker task. There were 10 English monolinguals and 10 English–other bilinguals. With respect to the RT data there was neither a main effect of group nor a significant Group × Trial Type interaction. There was a nonsignificant trend toward a global RT advantage for the bilinguals, but the groups were nearly identical in terms of their flanker effect.

Although the Simon and flanker tasks are often used interchangeably the two effects do not correlate with each other (see review in Paap & Greenberg, 2013). Given that fact, it is not completely surprising to discover that the two tasks differ with respect to how bilinguals neurally process congruent and incongruent trials. To review, Bailystok, Craik, et al. found the bilinguals and monolingual differ with respect to the loci of fast responding in the Simon task, but these group differences were the same for both types of trials. In contrast, Luk et al. reported that the loci associated with faster responding on congruent trials was the same for both monolinguals and bilinguals, but that faster responding on incongruent trials was associated with different areas (bilateral cerebellum, bilateral superior temporal gyri, left supramarginal gyri, bilateral post-central and bilateral precuneous), – but only for bilinguals. This additional and different pathway employed by bilinguals on incongruent flanker trials apparently led Luk et al. to an unjustified conclusion that bilinguals have superior inhibitory control: "….these results support the proposition that bilingualism influences cognitive control of inhibition…." (p. 356) and that "*differential engagement of this more extensive set of regions during incongruent trials in the two groups suggests that bilinguals can recruit this control network for interference suppression more effectively than monolinguals, consistent with their tendency to show less interference in terms of RT*" (p. 356).[2] The reference to showing "*less interference in terms of RT*" cannot of

[2] Hilchey & Klein point out that from a resource perspective the additional neural pathway employed by bilinguals on incongruent trials might lead to benefits on both types of trials, but this too is inconsistent with the pattern of behavioral differences across the three language groups tested by Bialystok, Craik, et al.

course refer to the concurrent behavioral performance because the flanker effect for the two groups was nonsignificant and nearly identical. The likely referent is the studies cited in their introduction that reported significant bilingual advantages in the magnitude of the Simon, Stroop, and flanker tasks. However, this advantage is properly hedged as a "tendency", at best, given the litany of null results reviewed by Hilchey and Klein (2011) and updated above.

Hilchey and Klein hold the same position as we do that the existence of behavioral phenomena must be adjudicated at the behavioral level. In reference to the two neuroimaging studies discussed above they pose the following: "*Moreover, the fact that there is no apparent behavioral advantage whatsoever for at least some bilingual groups in these previously described studies begs the question: How can we say, with any confidence, that these differences in neural circuitry underlie any bilingual advantage*" (p. 650).

Another way of thinking about the relationship between brain and behavior is to start where there is universal agreement, namely that behavior is caused by underlying neural activity. Thus, even if the faster responding of the Cantonese–English bilinguals is due to sampling variation (as Bialystok, Craik, et al. speculate) some unidentified neural circuits of this group of bilinguals must be more efficient than those in the group of French–English bilinguals (or English monolinguals). The fact that the neuroimaging methods used by Bialystok, Craik, et al. identify similar neural pathways for both groups of bilinguals means that these methods have failed to identify the neural cause of the behavioral difference that actually occurred. Differences between bilinguals and monolinguals in neural processing alone cannot support the hypothesis that there are bilingual advantages in conflict monitoring tasks such as Simon or Stroop. As a first step the neural differences must be linked to the behavioral differences of interest. At this point in time that link has not been forged as the behavioral differences reported by Bialystok, Craik, et al. for the three language groups mismatch the neural differences. Likewise, Luk et al. report neural differences between groups of bilinguals and monolinguals, but there are no behavioral differences. In summary, the current evidence linking performance in conflict monitoring tasks to neuroimaging differences does not, in our view, *drastically reinforce* the relatively rare reports of bilingual advantages in performance of the Simon and flanker tasks. In fact, there is a surprising disconnect between the brain differences and behavior differences that co-occur in the same experiment.

3. BICA in language processing

Paap and Greenberg (2013) discussed three possible reasons why the special experiences of bilinguals may not enhance EP and most relevant to Experiment 1 is the possibility that the inhibitory control exercised in language processing and comprehension is different from or encapsulated from the inhibitory control component in general EP. If that were the case then bilingual advantages may be more consistently obtained when suppression or inhibition is required during language processing.

4. Experiment 1: language group differences in homograph suppression

Gernsbacher, Varner, and Faust (1990) anticipated the role of individual differences in inhibitory control in language processing by hypothesizing that more skilled comprehenders build representations of a story by enhancing the activation of relevant information while suppressing the activation of less relevant information. This hypothesis was tested in their Experiment 4 with a task that required participants to read a sentence ending in either a homograph or control word (e.g., He dug with the *spade/shovel*.) and then judge if a test word (e.g., ACE) matched the meaning of the sentence they just read. Although the test word (e.g., ACE) is semantically related to one meaning of the homograph (*spade*) it is not the meaning supported by the sentence context (e.g., He dug with the…) and, consequently, this alternative meaning must be suppressed in order to correctly respond "no". Thus, a measure of homograph interference can be computed by comparing the mean RT for sentences ending in a homograph (e.g., *spade.* ACE) to those ending in a control word (e.g., *shovel* ACE). In order to examine the time course of homograph suppression Gernsbacher et al. presented the test word either immediately (ISI = 100 ms) or after a delay (ISI = 850 ms).

Gernsbacher et al.'s participants were partitioned into more-skilled (top third) and less-skilled (bottom third) comprehenders on the basis of their cumulative performance on a multi-media

comprehension battery consisting of six stories presented in written, auditory, and picture format. Both groups showed about the same amount (46 ms) of homograph interference when the test word was presented immediately, but the more skilled comprehenders showed minimal (7 ms) interference after the delay while the less skilled comprehenders showed no improvement at all. The authors conclude that less skilled comprehenders have less efficient, or perhaps simply less rapid suppression mechanisms.

Selecting the context appropriate meaning of an ambiguous homograph is both similar and dissimilar to the bilingual's task of selecting an appropriate word form (translation equivalent) when the conversational context has triggered a switch from one language to another. At a general level both situations require the language user to select and maintain the activation of context appropriate representations and to inhibit inappropriate representations. At a more specific level the type of inhibited representation differs. That is, in the homograph-interference task the inappropriate representation is conceptual (the alternative meaning of the homograph) whereas in the bilingual case the context-inappropriate translation equivalent has the correct meaning, but the incorrect form (e.g., phonological or orthographic form). If the bilingual's extensive experience in language switching only improves the specific abilities to switch languages and suppress translation equivalents then one would not expect bilinguals to show less homograph interference in the delayed condition. If, on the other hand, bilinguals have honed a general advantage in inhibitory control then they should show less homograph interference when the test word is delayed.

Gernsbacher et al. present a very general description of the cognitive architecture that results in high-skilled comprehenders having a greater ability to suppress the sentence-inappropriate meaning of a homograph, but Gernsbacher and Faust (1991) do provide more details. "… the mechanism of suppression dampens the activation of the less likely meanings… According to the structure building framework ambiguous words are accurately understood because the memory cells r*epresenting the semantic context*, the syntactic context, or other source of information transmit processing signals; these processing signals suppress the contextually inappropriate meanings" (p. 246). This is consistent with the processing architecture for the homograph-interference task shown in Fig. 1 where the memory cells representing the semantic context are represented as a propositional structure where the prior context has activated the *Dig* predicate, instantiated *He* in the agent slot, and established the expectation of an instrument (represented as X). Consistent with the quote above this representation of the prior context activates the appropriate meaning (tapered shovel) and inhibits the inappropriate meaning (card suit). The further assumption is that the relative degree of enhancement and suppression is modulated by individual differences in these two mechanisms. In summary, this hypothesis sounds like a precursor to the view that some individuals have superior executive attention. The

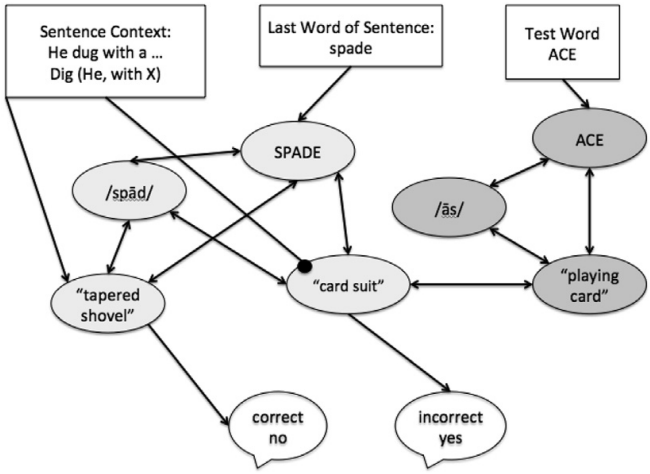


Fig. 1. A schematic of the cognitive architecture proposed by Gernsbacher et al. to explain how comprehension skill modulates the suppression of the sentence inappropriate meaning of a homograph. Note the presence of an inhibitory link (ending in a solid dot) from the sentence context to the inappropriate meaning ("card suit") and the absence of direct inhibitory links between the two alternative meanings of the homograph.

purpose of Experiment 1 is to determine if bilinguals can suppress the context inappropriate meaning of sentence-ending homographs more effectively than monolinguals in the delayed condition.

4.1.. Participants

Basic language characteristics for the participants in Experiment 1 are shown in Table 1 for both the complete pool of participants and the subset matched on overall task accuracy. Proficiency in a spoken language was self-rated using the 7-point scale described in Paap and Greenberg (2013) where a rating of 6 represents *Fluent: As good as a typical native speaker* and a rating of 7 represents *Super Fluency: Better than a typical native speaker.* Of the 34 bilinguals, 7 were simultaneous bilinguals (have 2 native languages), 4 acquired a language other than English as an L2, and 23 acquired English as an L2. Eight of the 34 participants classified as bilinguals actually spoke three or more languages. Like all our participants these bilinguals are upper-division psychology majors at a university where English is the language of instruction. Unlike most research universities in the United States, San Francisco State University is a commuter school where undergraduate students typically return every day to homes and communities where their non-English language is spoken. Note in Table 1 that the mean percentage of English currently used by our bilinguals is 71%. This reflects greater balance between languages than, for example, the 75% use of Catalan reported by Hernández, Martin, Barceló, and Costa (2013) for Spanish–Catalan bilinguals living in Barcelona. For the subset of our bilinguals who acquired English as an L2, the mean rated proficiency was exactly the same (6.0) for both English and their other language and a rating of 6 is "*as fluent as a typical native speak*er". A self-rating of reading comprehension in English was also obtained with a 5-point rating scale. The means were 3.8 and 4.0 for bilinguals and monolinguals, respectively, and the scale value of 4 was labeled: In comparison to other college students my ability to read and comprehend books written in English is somewhat above average. In summary, the participants classified as "bilingual" are fluent in at least two languages and actively use at least two languages. If managing two languages for many years enhances inhibitory control, then this group should by better at conflict resolution.

