2-2016

Re-Examining the Bilingual Advantage on Interference-Control and Task-Switching Tasks: A Meta-Analysis

Seamus Donnelly

Graduate Center, City University of New York

How does access to this work benefit you? Let us know!

Follow this and additional works at: https://academicworks.cuny.edu gc_etds

🔗 Part of the Cognitive Psychology Commons, First and Second Language Acquisition Commons, and the Psycholinguistics and Neurolinguistics Commons

Recommended Citation

https://academicworks.cuny.edu gc_etds/762

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
RE-EXAMINING THE BILINGUAL ADVANTAGE ON INTERFERENCE-CONTROL AND TASK-SWITCHING TASKS: A META-ANALYSIS

by

SEAMUS DONNELLY

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York 2016
RE-EXAMINING THE BILINGUAL ADVANTAGE ON
INTERFERENCE-CONTROL AND TASK-SWITCHING TASKS: A META-ANALYSIS

by

SEAMUS DONNELLY

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Bruce Homer
Chair of Examining Committee

Bruce Homer
Executive Officer

Dr. Bruce Homer
Dr. Patricia Brooks
Dr. Jay Verkuilen
Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
Abstract

RE-EXAMINING THE BILINGUAL ADVANTAGE ON INTERFERENCE-CONTROL AND TASK-SWITCHING TASKS: A META-ANALYSIS

by

Seamus Donnelly

Advisor: Professor Bruce Homer

A much-debated topic in psycholinguistics is whether lifelong bilingualism enhances executive functions (EF), the set of higher-order cognitive processes involved in the control of thought and action. Several researchers have predicted bilingual advantages on various EF tasks, especially interference-control and task-switching tasks. Many studies have tested these predictions, but results have proven unreliable. As a complementary approach to recent quantitative syntheses on this topic, the present dissertation tests whether the bilingual advantage is moderated by a number of theoretically significant variables: dependent variable (DV), task, age, age of L2 acquisition and lab.

Two meta-analyses were conducted. Study 1 considered interference-control tasks. It synthesized 168 effect sizes from 43 studies. There was a significant interaction between age and dependent variable: the bilingual advantage was larger for children than young adults on global reaction times (global RTs), and larger for older adults than younger adults on both dependent variables. There was also a significant interaction between age of acquisition and dependent variable: samples with bilinguals that learned their second language early exhibited larger effect sizes on global RTs than samples with bilinguals that learned their second language later. However, both of these interactions could be explained in terms of differential publication bias. There was also a strong lab effect. Study 2 considered task-switching tasks. It synthesized 30
effect sizes from 10 studies. However, it yielded a non-significant overall effect size that was not moderated by DV or lab.

Overall, the two studies yielded relatively inconclusive evidence for the bilingual advantage. While Study 1 revealed some coherent patterns of moderation, all of these effects could be due to publication bias. Furthermore, Study 2 revealed no evidence for an advantage on task-switching tasks. Various limitations of the present analysis and the literature more broadly may have obscured bilingual advantages if they do exist.
# TABLE OF CONTENTS

Copyright Page ii  
Approval Page iii  
Abstract iv  
Acknowledgements vii  
List of Tables ix  
List of Figures x  

Chapter

1 INTRODUCTION: LITERATURE REVIEW AND JUSTIFICATION 1  
How might Executive Functions Play a Role in Bilingual Language Processing? 1  
Interference Control Tasks 4  
Task-Switching Tasks 8  
Attempts to Synthesize the Literature on Bilingualism and EF 11  

2 MODERATORS AND PREDICTIONS 16  
Dependent Variable 16  
Age 16  
Task 18  
Age of Acquisition 18  
Lab 21  

3 STUDY 1 23  
Method 24  
Results 41  
Summary of Results 57
4 STUDY 2 60

Method 60

Results 65

Summary of Results 70

5 DISCUSSION 71

Discussion and Theoretical Implications of Study 1 71

Discussion and Theoretical Implications of Study 2 76

Limitations and Future Directions 77

Conclusion 71

REFERENCES
Acknowledgments

I would first like to thank my advisor, Dr. Bruce Homer, for all your support over the past seven years and for giving me the freedom to explore varied topics. I would also like to thank Dr. Patricia Brooks, for all your time and feedback on various earlier version of this project, and Dr. Jay Verkuilen for always being willing to discuss statistics. I would never have been able to complete this project without help from all of you.

I would also like to thank my readers Dr. Mark Lauterbach and Dr. Howard Everson, for the their helpful comments during the dissertation defense. I would also like to thank Dr. Everson for all of his support over the last nine years.

Thank you to all my Graduate Center colleagues, Jeremy Sawyer, Eve Higby, Katie Pace Miles, Anna Schwartz. I would like to especially thank Russell Miller, for countless conversations about theory and method. Thank you also to the faculty of the Program in Educational Psychology, particularly Dr. Rindskopf and Dr. Ehri.

Thanks you to my friends, Kristie Bailey, Benham Jones, Natalia Mitrofanova, Stefan Hench and Scott Abrahams, for their encouragement for over the past seven years. Finally, thank you to my mother, my father and my grandmother for their endless support.
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Effect Sizes and Moderators for Each of the 84 Comparisons for Interference-Control Tasks</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>Overall Fit Statistics for Interference-Control Moderator Models</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>Regression Coefficients from Interference-Control Moderator Models</td>
<td>43</td>
</tr>
<tr>
<td>4</td>
<td>Estimated Effect Sizes and Confidence Intervals for Each Task</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>Estimated Effect Sizes and Confidence Intervals for Each Age Group</td>
<td>46</td>
</tr>
<tr>
<td>6</td>
<td>Estimated Effect Sizes and Confidence Intervals for Each AoA Group</td>
<td>53</td>
</tr>
<tr>
<td>7</td>
<td>Estimated Effect Sizes and Confidence Intervals for Lab Group</td>
<td>54</td>
</tr>
<tr>
<td>8</td>
<td>Effect Sizes and Moderators for Each of the 84 Comparisons for Task-Switching Tasks</td>
<td>63</td>
</tr>
<tr>
<td>9</td>
<td>Overall Fit Statistics for Moderator Models Task Switching Tasks</td>
<td>66</td>
</tr>
<tr>
<td>10</td>
<td>Regression Coefficients from Moderator Models Switch Tasks</td>
<td>67</td>
</tr>
<tr>
<td>11</td>
<td>Estimated Effect Sizes and Confidence Intervals for Each AoA Group</td>
<td>68</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1  Schematic of Data Structure Using Paap and Sawi (2014) as an Example  26
2  Forest Plot for Interference Costs  41
3  Forest Plot for Global RTs  42
4  Funnel Plots for each Age Group  46
5  Funnel Plots for Recoded Age Group  47
6  Funnel Plots for Interference Costs and Global RTs for Each Age Group  49
7  Funnel Plots for Interference Costs and Global RTs for Recoded Age Group  51
8  Funnel Plots for Interference Costs and Global RTs by Age of Acquisition  52
9  Forest Plots for Switch Costs  65
10 Forest Plots for Mixing Costs  66
Chapter 1: Literature Review and Justification

Much research over the last decade and a half has sought to identify the cognitive benefits of lifelong bilingualism. Specifically, many researchers have argued that mentally juggling two languages exercises and, consequently, strengthens the set of domain-general processes for controlling attention and action, sometimes called executive functions (Bialystok & Kroll, 2013; Hilchey & Klein, 2011; Kroll, Dussais, Bogulski & Kroff, 2012). While this prediction has been tested several times, results have been inconsistent (see Paap, Sawi & Johnson, 2015, for a review). This dissertation seeks to clarify the relationship between bilingualism and executive functions by conducting a meta-analysis, a quantitative research synthesis for assessing the reliability of an effect across studies. This introduction begins with a discussion of the executive function system and then outlines several proposals for how bilingualism may affect this system.

How might executive functioning play a role in bilingual language processing?

The executive function system (EF), sometimes referred to as executive control, is the set of domain-general cognitive processes for regulating thought and action (Miyake & Friedman, 2012). Behaviorally, such processes are measured by tasks that require the suppression of distracting information, switching between tasks, and the updating of information in working memory (Miyake et al., 2000). Neurophysiologically, such processes are strongly associated with pre-frontal areas and possibly some sub-cortical regions as well (Blair, 2006).

While the specifics of this system are poorly understood, two broad conclusions seem warranted. First, structural equation modeling studies conducted on a broad array of EF tasks have demonstrated that the covariance on EF tasks can be decomposed into several distinct, though correlated, subcomponents (Miyake et al., 2000). This finding is typically taken as
evidence that the EF system contains several separable processes that share some neurocognitive resources. Some of the earliest of these studies identified three correlated but separated factors corresponding to inhibition, switching and updating (Miyake et al., 2000) More recent research has found that the inhibition tasks load on a general EF factor that accounts for variance in the shifting and updating factors (for an overview see Friedman & Miyake (2012)). This common EF component is often interpreted as goal-maintenance, since such a construct would be involved in all the EF tasks given in a single battery. These three constructs do not exhaust the concept of EF, and others potential components have been posited, such as monitoring (Botvinick, Cohen & Carter, 2004), planning (Miyake & Friedman, 2012). Moreover, these components uniquely predict other cognitive outcomes, such as IQ (Miyake & Friedman, 2012) and higher-order EF tasks (Miyake et al., 2000), furthering suggesting they are separable.