Some researchers are initially skeptical about the accuracy of self-ratings of language proficiency, but self-ratings are highly correlated with a range of objective and standardized measures of language proficiency. For example, a study by Marian, Blumenfeld, & Kaushanskaya (2007) correlated self-report measures of reading, speaking, and listening proficiency (obtained with the LEAP-Q questionnaire which is very similar to our scale) with eight different standardized measures of language skill involving reading, writing, speaking, and listening and covering both comprehension and production. These correlations were obtained for both L1 and L2 where L1 was defined as the language a bilingual acquired first. For L2 (the proficiency of greatest concern in classifying an individual as bilingual) all 24 correlations between the three subjective measures and the eight objective measures were significant with Pearson r values ranging from .29 to .74 with a mean of .59. Taking all of their results into account Marian et al. concluded that self ratings are "*an effective, efficient, valid, and reliable tool for assessing bilingual language status.*" p. 960.

Although our bilinguals are fluent in two languages, their rated proficiency in English ($M = 6.2$) is less than that of the monolinguals ($M = 6.7$), $t(84) = -4.16$, $p < .001$. This occurs because more than half of the monolinguals consider themselves to be "super fluent", that is, better than a typical native speaker. Experiment 1 also included an objective measure of category fluency (Gollan, Montoya, & Werner, 2002) as participants were asked to name as many instances in 1 min for the following categories: musical instruments, vegetables, and animals. The mean number of correct responses for both

Table 1
Language characteristics of monolinguals (M) and bilinguals (B) in Experiment 1: mean (SD).

Group	n	English Pro.	Other Pro.	English reading	English AoA	Other L AoA	% English use	Category fluency
All B	34	6.2 (0.7)	5.8 (1.2)	3.8 (0.7)	5.8 (5.4)	1.0 (3.4)	71.3 (18.9)	41.0 (1.7)
All M	55	6.7 (0.5)	0.7 (0.5)	4.0 (0.7)	0 (0.0)	10.4 (5.6)	100.0 (0.0)	49.2 (1.2)
Matched B	26	6.4 (0.6)	5.7 (1.2)	3.8 (0.7)	4.3 (4.6)	1.5 (4.1)	73.4 (18.6)	43.5 (1.9)
Matched M	32	6.6 (0.5)	0.1 (0.7)	3.8 (0.7)	0 (0.0)	11.3 (5.7)	100.0 (0.0)	46.9 (1.3)

Note. n = sample size; SD = standard deviation; Pro. = proficiency; AoA = age of acquisition.

groups is shown in Table 1. An independent samples *t*-test confirmed that the mean for monolinguals is greater than the mean for bilinguals, $t(84) = 3.92$, $p < .001$. Thus, the group differences in self-reported English proficiency correctly predict the group differences in the category fluency task. Bilingual disadvantages in category fluency and picture naming are commonplace even when bilinguals have been selected on the basis of an extensive battery of subjective and objective tests (Gollan et al., 2002; Gollan, Montoya, Fennema-Notestine, & Morris, 2005 are excellent examples.) The important point of this discussion is that it is anticipated that bilinguals will not be as fast or accurate as monolinguals in the immediate condition of the homograph-interference task even though their speaking and listening skills are at a high level.

4.2. Materials

A master set of 120 sentences and test words were constructed, each ending in a homograph (e.g., He dug with a *spade*. ACE). The relative strength of the context appropriate meaning of the homograph (tool meaning of spade) and the inappropriate meaning that was semantically-related to the test word (playing cards meaning of spade) were scored using the University of Alberta norms of relative meaning frequency (Twilley, Dixon, Taylor, & Clark, 1994). Because many homographs have more than two distinct meanings the proportions of these two meanings do not always sum to one. Across all 120 sentences the relative strength (often referred to as meaning frequency or dominance) of the inappropriate meaning was .58 and that for the appropriate meaning was .25. Three judges selected the 120 homographs from the pool of 566 because they were judged to have two distinct meanings that would both be known to fluent L2 speakers of English. Because the goal was to investigate if bilingualism modulates the interference effect it is important to produce a robust interference effect. To this end sentences were constructed to usually support the weaker of the two meanings and consequently the stronger (but sentence inappropriate) meaning would have to be suppressed in order to correctly respond "no". If the sentence context supports the dominant meaning, then the weaker meaning may not be activated. This concern is consistent with, for example, the report by Duffy, Morris, and Rayner (1988) that gaze durations on homographs were longer than on control words when the context supported the less common meaning, but not when it supported the dominant meaning.

Gernsbacher et al. state that their 80 homographs were selected with the constraint that at least two of its meanings were relatively equal in frequency. Their materials were not presented in the original article but do appear on Gernsbacher's web site. When the materials on the web site are coded using the Twilley et al. norms, the outcome showed that Gernsbacher's materials were somewhat less difficult in that the relative strength of the inappropriate (.39) and appropriate (.40) meanings were about the same rather than favoring the sentence-inappropriate meaning. Thus, all other factors being equal, the homograph-interference effect should be larger in our experiment.

4.3. Procedure

The procedure was patterned after that used by Gernsbacher et al. Each trial began with a centered warning signal (+) that appeared for 850 ms. Then each sentence was presented, one word at a time, in the center of the screen, with each successive word replacing the previous one. Each word's duration was a function of its number of characters (16.7 ms per character) plus a constant (300 ms). The blank interval between words was 150 ms. After the offset of the final word in each sentence a centered test word was presented for either 100 ms (*immediate*) or 850 ms (*delayed*) later. Each test word was presented in uppercase and flanked by asterisks (e.g., **ACE **). The test word remained on the screen until the participant responded by pressing the "/" key with the right index finger if the test word *matched the meaning of the sentence just read* and the "Z" key with the left index finger if it did not. A beep sounded if the response was wrong.

4.4. Design

There were eight conditions formed by the factorial combination of test words requiring either a "yes" or "no" response, two types of final words (homograph/control), and two levels of ISI (immediate/

delayed). Within each block the trials were randomized. A practice block of 24 trials included 3 tokens of each condition. The practice block was followed by three 80-trial blocks that included 10 tokens of each condition. Thus, collapsing across blocks each condition mean was based on 30 trials.

Gernsbacher considered the four "no" conditions to be critical and the "yes" conditions as filler material. The master list of materials included a non-homograph control word (He dug with a *shovel*. ACE) for each of the 120 homographs (He dug with a *spade*. ACE). No sentence-ending word (homograph or control) or test word was presented to a given participant more than once. This necessitated the formation of 4 lists of "no" materials so that each specific pair of sentences (and their presentation at either 100 ms or 850 ms ISI) could be rotated onto separate lists. Thus each individual participant saw 30 unique sentence-test word pairs in each of the four critical conditions and item variation was counterbalanced only when condition means were averaged across lists. The 120 filler items that required a "yes" response were the same for each participant.

4.5. Results

If there is a bilingual advantage in inhibitory control then, other factors equal, bilinguals should be less vulnerable to interference from the context-inappropriate meaning of a homograph, but only when the test is delayed and the sentence context has sufficient time to suppress the inappropriate meaning. The first analysis comparing the interference effect for bilinguals and monolinguals includes 55 monolinguals and 34 bilinguals. The condition means and SE for each group are shown in Table 2. The data were analyzed with a mixed ANOVA with last word (homograph versus control) and delay as within-subject factors and group as a between-subject factor. All three main effects were significant: homographs resulted in longer RTs (61 ms) than control words, $F(1,87) = 84.73$, $p < .001$, partial $\eta^2 = .49$; responses were faster (52 ms) after a delay compared to immediate, $F(1, 87) = 55.61$, $p < .001$, partial $\eta^2 = .39$; and monolinguals were faster (238 ms) than bilinguals, $F(1, 87) = 30.81$, $p < .001$, partial $\eta^2 = .26$. The Last Word × Group interaction was also significant, $F(1, 87) = 7.84$, $p = .006$, partial $\eta^2 = .08$: the homograph-interference effect is larger for bilinguals ($M = 79$ ms) than monolinguals ($M = 42$ ms).

Although neither the Group × Delay nor the Group × Delay × Last Word interaction were significant, inspection of Table 2 (and the left side of Fig. 2) shows that after a delay monolinguals have visibly reduced the amount of homograph interference whereas the bilinguals had not. These trends are consistent with a bilingual disadvantage in EF rather than an advantage. However, the comparison is complicated because the bilinguals experienced far more immediate interference than the monolinguals and also were less accurate: 88.3% correct (collapsed across both "yes" and "no" responses) for bilinguals compared to 93.5% for the monolinguals, $t(87) = -5.27$, $p < .001$. These differences in accuracy are consistent with the differences in self-ratings of both spoken language proficiency and reading comprehension in English.

The mean proportion correct for only the experimental ("no" response) trials were submitted to the same mixed ANOVA. There were significant main effects of the last word (controls 9.9% more accurate),

Table 2
Response-times (ms and z-transformed) and the homograph-interference effect for monolinguals and bilinguals and for immediate and delayed targets: mean (SE).

Group	*n*	Immediate targets			Delayed targets		
		Control	Homo.	Diff.	Control	Homo.	Diff.
All M	55	810 (29)	861 (28)	51	773 (25)	807 (27)	34
All B	34	1040 (37)	1121 (35)	81	983 (32)	1060 (35)	77
All Mz	55	−.02 (.02)	.26 (.03)	.28	−.19 (.03)	−.03 (.03)	.16
All Bz	34	−.04 (.03)	.28 (.04)	.32	−.21 (.03)	.04 (.03)	.25
Matched M	32	832 (34)	883 (35)	51	794 (29)	827 (33)	33
Matched B	26	943 (38)	1032 (39)	89	891 (32)	953 (36)	62
Matched Mz	32	−.01 (.03)	.25 (.04)	.27	−.16 (.04)	−.05 (.03)	.11
Matched Bz	26	−.04 (.03)	.32 (.04)	.36	−.22 (.04)	−.01 (.03)	.22

SE = standard error; *n* = sample size; Homo. = homograph; M = monolingual; B = bilingual; Mz = standardized monolingual RTs, Bz = standardized bilingual RTs; Diff. = homograph interference effect.