Second, these tasks seem to be related to many significant life outcomes (Friedman & Miyake, 2012). Variation in EF has been associated with academic skills, including math knowledge amongst children (Blair, Urasche, Greenberg, Vernon-Fegans & The Family Life Project Investigators, 2015), and procrastination amongst college-aged participants (Gustavason, Miyake, Hewitt & Friedman, 2015). They have also been associated with externalizing disorders, such as attention deficit hyper disorder and substance abuse (Young et al., 2009), and socially significant behaviors such as implicit racism (Ito et al., 2015).

Given the practical significance of EF, many researchers have asked what sorts of experiences affect EF (Diamond & Lee, 2011). One commonly proposed influence on EF is bilingualism (Bialystok et al., 2010). There are at least two reasons why EF might be recruited during bilingual language processing. First, there is strong evidence that each of bilingual’s two languages remain active even while a single language is in use. For example, in picture-naming
and word-reading studies bilinguals name cognates in each of their two languages more quickly than unrelated words (van Hell & Djikstra, 2002) but monolinguals do not (Jared & Szucs, 2002). This finding suggests that cognates activate word representations in each of a bilingual’s languages, which reinforce one another. Likewise, bilinguals name cross-language homographs, words with different meanings but the same form, in each of their two languages more slowly than control words, but monolinguals do not, suggesting interference between the two activated word forms (de Groot, Delmaar & Lupker, 2000). These effects can even be observed when the target word occurs within a sentence, suggesting they are not artifacts of isolated-word naming (for a review see Kroll, Dussais, Bogulski & Kroff (2012)). Many models of bilingual lexical access assume, or accommodate the assumption, that these representations compete and that this competition is resolved through some control process (e.g., Green, 1998; Djiksta & Van Heuven, 2002). If competition between these representations is resolved by EF, then it follows that these processes might be strengthened through life-long bilingualism (Bialystok et al., 2009). As a result, researchers have hypothesized that bilinguals may have an advantage over monolinguals on tasks that require EF, particularly those that have high amounts of interference.

A second possibility is agnostic to the nature of bilingual lexical representation. Regardless of whether words from each language compete, bilinguals regularly use and, therefore, switch between each of their two languages (Prior & Gollan, 2011). It is possible that EFs are recruited for language switching, much in the way they are recruited in domain-general task switching. Indeed language-switching experiments reveal patterns of results very similar to those from domain-general task-switching experiments, suggesting a possible shared mechanism (for a review see Hernández et al. (2013)).
These two possibilities are distinct but not mutually exclusive, and they have lead to predictions for bilingual advantages on two classes of tasks, which align with the two explanations described above: interference-control tasks and task-switching tasks. The following sections reviews results for studies comparing monolinguals and bilinguals on interference-control tasks and task-switching tasks.

**Interference Control Tasks.**

Early theories of bilingual language control proposed that competition between representations from each of a bilingual’s two languages was resolved by domain-general inhibition mechanism (Green, 1998). From this perspective both target and non-target language representations are active even while only one is being used, and the non-target language representations are inhibited. Evidence for the engagement of such a mechanism includes a correlation between performance on tasks of inhibition and cross-language intrusions (Festman, Rodriguez-Fornells & Munte, 2011). This leads to a straightforward prediction: bilinguals should exhibit smaller interference costs than monolinguals on interference-control tasks.

Examples of interference-control tasks include the Simon, Flanker (or ANT) and Stroop tasks. In all these tasks, participants respond to visually presented stimuli with target and distracter dimensions. For example, in each trial of the Simon task, participants see a red or green square on either the right or left side of the screen. They respond to the color of the square by pressing a key on the left side of the keyboard for red squares and a key on the right side of the keyboard for green arrows (or vice versa). It is assumed that the color (the target dimension) and location (the distracter dimension) each prime responses, and that, when these responses differ, the response primed by the distracter dimension interferes with that primed by the target dimension. Similarly in each trial of the flanker task, participants see a row of arrows, one of
which is the target and the remaining are distracters. On some trials the target and distracters point in the same direction and in others they point in different directions. As was the case with the Simon task, it is assumed that the target and distracter arrows prime responses and that the response primed by the distracter arrows interferes with that primed by the target arrows. In the Stroop task, participants see color words that are printed in different colors. They must indicate the color of the font (the target dimension) while ignoring the color word itself (the distracter dimension). For all these tasks, trials in which the target and distracter dimensions prime different responses are called incongruent trials and trials in which they prime the same response are called congruent trials. Generally speaking, reaction times (RTs) are longer on the incongruent than congruent trials, because of the extra time necessary to inhibit the non-target response. The difference between mean RTs of incongruent and congruent trials is called the interference cost, and is often viewed as an indicator of the efficiency of inhibition. If lifelong bilingualism strengthens the inhibition mechanism, then bilinguals should have smaller interference cost than monolinguals.

Early work suggested that bilinguals exhibit smaller interference costs than monolinguals. In a seminal paper, Bialystok et al. (2004) reported on three studies comparing monolingual and bilingual middle-aged and older adults on the Simon task. Across all three studies, bilinguals exhibited smaller interference costs than monolinguals, and this difference was larger for older than middle-aged adults. However, results of studies attempting to replicate this finding have been mixed. In a review of 13 studies published before 2011, Hilchey and Klein (2011) concluded that a bilingual advantage on interference costs was elusive, only appearing in the minority of samples. Since this review, however, Luk, De Sa and Bialystok (2011) have observed smaller interference costs amongst bilingual college students than their monolingual
peers, an effect that was moderated by the age of onset of the second language. Furthermore, two studies have reported a bilingual advantage on interference costs amongst children (Poarch & Van Hell, 2013; Yang, Yang & Lust, 2011). On the other hand, Paap and Greenberg (2013) found no bilingual advantage in interference costs for mixed-language bilingual college students, and three large-scale studies found no advantage in children (Antón et al., 2014; Duñabeitia et al., 2014; Gathercole et al., 2014).

An important caveat is in order. While interference costs from the three tasks are widely assumed to reflect the same construct, it is not clear that they do. While factor-analytic studies have reported that many tasks of inhibition load onto a single factor (Miyake et al., 2000), few studies have included all of the Simon, Fanker and Stroop task, and when they have, they have typically not used interference costs. For example, Miyake et al (2000) found that the difference between incongruent and neutral trials on the Stroop task (trials without distracters) loaded on an inhibition factor (later re-named common EF (Friedman & Miyake, 2012), along with the stop-signal and anti-saccade task. In a subsequent study, Friedman and Miyake (2004) found that the difference in RT between congruent and neutral costs on the Flanker task loaded on a shared factor with the analogous cost from the Stroop task. It is not clear whether interference costs would follow the same pattern. This point is discussed further by Paap and Sawi (2014).

Furthermore, studies that have investigated the similarity of the interference costs across tasks have generally found non-significant, near-zero correlations (Humphrey & Valian, 2012; Paap & Greenberg, 2013) Analyses of interference costs, then, ought to consider results from each of these tasks separately, as well as together, as will be described further in Chapter 2.

In the aforementioned review of studies of bilingual advantages on interference-control tasks, Hilchey and Klein (2011) noted that, while bilinguals often did not exhibit smaller
interference costs than monolinguals, they did often exhibit smaller global RTs (the average reaction time across both congruent and incongruent trials). Indeed, since the earliest studies on the topic, authors have pointed to this finding as additional evidence for a bilingual advantage in EF (Bialystok et al., 2004). While global RTs do not have an agreed-upon interpretation in the cognitive psychology literature, several authors have put forward explanations for a bilingual advantage on this measure that invoke goal maintenance and subsequent detection of goal-relevant stimuli (Hilchey and Klein, 2011; Bialystok et al., 2012). Specifically, the competition between representations from each language would require stronger goal activation for lexical selection in the target language, leading to up-regulation of cognitive control, regardless of whether this control is achieved via inhibition or some other mechanism. Adaptations to the regular conflict in language use might extend to any high-conflict environment. In other words, one could follow the same theoretical trail laid out for the bilingual advantage on interference costs to a slightly different dependent variable. While this hypothesis is plausible and worthy of consideration, it is important to remember that this interpretation of global RTs was initially post hoc and is unconventional in the broader cognitive psychology literature. Indeed, while the general EF score in Friedman and Miyake’s (2012) model, which they interpret as a measure of goal maintenance, contains the Stroop task, these authors include the difference between incongruent trials and baseline trials, not global RTs.