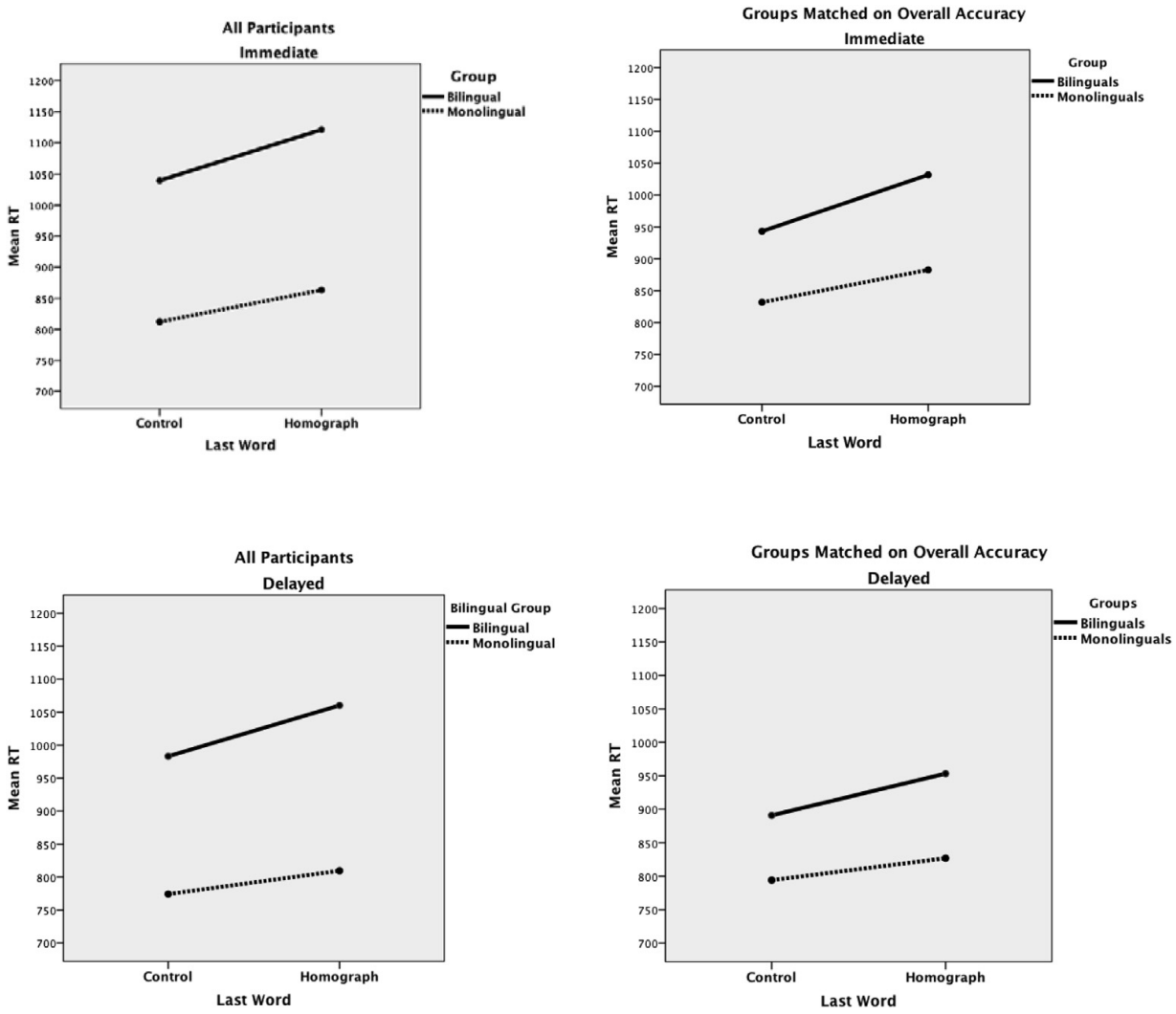


Fig. 2. Mean RT for controls and homographs with language groups as separate lines. The data from all participants are on the left and groups match on overall task accuracy are on the right. The top graphs show the means for the immediate presentation of the test word and the bottom graphs for the delayed test.

group (monolinguals 6.4% more accurate), and a significant Group × Last Word interaction, $F(1, 87) = 14.60$, $p < .001$, partial $\eta^2 = .14$. The significant interaction occurred because the homograph interference effect is 13.2% for the bilinguals compared to only 6.6% for the monolinguals.

The larger homograph-interference effect in RT obtained with bilinguals may, in part, reflect differences in confidence regarding the correctness of their decision. At the risk of testing the same hypothesis (viz., that there is a bilingual advantage in inhibitory control) multiple times, we explored three additional analyses that level the playing field for the bilingual group. The first was to standardize the RT scores for each individual participant based on the mean and standard deviation of all correct responses on both "yes" and "no" trials. The standardized means are shown in Table 2. When the same mixed ANOVA is performed on the standardized data the only significant interaction is the Last Word × Delay interaction, $F(1,87) = 5.62$, $p < 020$, partial $\eta^2 = .06$. The homograph interference effect observed in the immediate condition ($M = .30$ z) was significantly reduced after the delay ($M = .20$ z). This is exactly the pattern of interaction one would expect on the basis of the Gernsbacher et al. results assuming that our total set of participants was a mixture of *more skilled* and *less skilled* comprehenders.

The primary purpose of the *z*-score analysis was to test the hypothesis that bilinguals could reduce the size of their standardized interference effect (more so than monolinguals) when there was a delay and inhibitory control had more time to act. The left panel of Fig. 3 shows the mean standardized RT scores for the controls and homographs for both groups in the immediate (top) and delayed (bottom) condition. Not only is the Group × Last Word × Delay interaction far from statistically significant,

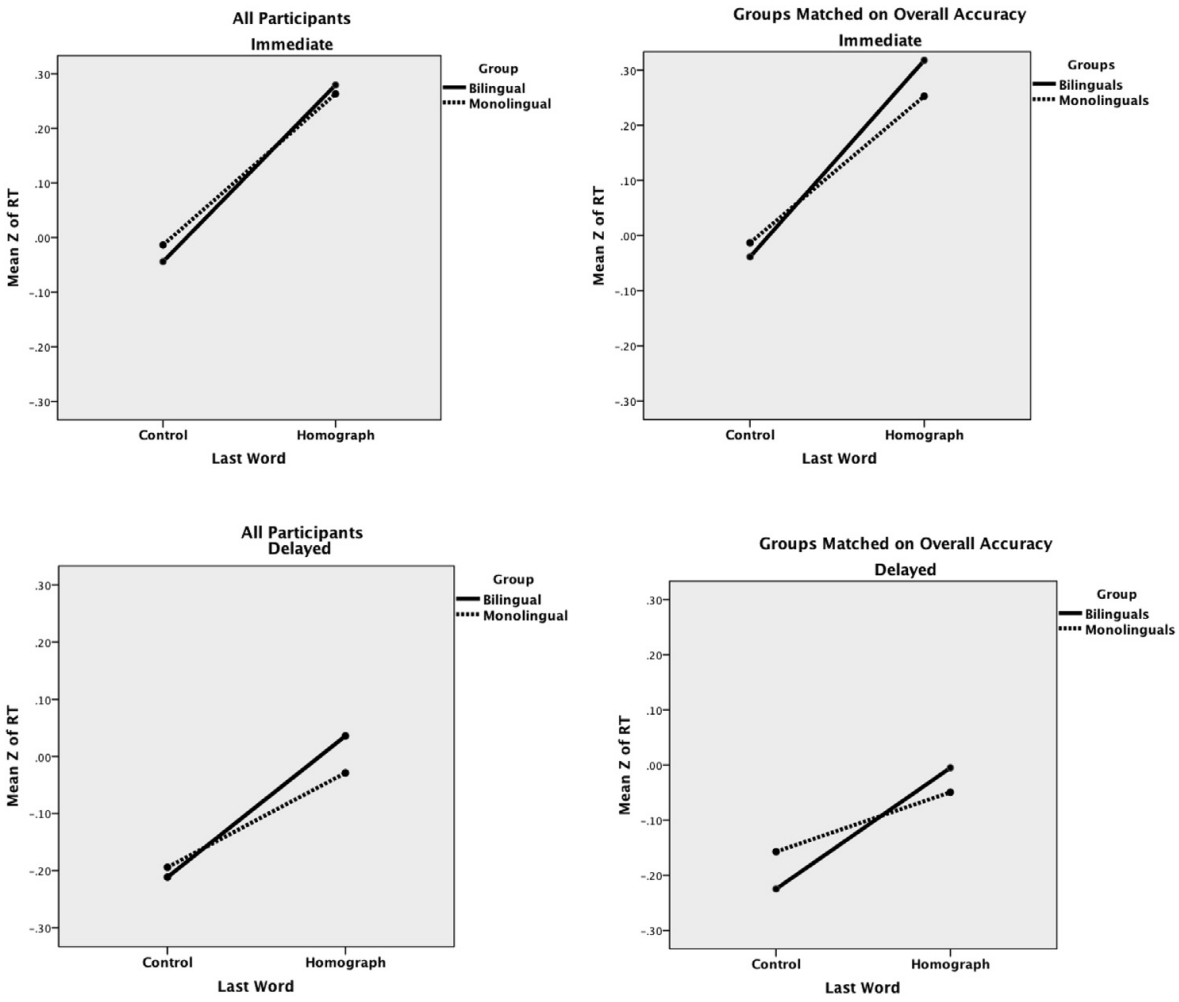


Fig. 3. Mean of *z*-transformed RTs for control and homographs with language group as separate lines. The data from all participants are on the left and groups matched on task accuracy are on the right. The top graphs show the means for the immediate presentation of the test word and the bottom graphs for the delayed test.

$F(1,87) = 0.27$, $p = .603$, but inspection of Fig. 3 shows that the trend favors the monolinguals. That is, both groups have nearly identical slopes when the test is immediate, but it is the monolingual slope that slightly decreases with additional time to suppress the context inappropriate meaning.

Another supplementary test for a bilingual advantage involved selecting subsets of participants who performed at the same level of overall accuracy in the task. To this end bilinguals who made fewer than 200 correct responses (83% correct) or monolinguals who made more than 230 correct responses (96% correct) were pruned from the pool of participants. These cutoffs resulted in a reduced pool of 26 bilinguals (91.0% correct) and 32 monolinguals (91.4% correct) with nearly identical accuracy levels. The same mixed ANOVA was run on the groups matched on overall accuracy. The main effects of last word, $F(1, 56) = 54.89$, $p < .001$, partial $\eta^2 = .50$; delay, $F(1, 56) = 39.84$, $p < .001$, partial $\eta^2 = .42$; and group, $F(1, 56) = 6.58$, $p = .013$, partial $\eta^2 = .10$ were all significant. Homographs took longer than control words (a 59 ms homograph-interference effect), responses were faster by 56 ms after a delay, and monolinguals were faster by 122 ms. Note that the matching on accuracy did not eliminate the statistically significant differences in group RT, but it did reduce the partial η^2 from .26 (238 ms) to .10 (122 ms).

The Group × Last Word interaction was significant, $F(1, 56) = 4.25$, $p = .038$, partial $\eta^2 = .07$ because the homograph-interference effect is larger for bilinguals ($M = 76$ ms) than monolinguals ($M = 42$ ms). The main purpose of matching the groups on overall task accuracy was to use a more sensitive test of the Group × Last Word × Delay interaction that might reveal more effective inhibitory control by the bilinguals when there is more time for suppression to act. The critical Group × Last Word × Delay

interaction was not significant, $F(1, 56) = 0.11$, $p = .741$ and the right panel of Fig. 2 shows that when the groups are matched on overall accuracy the slopes for bilinguals and monolinguals are very similar both with immediate (top) and delayed (bottom) presentation of the test word.

Given that matching on overall task accuracy did not completely eliminate the main effect of group on RTs, in a final test we combined both matching and standardization by analyzing the z transforms for the matched groups. The main effects of last word, $F(1, 56) = 66.85$, $p < .001$, partial $\eta^2 = .54$ and delay, $F(1, 56) = 53.17$, $p < .001$, partial $\eta^2 = .49$ were significant. The only significant interaction was the Last Word × Delay interaction, $F(1,56) = 9.47$, $p = .003$, partial $\eta^2 = .15$. The interaction occurred because the homograph interference effect was greater immediately (.31 z) than after a delay (.16 z). The critical Group × Last Word × Delay interaction was not significant, $F(1,56) = .05$, $p = .824$. This potential interaction is shown on the right side of Fig. 3. The nonsignificant trend is for monolinguals to show less homograph interference (more shallow slope) after a delay.

4.6. Discussion of the homograph-interference effect

The results of the homograph-interference task showed that when all participants were included in an analysis of mean RTs, bilinguals showed more homograph interference compared to monolinguals and this is true with both immediate and delayed presentation of the test word. This pattern holds when individual RT scores are standardized. When participants are matched on overall task accuracy the global RT advantage favoring the monolinguals is substantially reduced, but not eliminated. The primary difference between the full set and matched set is that for both groups in the matched set the amount of interference significantly declines when there is more time for the sentence context to inhibit the inappropriate meaning of the homograph. This is true for both mean RTs and the z-transformed RTs. Most important, the three-way interaction was not significant and this is consistent with the conclusion that both groups were more-or-less equally adept at suppressing some of the initial activation of the inappropriate meaning during the delay interval. The fact that both groups of participants showed a significant decrease in the amount of homograph interference when the test word was delayed, but still showed substantial interference in the delayed condition suggests that both groups are not as good at suppressing the inappropriate meaning as Gernsbacher et al.'s high-skill group, but they are better than Gernsbacher et al.'s low-skill group who showed no decline.

It is worthwhile to consider why bilinguals show more immediate homograph interference compared to monolinguals and what implications this has for the hypothesis that the special experience of managing two languages induces a domain-independent superiority in inhibitory control. A compelling account is based on *the lexical quality hypothesis* (Hart & Perfetti, 2008; Perfetti & Hart, 2002). The lexical quality hypothesis proposes that experience with a word determines the quality of its lexical representation and that higher quality representations are activated more rapidly and result in better reading comprehension. The lexical representation is assumed to consist of orthographic, phonological, and semantic constituents. Highly skilled readers have high quality (coherent) lexical representations that are marked by having fully specified codes for each constituent and strong links between them. Thus, highly skilled readers quickly and efficiently activate lexical representations and this facilitates the construction of sentence- and text-level representations. A further critical assumption is that lexical quality increases with specific experience. More voracious readers encounter words more often and in more varied contexts and, consequently, hone the quality of their lexical representations and become highly-skilled readers. Thus, other factors the same, the more often readers traverse the orthography to meaning pathway the more quickly and precisely they can recognize words and activate their precise meanings.

The lexical quality hypothesis has clear implications for bilingualism because bilinguals divide their language use between two lexicons and, in the simplifying case of a balanced bilingual who reads both languages equally often, the quality of the lexical representations will be strengthened on the basis of only half the number of encounters. Note that even if fluent bilinguals read English far more often than their other language (as many of our student bilinguals do) the lexical quality hypothesis still puts them at a disadvantage because both listening to English and reading English facilitates the acquisition of precise meanings. For example, both listening and speaking experiences contribute to the meaning of low-frequency English words like grits or gumbo. Because our bilinguals speak English only part of the

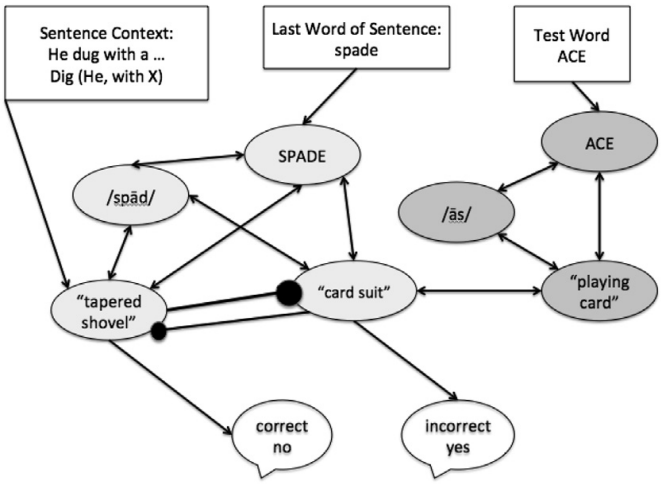


Fig. 4. A schematic of the cognitive architecture proposed by Hart and Perfetti to explain how experience creates and strengthens inhibitory links (ending in a solid dot) between the two meanings of a homograph. Note the absence of an inhibitory connection from sentence context to the sentence inappropriate meaning.

time they have less relevant experience than a monolingual in building complete semantic representations. Another aspect of bilingualism that degrades lexical quality are false cognates and, in milder form, concepts that are translation equivalents (e.g., *rice* in English versus *arroz* in Spanish) but differ in their precise meanings because each connects to somewhat different constellations of semantic features. In summary, the lexical quality hypothesis provides an explanation for why our monolinguals had faster RTs than bilinguals on all types of sentences.

Furthermore, the lexical quality hypothesis can also provide an explanation for the Group × Last Word interaction. A key to this explanation is that Hart and Perfetti (Fig. 4) assume a different architecture for the resolution of lexical ambiguity than Gernsbacher et al. As shown in Fig. 1 Gernsbacher et al. assume that the sentence inappropriate meaning of a homograph is directly suppressed by the expectations established by the preceding sentence context. In contrast, for Hart and Perfetti (cf. Fig. 2 is an elaboration of their Fig. 6.2) the sentence inappropriate meaning is indirectly inhibited because the preceding context fortifies the appropriate meaning which, in turn, directly inhibits the alternative meaning of the homograph.[3]

A key question is when and how are these inhibitory connections between the alternative meanings of a homograph acquired? In most models of semantic memory related concepts are assumed to have excitatory connections in proportion to the degree of shared semantic features. For example, *tiger* and *lion* would have a strong excitatory connection, *cat* and *tiger* a more moderate weight, and *brick* and *tiger* no connection or a connection with a zero weight. Why should *tapered spade* and *card suit* (Fig. 2) have an inhibitory connection rather than simply a zero weight? If these inhibitory links are learned, as Perfetti and Hart assume, it must be triggered by the discovery that the same word form (e.g. *SPADE*) maps to two distinct meanings and that on any specific encounter of the word form the intended meaning is one of the two alternatives. From that trigger point each encounter of the dominant meaning increments the inhibitory weight of the subordinate meaning and vice versa. In summary, inhibitory connections between meaning concepts are restricted to homographs and are acquired after the establishment of excitatory connections between the word form (e.g., *SPADE*) and the two alternative meanings (e.g., *tapered shovel* and *card suit*). The development of these inhibitory links should be particularly compromised in bilinguals because words that are

[3] These inhibitory links between the two meanings of a homograph are explicitly denied by Gernsbacher & Faust: "Some theories assume that the inappropriate meanings of ambiguous words become less activated in other ways. For instance, *according to some theories, the inappropriate meanings are inhibited by the appropriate meanings* (McClelland & Kawamoto, 1986; Waltz & Pollack, 1985), and according to others the inappropriate meanings simply decay (Anderson, 1983). Unfortunately, neither assumption is strongly supported by empirical data (Gernsbacher & Faust, 1991) p. 246."

homographs in one language will either be unambiguous in the other or, if ambiguous, will map to a different alternative meaning. The upshot is that the inhibitory links between the alternative meanings are likely to be weaker in bilinguals and this will result in a greater homograph-interference effect for bilinguals compared to monolinguals.

In summary, the lexical-quality hypothesis provides not only an account of why monolinguals are faster than bilinguals overall, but also an account of why the homograph interference-effect would be greater for bilinguals. According to this account individual differences between low-skill and high-skill readers or between bilinguals and monolinguals are experience based, that is, simply determined by the frequency of use and not due to differences in executive attention or a general ability to suppress irrelevant information. This does not mean that a general suppression ability does not also modulate the degree to which the sentence context inhibits the inappropriate meaning of a homograph. In a recent review of ambiguity resolution within- and between languages Degani and Tokowicz (2010) agree that "…the two classes of individual differences accounts are not mutually exclusive. The two sources (cognitive ability and experience/proficiency) may work in parallel or may interact." p. 1287. Hart and Perfetti suggest that their lexical quality hypothesis is more "… *parsimonious in that it does not require the invoking of an additional suppression mechanism*" p. 115. In fairness one should note that Gernsbacher and Varner propose a different inhibitory mechanism (from sentence context to the sentence inappropriate meaning of the homograph) not an additional mechanism as they explicitly reject the assumption of direct inhibitory links between the two meanings. Whatever the resolution, the results of our homograph-interference task provide absolutely no support for the hypothesis that there is a bilingual advantage in executive attention that enhances the sentence appropriate meaning and inhibits the inappropriate meaning. The bilingual's ubiquitous experiences in language switching that leads to the activation of one lexicon and the inhibition of the other appears to offer no positive transfer to the within-language task of activating one meaning of a homograph and inhibiting the other.

5. Experiment 2: replication and extension of Moreno, S. et al.'s behavioral study

The conclusion from our homograph interference experiment that there is no evidence for a bilingual advantage in inhibitory control during sentence processing is contradicted in a recent article by Moreno, Bialystok, Wodniecka, and Alain (2010). They concluded that their study provides evidence that bilinguals have enhanced executive control because bilinguals close the performance gap when a sentence judgment task involves conflict resolution. This is explained more fully below.

The Moreno, S. et al. study involves both ERP and behavioral data and is quite complex in both design and interpretation. Bilinguals and English-speaking monolinguals participated in both an acceptability and grammaticality task that includes four sentence types: (1) correct in both semantics and syntax, (2) semantically anomalous, but syntactically correct, (3) semantically fine, but with a syntax violation, and (4) contain both a semantic anomaly and a syntax violation. In the acceptability task participants press one key if there is nothing wrong with the sentence (Type 1 sentences) and a different key if there is anything wrong (Types 2, 3, and 4). In the grammaticality task participants press one key if there is a syntax violation (Type 3 or 4 sentences) and the other key if the syntax is correct (Types 1 or 2). For both tasks the participant does not respond until a prompt is displayed 1450 ms after the last word of the sentence. Because of the long delay the authors assume that accuracy is meaningful, but that RT is not and, consequently, RTs are not reported.

Moreno, S. et al. report separate ANOVAs for each of the two tasks. The main effect of language group favoring the monolinguals (+4.3%) is significant in the acceptability task but non-significant (+2.9%) in the grammaticality task. Moreno et al. concluded that the bilinguals superior EP enabled them to overcome the monolingual advantage shown in the easy acceptability task by employing their superior inhibitory control in the more difficult and EP demanding grammaticality task. Given that the monolingual advantage in the grammaticality task is 67% of that obtained in the acceptability task it is quite likely that these group differences across tasks do not differ from one another and that a ANOVA including task as a third factor would not show a significant Group × Task interaction. The strong possibility that the Group × Task interaction is not significant will be tested in our Experiment 2, a replication of the behavioral component of Moreno et al.

Moreno, S. et al. critically assume that the grammaticality task is more difficult than the acceptability task because the grammaticality task requires substantial amounts of EP. There is no empirical evidence for this critical assumption. In fact, accuracy is higher in the grammaticality task and this is, of course, consistent with the opposing assumption that the grammaticality task is easier than the acceptability task. If the grammaticality task is easy this explains why both groups are pressed against a performance ceiling (about 95% correct) that reduces the size of the monolingual advantage. Furthermore, if it is actually the acceptability task that is more difficult, then the monolinguals' superior English proficiency becomes apparent because accuracy is no longer near the performance ceiling. We offer this as a clearly plausible and more parsimonious alternative explanation of the smaller monolingual advantage obtained in the grammaticality task.

Moreno, S. et al.'s assumption that the grammaticality task is more difficult and requires more executive attention stems mostly from consideration of the sentences with semantic anomalies (e.g., "*A new computer will paint for many years.*") and syntactic violations (e.g., "*A new computer will lasting for many years*"). Given the instructions in the grammaticality task, participants must select the action plan consistent with the syntax and suppress the opposing plan activated by the semantics. In contrast to the grammaticality task, Moreno, S. et al. reason that in the acceptability task any global problem with the test sentence can be used to select the "no" response and there is no need to identify and keep separate the two types of errors.

This is a plausible line of reasoning and might be correct if participants in the grammaticality task are really forced to monitor for a variety of syntactic violations while holding the outcome of the semantic analysis at bay. However, both the accuracy data and the way in which the sentences were constructed suggest that evaluating the syntax was easier than evaluating the semantics. This is evident in the acceptability task where sentences with semantic anomalies are the most difficult for both groups (81% and 87% correct for bilinguals and monolinguals, respectively) compared to sentences with a syntax violation (93% and 95% correct for bilinguals and monolinguals, respectively). Why are the syntax violations so easy to detect, even for the less proficient bilinguals? The syntactically incorrect version always involved a modal verb followed by a present participle (-ing) form of the verb. Thus, 60 of the 120 test sentences contained exactly this type of verb tense violation. Given that the present participle comes in a predictable location and in a single form (-ing) participants may quickly learn to check for this specific violation. The semantic anomalies involved an unsuitable pairing of agents and actions, but they are varied in form and open to acceptable metaphorical interpretations (e.g., *…computer will paint…sea lions can edit…*).

To summarize, the accuracy data supports the conclusion that monitoring for a syntax violation was easier than monitoring for a semantic anomaly and that this, in turn, made the grammaticality task easy compared to the acceptability task where the correct response often requires taking into account a full semantic analysis of the sentence. We will directly test Moreno et al.'s assumption that the grammaticality task is more difficult and requires more EP in Experiment 2 by including latency as well as accuracy measures.

The predictions for both the behavioral and ERP measures critically compare the magnitude of language-group differences observed in one task with those obtained in the other task. Consequently, it is unfortunate that the grammaticality task was always run first and that task and task order were confounded. Although there could be practice effects favoring the task that comes second (viz., the acceptability task), there are two reasons for expecting poorer performance and greater difficulty in integrating the critical word into the preceding context. The first is a potential "surprise" effect based on the preceding context (e.g., "The sea lions can…") and memory for the way the stem was completed on the first task (e.g., "…edit on the beach all day."). Given the counterbalancing of the sentences across lists and tasks, the stem will be completed with a different critical word (e.g., "…bask on the beach all day."). If participants remembered what the sea lions were doing before, they will be surprised to read what they are up to now. That is, participants may experience more disconfirmed expectancies in the acceptability task (always second) compared to the grammaticality task.

A second potential source of negative transfer is the change in the decision rule for the anomalous sentences from "OK" in the initial grammaticality task to "not OK" in the acceptability task. This leads to the distinct possibly that it is the anomalous sentences in the acceptability task that most require suppression of a conflicting action plan. Thus, it is not possible to determine whether intrinsic task

differences or task order is responsible for differences in the magnitude of the monolingual advantage across the two tasks. Experiment 2 counterbalances the order of the two tasks, eliminating the confounding and enabling one to investigate the effects of order.

Another modification introduced in Experiment 2 was to use self-paced reading (SPR) rather than rapid serial visual presentation (RSVP). RSVP is often used in ERP studies of sentence processing because it ensures that all early components to upcoming words occur at a regular interval. However, as Ditman, Holcomb, and Kuperberg (2007) point out there are also disadvantages to RSVP. For example, RSVP is not a natural way to read and may distort the processing that occurs during normal reading for comprehension. Independent of naturalness, it has been repeatedly demonstrated that faster presentations are associated with an enhancement of lexical factors and slower rates with increases in sentence-level and discourse-level factors (Ledoux, Gordon, Camblin, & Swaab, 2007; Swaab, Camblin, & Gordon, 2004). Another disadvantage of RSVP is that it forces all participants to read at the same rate and does not allow for adjustments due to general levels of reading skill or for local perturbations in the ease of semantic or syntactic integration.

Although SPR is not a perfect panacea for these problems it is a good step in the right direction. A very important contribution of the Ditman et al. study is their demonstration that the motor response required to advance to the next word and the varying onset asynchronies between words did not alter the typical pattern of N400 and P600 responses to sentences that were correct compared to those containing either a semantic or syntactic violation. Another advantage of SPR is that one can explore the relationship between ERP components to the critical word and the processing load it creates as reflected in reading time (i.e., the average duration of the critical word or to words farther down steam). Ditman et al. observe that the ERP data confirm that readers immediately detect both syntactic and semantic violations as they occur at the critical-word location, but that the reading times show that only the syntactic violations induce an immediate processing load at the critical word. In contrast, the processing load induced by a semantic anomaly did not show up as longer reading times until the sentence final word.

5.1. Participants

Twenty-four English monolingual and 24 bilingual San Francisco State University students were recruited in order to fulfill a course assignment or receive extra credit. Table 3 shows the means and standard deviations for the two groups on several different language characteristics. Proficiency in English and any other language spoken was self-rated using the same 7-point scale described earlier. The mean for both groups is about 6.5. A rating of 7 represents "*Super Fluency: Better than a typical native speaker.*" About half our student participants, in each language group, rate themselves as super fluent in English. The mean rating of our bilingual group in their other (non-English) language is somewhat lower than their self rating in English, but the mean of 5.9 is very near the modal and median rating of 6 which represents "*as good as a typical native speaker.*" As a group our bilinguals are highly fluent in at least two languages. Seven have two native languages (with English being one of the two), five bilinguals have English as a native language, and for the remaining 12 English is an L2. The median age-of-acquisition for the bilinguals who acquired an L2 was 4.5 years of age. In addition to English our bilingual group included fluent speakers of Spanish (13), Japanese (2), Arabic (2), French, German, Dutch, Bosnian, Turkish, Farsi, Tungan, Punjabi, Hindi, Urdu, and Cantonese.

A self-rating of reading comprehension in English was obtained with a 5-point rating scale. The modal and median rating for both language groups is a 4 and this scale value was labeled: "*In comparison to other college students my ability to read and comprehend books written in English is somewhat above average.*"

Table 3
Language characteristics of participants in Experiment 2.

Group	English Pro.	Other Pro	English reading	English AoA	Other L AoA	% english use	Switch frequency
BiL	6.5 (0.7)	5.9 (1.2)	3.9 (0.7)	3.7 (0.1)	1.4 (3.7)	70.8 (20.2)	3.5 (1.2)
MonoL	6.6 (0.5)	0.8 (1.0)	4.0 (1.0)	0 (0.0)	11.7 (6.0)	99.6 (1.4)	0.2 (0.6)

Note. Pro. = proficiency; AoA = age of acquisition; L = language; BiL = bilingual; MonoL = monolingual.

Table 4 shows the means and standard deviations for the two language groups on six characteristics that are not related to language, but that may influence task performance. These include a short-form computerized version of the Ravens' test of nonverbal intelligence (described in Paap & Greenberg, 2013), the level of education of the participant's most highly educated parent (PED), and age. The measure Frequency Multitasking is a composite of responses to four items from our background questionnaire that tap into the individual's multitasking experiences. Each item requires a response on a 6-point scale with higher scale values associated with greater frequencies of playing computer games, exercising while listening to music, doing homework while listening to music, texting, conversing, etc. The composite Multitasking Frequency measure has a minimum value of 6 (responses of 1 to all 4 items) and a maximum of 24 (responses of 6 to all 4 items). Another characteristic shown in Table 4 is a self-rating on a 5-point scale of the degree to which the individual excels at team sports. The final characteristic assesses the individual's attitude toward multitasking rather than the frequency of actual behaviors. The only significant difference between the two groups is that bilingual's express a more positive attitude toward multitasking, $t(46) = 2.42$, $p = .02$. This item reads: "*How do you feel when you need to focus on an important task but there are lots of things going on that could be distracting? I find these situations:*" The modal and median responses of the bilinguals align with "*neither frustrating nor stimulating*" whereas the mode and median for the monolinguals is "*somewhat frustrating and my performance is sometimes not as good as it could be.*"

5.2. Materials

The same sentences used by Moreno, S. et al. were used in Experiment 1. The sentence frames are available in the appendix to Osterhout and Nicol (1999). Participants read, in random order, 30 sentences of each of the four types. Like Moreno, S. et al., we consider the sentences with both syntactic and semantics errors to be fillers. The 90 sentence frames were partitioned into three sets of 30 and then rotated across the participants into the three experimental conditions: correct sentences, sentences with a syntax violation, and sentences with a semantic anomaly. Thus, within a single task each sentence frame was used only once.

The same sentence frames were used in the second task. Consequently, each sentence frame was repeated in the second task with a different critical word that changed its type (e.g., replacing *edit* with *bask* changed the sentence from one that had a semantic anomaly to one that was completely correct). The repetition of the same sentence frame across tasks may lead to negative transfer effects. To investigate the transfer question, we also counterbalanced the order of tasks rather than having all participants complete the grammaticality task first.

Controlling for item specific effects required the formation of three lists so that a sentence frame that was correct in list 1 had a semantic anomaly in list 2 and a syntax violation in list 3. Because task order was also counterbalanced, 12 participants were required to cover all combinations of lists and task order. Consequently the 24 participants in each language group represent two complete cycles of counterbalancing.

5.3. Design and procedure

The sequence of events for each trial was the same in both the grammaticality and acceptability task except for the final prompt. Each trial was initiated with a fixation point (+) in the center of the screen. Self-paced reading was enabled and the first word (and each successive word) was presented when the participant pressed the space bar with their right hand. Each new word was centered and replaced the

Table 4
Other characteristics of bilinguals and monolinguals in Experiment 2: mean (SE).

Group	Raven's	PED	Age	Frequency multitasking	Excel team sports	Attitude multitasking
Bilingual	8.4 (2.2)	4.1 (1.6)	24.0 (6.5)	16.0 (3.9)	2.6 (1.1)	2.6 (1.0)
Monolingual	9.2 (1.5)	4.4 (1.4)	23.0 (4.2)	14.0 (3.8)	2.5 (0.9)	1.9 (0.9)

Note. PED = parent's educational level.

previous word. There was a 700 ms ISI between the sentence-final word (that include the full stop) and the prompt. (This contrasts with the 1450 ms ISI used by Moreno, S. et al.) For the grammaticality task the prompt was "*Grammar OK?*" For the acceptability task the prompt was "*Sentence OK?*". To answer the question participants used two fingers of their left hand to press either the *Yes* button (the Z key) or *No* button (X key).

A set of 32 practice sentences (8 of each type) was randomly ordered and presented as the first block of each task. This was followed by an experimental block of 120 randomly ordered trials that included 30 sentences of each type. Upon completion of the first task, the Ravens task was administered. The second sentence judgment task immediately followed the Ravens task.

5.4. Accuracy results

The only behavioral measure reported by Moreno, S. et al. was the proportion correct to the prompt following each test sentence. In review, in separate ANOVAs on each task, they reported a significant bilingual advantage in the acceptability task and a non-significant advantage in the grammaticality task. Because there are plausible reasons for expecting asymmetrical negative transfer between the two tasks it is prudent to start with an analysis of the subset of participants who received the tasks in the same order that Moreno, S., et al. used (viz., grammaticality first). The condition means for the Moreno, S. et al. experiment are shown in the first row of Table 5 and those for our subset using the same task order are shown in the second row. The condition means in our Experiment 1 show the same pattern as there is a 3.4% monolingual advantage in the acceptability task compared to 2.6% advantage in the grammaticality task. Our monolingual advantage (3.4%) in the acceptability task is slightly smaller than that reported by Moreno et al. (4.3%) and falls just short of a conventional alpha level of .05, $F(1, 22) = 3.90$, $p = .061$, partial $\eta^2 = .151$. Like Moreno, S. et al., the smaller bilingual advantage in the grammaticality task (2.6%) is not significant, $F(1, 22) = 2.66$, $p = .117$, partial $\eta^2 = .108$. In summary, our accuracy data comes very close to replicating the statistical pattern reported by Moreno, S. et al.

A key question we posed of the Moreno, S. et al. findings is whether the Group × Task interaction is significant when the task factor is included in the ANOVA. When we include all three factors in the analysis of our data the Group × Task and the Group × Task × Sentence interactions show no evidence of interaction with exact p's $> .75$. In summary, there is no statistical evidence in the appropriate analysis of our accuracy data that bilinguals "catch up" in the grammaticality task. This is, very fundamentally, our most important result as it undermines the only reason for believing that Moreno et al. have any evidence for a bilingual advantage in inhibitory control.

The bottom row of Table 5 shows the condition means for the subset of subjects that did the acceptability task first. The most important outcome here is that the reverse order shows identical monolingual advantages of 6.6% for both tasks! The bilinguals do not "catch up" or "close the gap" in the grammaticality task. Thus, our accuracy data does not support the hypothesis that bilinguals enjoy a relative benefit compared to monolinguals in the grammaticality task because of their enhanced EP skills.

A full factorial analysis of language group, sentence type, task, and task order was also conducted. All of the main effects were significant: monolinguals ($M = 95.3\%$) were more accurate than bilinguals ($M = 90.3\%$), $F(1, 44) = 20.51$, $p = .001$, partial $\eta^2 = .318$; semantic anomalies ($M = 87.7\%$ correct) were more error prone than correct sentences ($M = 94.7\%$) or those with syntactic violations (95.9%), $F(2, 88) = 27.39$, $p < .001$, partial $\eta^2 = .384$; the grammaticality task ($M = 93.8\%$) was easier than the acceptability task ($M = 91.7\%$), $F(1, 44) = 4.60$, $p = .037$, partial $\eta^2 = .095$; and the order used by Moreno,

Table 5
Mean proportion correct to final question for bilinguals and monolinguals in the grammaticality and acceptability tasks.

Data source	Grammaticality task			Acceptability task		
	B	M	Diff.	B	M	Diff.
Moreno et al.	.940	.960	.020	.880	.923	.043
Same order	.946	.972	.026	.914	.948	.034
Reverse order	.881	.947	.066	.870	.936	.066

Note. B = bilingual; M = monolingual; Diff. = M − B.

S. et al. ($M = 94.7\%$) was easier than the reverse order ($M = 90.8\%$) when the acceptability task was done first, $F(1, 44) = 9.81$, $p = .003$, partial $\eta^2 = .182$. Critical to the question of the hypothesized interaction between groups (due to an assumed bilingual advantage in EP) and tasks (due to assumed greater demands of EP in the grammaticality task), the Group × Task interaction and the three higher-order interactions that include both Group and Task are all non-significant, p's $> .05$. In summary, there is no evidence for a bilingual advantage in conflict resolution during sentence processing from the accuracy data in Experiment 1.

5.5. Reaction time results

Given that Moreno et al. did not report response times to the prompt, only the results of the full factorial analysis including all four factors (group, sentence type, task, and task order) will be reported. Interestingly there was no main effect of group indicating that although the bilinguals were less accurate in responding to the prompt they were not significantly slower, $F(1, 44) = 0.299$, $p = .588$, partial $\eta^2 = .007$. Nor did group enter into any significant interactions, all p's $> .05$. Thus, the RT data provide no evidence that monolinguals and bilinguals differ in the speed of making judgments about the grammaticality or acceptability of English sentences. The main effect of task was significant, $F(1, 44) = 7.629$, $p = .008$, partial $\eta^2 = .148$ showing the mean RT in the grammaticality task (629 ms) was faster than the mean in the acceptability task (681 ms). This is consistent with the task differences in accuracy and with our alternative hypothesis that, in general, it is the acceptability task (not the grammaticality task) that is more difficult.

The absence of any group effects cannot be attributed to the lack of sensitivity of the RT measure because there were highly significant main effects of both sentence type and task and two significant two-way interactions. These interactions are not germane to the primary purpose of examining differences in sentence processing between bilinguals and monolinguals as these interactions do not involve the language group factor, but they do merit a brief description.

The Task × Sentence interaction, $F(2, 88) = 12.697$, $p < .001$, partial $\eta^2 = .224$, is shown in Fig. 5. In both tasks responses to semantic anomalies are slow and responses to syntactic violations are fast. However, responses to the intact sentences differ across the two tasks. In the grammaticality task where readers can solely focus on detecting the fairly obvious syntactic violations, responses to correct responses are fast, just as fast as responding "no" to the syntactic violations. In contrast, the acceptability task requires readers to also monitor for semantic as well as syntactic errors. Resolution of possible anomalies within the correct sentences apparently spills over the end of the sentence and significantly slows responses to the "Sentence OK?" prompt in the acceptability task.

The Task × Order interaction was also significant, $F(1, 44) = 23.186$, $p < .001$, partial $\eta^2 = .345$. Inspection of Fig. 5 (right side) shows that response times in the two tasks are about the same when the

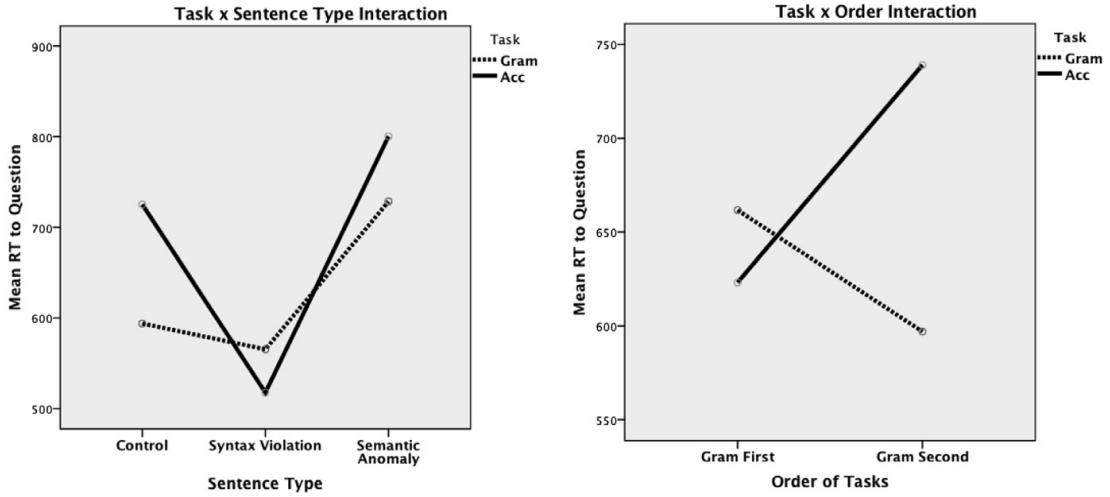


Fig. 5. Mean RTs to the final question in the grammaticality and acceptability tasks. The significant Task × Sentence Type interaction is shown on the left and the significant Task × Order interaction is on the right.

grammaticality task is first, but that the acceptability judgment takes substantially longer than the grammaticality judgment when acceptability is the first task. Having a chance to focus exclusively on the grammatical judgment when it is the first task may make it easier to step into the more difficult task that requires one to also monitor for the subtle semantic anomalies.

5.6. Reading time to critical word

Reading time to the critical word was adjusted for both word length and individual reading speed. The first step was to compute, for each participant, the mean reading times for words of each length, using all reading times other than those to the first and last word. Next, the best-fitting straight line was used to predict the reading time for each critical word. The adjusted times reported and analyzed were the observed times minus the predicted times. Thus, positive adjusted times reflect longer reading times adjusted for both the readers average reading speed and the length of the critical word.

A mixed factorial ANOVA was conducted on the adjusted reading times with group, sentence, task, and task order as factors. The only significant effect was the main effect of sentence type, $F(2, 88) = 3.62$, $p = 3.62$, partial $\eta^2 = .076$. Adjusted reading times to the sentences with syntactic violations take longer to read ($M = +28.6$ ms) compared to either correct sentences ($M = -.5.6$ ms) or sentences with semantic anomalies ($M = +.2$ ms). This is precisely the same pattern observed by Ditman et al. (in an acceptability task) and is consistent with the interpretation that the syntactic violations are usually detected immediately and cause an immediate processing load. The higher N400 amplitudes reported by both Moreno, S. et al., and Ditman et al. indicate that the verbs creating a semantic anomaly are more difficult to semantically integrate, but not to the extent that they disrupt the pace of reading at the critical word location. The critical findings for present purposes is that reading times to the critical word are the same for both bilinguals and monolinguals, the same for both tasks, and that there is no Group × Task or Group × Task × Sentence interaction.

5.7. Reading time to sentence-final word

Reading time to the sentence-final word was adjusted for both word length and individual reading speed using the procedure applied to the critical-word reading times. The only significant main effect was the main effect of sentence type, $F(2, 88) = 40.20$, $p < .001$, partial $\eta^2 = .477$. Both correct sentences ($M = +123$ ms) and sentences with semantic anomalies ($M = +163$) took longer than predicted from their length, whereas syntax violations were faster than predicted ($M = -54$ ms). This was the same pattern obtained by Ditman et al. using an acceptability task and is consistent with the interpretation that readers are fully engaged in normal comprehension processes and trying to resolve possible anomalies in a wrap-up process, except for when the sentence has a syntax error. When a syntax violation is detected readers can skim through to the prompt as the presence of a syntax error always means the correct answer is “no” regardless of the task.

As shown in Fig. 6 there is also a Task × Sentence Type interaction, $F(2,88) = 6.18$, $p = .003$, partial $\eta^2 = .123$. It is somewhat surprising that the interaction is caused by task differences with the correct sentences rather than with the semantic anomalies. Semantic anomalies produce equivalent amount of disruption on the sentence-final word regardless of whether they are relevant (acceptability task) or irrelevant (grammaticality task). Perhaps when participants are reading for comprehension semantic anomalies (if detected) automatically cause a wrap-up load even when task irrelevant. In contrast, participants reading correct sentences may do an additional final check for semantic anomalies only when those errors are task relevant. The account provided of the overall results is coherent if one assumes that participants usually read the entire sentence for normal comprehension except when they detect a syntax violation during the grammaticality task, whereupon, they shift to a skim mode.

5.8. Discussion of Experiment 2 results

There were several reasons for replicating the behavioral component of the Moreno, S. et al. experiment. Because their entire argument for a bilingual advantage in conflict resolution during sentence processing rests on the conclusion that the magnitude of the monolingual advantage is

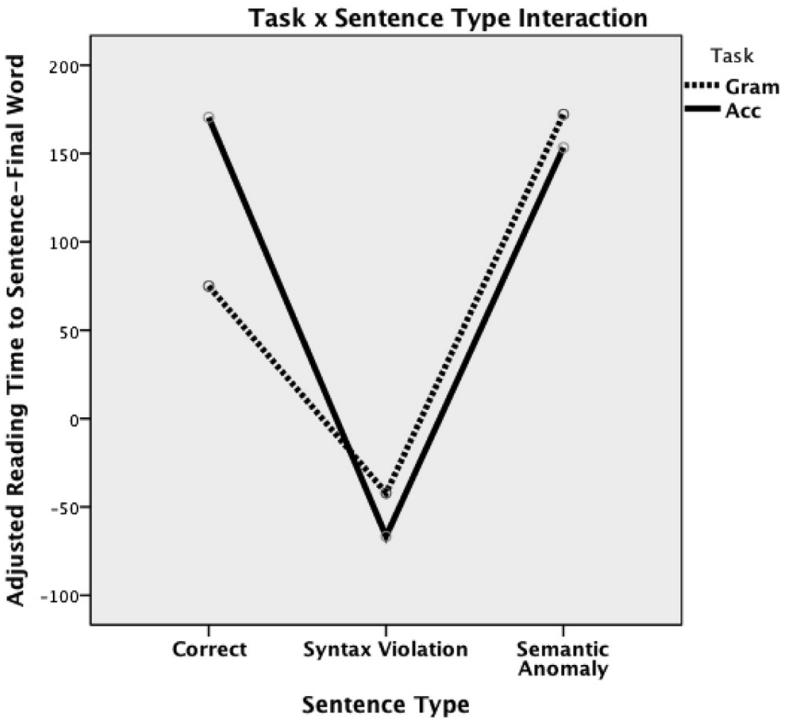


Fig. 6. The significant Task × Sentence Type interaction for the adjusted reading times to the sentence-final word.

significantly smaller in the EP-demanding grammaticality task, the most critical outcome was to test for a Group × Task interaction in accuracy to the final probe. Our results did produce the same pattern of monolingual advantages across tasks, but the Group × Task interaction was far from significant and thus provides no evidence for a bilingual advantage in EP. That is, the bilinguals did not close the gap in the grammaticality task regardless of the order of the two tasks. Likewise, by using the SPR procedure we were also able to test for a Group × Task interaction with three[4] additional dependent measures: RT to the probe question, adjusted reading time to the critical word, and adjusted reading time to the sentence-final word. The Group × Task interactions never approached significance, all p's $> .20$.

A secondary purpose of this replication was to further test Moreno et al.'s assumption that the grammaticality task was more EP-demanding than the acceptability task. If that assumption was true, then the acceptability task should yield higher accuracy, faster RTs to the probe question, and faster reading times compared to the acceptability task. As reported above, it was the grammaticality task that resulted in significantly better accuracy and faster RTs. There were no significant task differences with the reading time measures. The present results are inconsistent with the assumption that the grammaticality task was the more difficult task, as one would expect if it produced more conflict and required more EP.

Because Moreno, S. et al. confounded task with order, a tertiary purpose was to explore for order effects by counterbalancing the order of tasks. When the acceptability task was done first (reverse order) accuracy on both tasks declined and the size of the monolingual advantage grew in both tasks to the same magnitude for both tasks (6.6%). Although the Task × Order interaction was not significant, these trends are opposite of what one would expect if more difficult conditions allow bilinguals with superior EP abilities to close the gap compared to monolinguals who are slightly more proficient in English.

Experiment 2 shows that Moreno et al.'s critical behavioral test (based on separate ANOVAs for each of the two tasks) for a bilingual advantage in conflict resolution in sentence processing does not

[4] We also analyzed the unadjusted reading time to the critical word and sentence-final word. These also yield non-significant Group × Task interactions.

replicate (although our data show the same pattern). Furthermore, their failure to use the appropriate ANOVA to test for a Group × Task interaction and their failure to counterbalance the order of the tasks makes it likely that they never produced a statistically significant and general bilingual advantage to begin with.

6. A confirmation bias to find bilingual advantages in EP

Given the review of the literature on the bilingual advantage in inhibitory control in our introductory remarks, it appears that this specific research area has not escaped the replicability problem. Even worse, from our perspective, is that failures to replicate may have induced a widespread confirmation bias that sustains the theory of bilingual advantages in inhibitory control rather than leading to disconfirmation and falsification. The Ferguson and Heene (2012) article in the *Perspectives on Psychological Science* special issue on replicability evokes a strong image that appears to fit the current research practices on bilingual advantages all too well.

> "*There is a systematic discipline-wide problem in the way that disconfirming data is managed. Many theories, particularly those tied to politicized or "hot" topics, are not subjected to rigorous evaluation and, thus, are allowed to survive long past their utility. This is our use of the term "undead theory", a theory that continues in use, resisting attempts at falsification, ignoring disconfirming data, negating failed replications through the dubious use of meta-analysis or having simply maintained itself in a fluid state with shifting implicit assumptions such that falsification is not possible…. We suspect a good number of theories in popular use within psychology likely fit within this category; theories that explain better how scholars wish the world to be than how it actually is*" (p. 559).

Concisely stated, the prevailing view in the bilingualism field is that the ubiquitous practice of managing two languages enhances EP abilities and leads to bilingual advantages in both verbal and nonverbal tasks. Is this an "undead" theory? Although the diagnosis is debatable, it certainly has many of the undesirable symptoms. For example, there are strong tendencies to: (a) discuss outcomes that show advantages and ignore outcomes that do not, (b) accept any group differences (especially in ERP or fMRI) as evidence of a bilingual advantage, (c) select and report only those statistical analyses that support the advantage hypothesis, (d) ignore issues of test–retest reliability and convergent validity, (e) alter the interpretation of what a task measures in order to produce a confirmation, and (f) design and report underpowered experiments with risky small *n*'s.

The Moreno et al. study resonates with many of these pathological tendencies. For example, both the N400 and P600 data were interpreted as providing additional evidence that bilinguals exhibit enhanced inhibitory control during sentence processing. However, the next two sections show that there are highly plausible alternative interpretations for their ERP results.

6.1. Alternative interpretations of the N400 differences

During sentence processing the amplitude of the N400 component of the ERP to a critical word is usually assumed to index difficulty in semantic integration. Moreno, S., et al. report that "… our results show a larger N400 in bilinguals than in monolinguals in the grammaticality task… this larger negativity could be interpreted in terms of bilinguals dealing with conflict in which they were judging the syntax and answering 'correct' to a *semantically anomalous sentence*", p. 575. The underlying logic of this interpretation is not explicitly presented and the conclusion is, at best, counter intuitive. If superior executive attention allows one to filter the unwanted dimension then when semantics is irrelevant there should be a smaller effect of anomalous semantics, but this is the opposite of what they find: Semantic anomalies produced larger N400s in bilinguals suggesting that even when it is irrelevant semantics intruded on word processing more in bilinguals than monolinguals.

6.2. Alternative interpretations of the P600 group differences

The P600 amplitude generated by bilinguals was smaller in amplitude compared to monolinguals in the grammaticality task, but not in the acceptability task. Given that P600 is associated with the

detection, reanalysis, or repair of a syntax problem the smaller amplitudes for bilinguals is consistent with the interpretation that bilinguals invest fewer neural resources to deal with the syntactic violations. But what does this have to do with the hypothesis that bilinguals have better inhibitory control? Why should an enhanced ability to suppress a competing action plan based on semantics modulate P600, a component sensitive to difficulty in syntactic processing?

Another complication in interpreting the group differences is that the P600 amplitude is reduced or delayed at lower levels of L2 proficiency (Moreno, Rodriguez-Fornells, & Laine, 2008; Weber-Fox & Neville, 1996). Moreno, S. et al. (2010) discounted age-of-acquisition or proficiency as an account of the reduced amplitude of P600 for bilinguals in the grammaticality task by pointing out that the group differences in P600 in the acceptability task were not significant. Thus, the significant reduction in P600 amplitude observed in the grammaticality task can be attributed to the bilingual's having superior EP. This argument relies on accepting that the larger differences observed in the grammaticality task are significantly greater than the somewhat smaller differences obtained in the acceptability task. Unfortunately, an ANOVA that includes task as an additional within-subject factor is not reported and it is not known if there is a significant Task × Group interaction. Failure to report the more appropriate ANOVA repeats the problem described earlier for the behavioral differences in accuracy. In summary, there appear to be multiple alternative interpretations of the P600 data that do not attribute the group differences in P600 to a bilingual advantage in EP and to the dubious assumption that the grammaticality task is the more difficult and EP-demanding task.

6.3. *The shapeshifting grammaticality task*

Although assumptions about the degree to which a task requires EP may vary depending upon the adopted theoretical framework, researchers must be consistent across contemporaneous studies. Consequently, it is very surprising to discover that the same research group that assumes that the grammaticality task requires more EP than the acceptability task asserts a completely different view of the grammaticality task in a subsequent publication (Bialystok & Barac, 2012). In this article, the grammaticality task is asserted to be a measure of metalinguistic awareness akin to Berko's (1958) famous *wugs* task. While it is certainly true that the grammaticality task requires awareness of the distinction between syntactic and semantic violations, the intent to confirm a bilingual advantage compels Bialystok and Barac to juxtapose the grammaticality task as an EP-free task that can be contrasted to a prototypical task of inhibitory control (viz., the flanker interference effect). The morphing of the grammaticality task from an EP-demanding to an EP-free task, with neither mention nor justification, is a mystery; but does fit the profile of a firm confirmation bias.

6.4. *Underpowered small* n *studies*

All three dependent measures (accuracy, N400, and P600) are susceptible to spurious effects because there are only 14 participants in each group. This reduces the design's power to correctly reject the null hypothesis, but as Bakker, van Dijk, and Wicherts (2012) demonstrate with simulations, small *n*'s coupled with a publication bias against publishing null also results in an inflated rate of false positives. Similarly, in an analysis of studies in the neurosciences Button et al. (2013) show that the average statistical power is very low and that low power reduces the likelihood that a statistically significant result reflects a true effect. The *European Journal of Personality* in its recent recommendations for increasing replicability in psychological science urges increases in sample size and the avoidance of multiple underpowered studies (Asendorpf et al., 2013). Francis (2012) bluntly asserts that "*Studies with unnecessarily small sample sizes should not be published*" (p. 989). While sympathizing with the additional time required to collect ERP responses a cell size of 14 is a risky choice in testing for effect sizes that are likely to be moderate at best.

7. Conclusion

The primary purpose of this study was to test for bilingual advantages in conflict resolution during sentence processing. Experiment 1 examined the time-course of a homograph-interference effect

when test words were either presented immediately after the sentence-final word or after a delay. Bilinguals and monolinguals were equally adept at using the extra time to suppress the context-inappropriate meaning when the sentence-final word was a homograph. Experiment 2 was an unsuccessful replication of the behavioral component of the study by Moreno, S. et al. contrasting the *sentence acceptability* task to the *sentence grammaticality* task. When the tasks were given in the same order as Moreno, S. et al. the same pattern of means was obtained, but the magnitude of the bilingual disadvantage in RT in the grammaticality task was not significantly smaller than that observed in the acceptability task. When the tasks are given in the reverse order, there were no differences at all in terms of the magnitude of the bilingual disadvantage. Thus, neither experiment provided any evidence that bilinguals have superior conflict-resolution abilities during sentence processing. Several aspects of the Moreno, S. et al. study were examined for possible influences of a confirmation bias and this may explain the surprising and strong interpretations of their N400 and P600 findings that may, otherwise, appear to be unwarranted.

Acknowledgments

We thank the following members of the LACE (Language, Attention, & Cognitive Engineering) laboratory for their individual contributions to this project: Edgar Alcaine, Tavi Alvarez, Kirstin Anderson, Olimpia Andrade, Jack Darrow, Frank Du, Eddie Ferrero, Lynne Freeman, Jenesis Imai, Hunter Johnson, James Keenan, Sati Morgan, Lindsay Pellichia, Richard Farrell, Oliver Sawi, Ariz Sanchez, and Carlos Urtecho. We thank Michael Paap and Gus Woythaler for creating the materials used in the homograph interference task (Experiment 1).

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.
Asendorpf, J. B., Conner, M., De Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108–119.
Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives in Psychological Science, 7*, 534–554.
Berko, J. (1958). The child's learning of English morphology. *Word, 14*, 150–177.
Bialystok, E., & Barac, R. (2012). Emerging bilingualism: dissociating advantages for metalinguistic awareness and executive control. *Cognition, 122*, 67–73.
Bialystok, E., Craik, F. I. M., Grady, C., Chau, W., Ishii, Y., Gunji, A., et al. (2005). Effect of bilingualism on cognitive control in the Simon task: evidence from MEG. *NeuroImage, 24*, 40–49.
Bialystok, E., Craik, F. I. M., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and Aging, 19*, 290–303.
Bialystok, E., Craik, F., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Science, 16*(4), 240–250.
Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews, 14*, 365–376.
Degani, T., & Tokowicz, N. (2010). Semantic ambiguity within and across languages: an integrative review. *The Quarterly Journal of Experimental Psychology, 63*(7), 1266–1303.
Ditman, T., Holcomb, P. J., & Kuperberg, G. R. (2007). An investigation of concurrent ERP and self-paced reading methodologies. *Psychophysiology, 44*, 927–935.
Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and eye fixations in reading: a test of competing models of lexical ambiguity resolution. *Journal of Memory & Language, 27*, 429–446.
Duñabeitia, J. A., Hernández, J. A., Antón, E., Macizo, P., Estévez, A., et al. (2013, April). *The inhibitory advantage in bilingual children revisited* (in press).
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*(11), 1364–1371.
Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555–561.
Francis, G. (2012). Publication bias and failure of replication in experimental psychology. *Psychonomic Bulletin and Review, 19*(6), 975–991.
Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: a component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*(2), 245–262.
Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*(3), 430–445.
Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition, 33*(7), 1220–1234.
Gollan, T. H., Montoya, R. I., & Werner, G. (2002). Semantic and letter fluency in Spanish–English bilinguals. *Neuropsychology, 16*, 562–576.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*, 67–81.
Hart, L., & Perfetti, C. A. (2008). Learning words in Zekkish: implications for understanding lexical representation. In E. L. Grigorenko, & A. J. Naples (Eds.), *Single word reading: Behavioral and biological perspectives* (pp. 107–128). New York: Taylor & Francis.
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.
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*.
Kirk, N. W., Scott-Brown, K., & Kempe, V. (2013). Do older Gaelic-English bilinguals show an advantage in inhibitory control?. In *Proceedings of the 35th annual conference of the Cognitive Science Society* (in press).
Kousaie, S., & Phillips, N. A. (2012a). Conflict monitoring and resolution: are two languages better than one? Evidence from reaction time and event-related brain potentials. *Brain Research, 1446*, 71–90.
Kousaie, S., & Phillips, N. A. (2012b). Aging and bilingualism: absence of a "bilingual advantage" in Stroop interference in a nonimmigrant sample. *The Quarterly Journal of Experimental Psychology, 65*(2), 356–369.
Ledoux, K., Gordon, P. C., Camblin, C. C., & Swaab, T. Y. (2007). Coreference and lexical repetition: mechanisms of discourse integration. *Memory and Cognition, 35*(4), 801–815.
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.
Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*, 940–967.
McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: assigning roles to constituents of sentences. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Foundations: Vol. I. Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 272–325). Cambridge, MA: MIT Press.
Moreno, S., Bialystok, E., Wodniecka, Z., & Alain, C. (2010). Conflict resolution in sentence processing by bilinguals. *Journal of Neurolinguistics, 23*, 564–579.
Moreno, E. M., Rodriguez-Fornells, A., & Laine, M. (2008). Event-related potentials (ERPs) in the study of bilingual language processing. *Journal of Neurolinguistics, 21*(6), 477–508.
Osterhout, L., & Nicol, J. (1999). On the distinctiveness, independence and time course of the brain responses to syntactic and semantic anomalies. *Language and Cognitive Processes, 14*, 283–317.
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. & Sawi, O. (2013, July). Bilingual advantages in the antisaccade, flanker, Simon, and switching tasks: Are we measuring differences in cognitive control? (submitted for publication).
Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Ebro, & P. Reitsma (Eds.), *Precursors of functional literacy*. Amsterdam/Philadelphia: John Benjamin.
Ryskin, & Brown-Schmidt. (2012, November). A bilingual disadvantage in linguistic perspective adjustment. In *Presentation at the 53rd annual meeting of the Psychonomic Society. Minneapolis, MN*.
Sawi, O., & Paap, K. (2013, April). Test–retest reliability and convergent validity of measures of executive processing: evidence from the Simon, flanker, switching, and antisaccade task. In *Poster presented at the meeting of the Cognitive Neuroscience Society, San Francisco*.
Swaab, T. Y., Camblin, C. C., & Gordon, P. C. (2004). Electrophysiological evidence for reversed lexical repetition effects in language processing. *Journal of Cognitive Neuroscience, 16*, 715–726.
Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory and Cognition, 22*(1), 111–126.
Waltz, D. L., & Pollack, J. B. (1985). Massively parallel parsing: a strongly interactive model of natural language interpretation. *Cognitive Science, 9*, 51–74.
Weber-Fox, C., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8*(3), 231–256.