While Hilchey and Klein (2011) observed that this advantage seemed ubiquitous, two subsequent studies have reported a bilingual advantage in interference cost but not overall RT (Luk, De Sa & Bialystok, 2012; Yang, Yang & Lust, 2011), and several subsequent studies have observed no advantage on either (Antón et al., 2014; Duñabeita et al., 2014; Gathercole et al.,

In summary, there have been two proposals for bilingual advantages on interference-control tasks. One proposal predicts they will emerge on interference costs, while the other proposes they will emerge on global RTs. While both of these proposals have been extensively studied, the evidence is mixed for both.

**Task Switching Tasks.**

A second set of studies have tested the claim that switching attention between L1 and L2 representations might engage domain-general task-switching mechanisms. If this is so, bilingual advantages might emerge in domain-general task-switching tasks. Task-switching is typically studied in tasks that contain three blocks, all of which contain the same set of bivalent stimuli. For example, in the color-shape switching task, stimuli vary according to shape (for example, circle or square) and color (for example, red and green). In the first block, participants attend and respond to one dimension by, for example, pressing Q when the color is red and P when the color is green. In the second block, participants attend and respond to stimuli according to a second dimension by, for example, pressing Q when the shape is a circle and P when the shape is a square. In the third block, called the switch block, a cue prior to each trial signals the dimension of the stimulus that should be attended to. The switch block contains two types of trials: repeat trials, in which participants attend to the same dimension as the previous trial, and switch trials, in which participants attend to a different dimension from the previous trial. The switch cost, the difference in RT between switch and non-switch trials, is thought to represent the extra time it takes to initiate the new task-set.
Interestingly, as noted by Hernández et al. (2013) there is a large set of empirical phenomena from domain-general task-switching studies that have been replicated in language-switching tasks (a variant of task-switching tasks, in which participants name pictures in either their L1 or L2 according to a cue). For example, it has often been observed that switching to the less challenging task takes longer than switching to the more challenging task (Koch, Gade, Schuch & Philipp, 2010). Similarly, in language-switching tasks, it takes more time to switch to the stronger L1 than weaker L2 (Meuter & Alport, 1999). Additionally, both language- and task-switching paradigms typically yield so-called N2-repetition costs: Participants who complete three distinct tasks in consecutive blocks, switch into the third block more quickly than do participants who repeat the first task in the third block (Phillip & Koch, 2009). Furthermore, imaging studies of language-switching indicate that language-switching engages many of the cortical regions engaged in task-switching (Abutalebi & Green, 2008).

These similarities have led researchers to speculate that language-switching recruits the same neurocognitive mechanisms as domain-general task switching. If this is the case, then bilinguals might exhibit reduced switch costs relative to monolinguals, a hypothesis that has been tested in several studies.

Garbin et al. (2010) found that Spanish monolinguals exhibited a larger switch cost than did Spanish-Catalan bilinguals, whose switch cost did not statistically differ from zero. Prior and MacWhinney (2010) report similar results, and Prior and Gollan (2011) found that Spanish-English bilinguals, who reported regularly switching between languages, exhibited smaller switch costs than did the monolinguals, but Chinese-English bilinguals, who reported not regularly switching between languages did not. On the other hand, Paap and Greenberg (2013) did not observe a bilingual advantage in switch cost for a group of mixed-L2 bilinguals.
Furthermore, Hernández et al. (2013) administered several task-switching tasks, including one identical to that by Prior and MacWhinney (2010), to Spanish-Catalan bilinguals and Spanish monolinguals, and found no evidence of a bilingual advantage on any measure of switch cost. Importantly, the authors note, that the sociolinguistic context in Catalonia encourages regular language switching during ordinary conversation.

Task-switching tasks also permit the calculation of so-called mixing costs, which have also been studied by bilingual advantage researchers. The mixing cost, the difference in average RT between trials in single-task blocks and repeat trials in switch-task blocks, is thought to represent the amount of additional time needed to monitor for cues for which task set to engage (Paap & Greenberg, 2013). Hernández et al. (2013) speculated that bilinguals may constantly monitor for social and linguistic cues to which language to use, and this constant monitoring may lead to more a efficient domain-general monitoring system, which would be consistent with smaller mixing costs.

Little research has found an advantage for bilinguals on mixing costs, however. Barac and Bialystok (2012) found that three groups of bilingual children exhibited smaller mixing costs than their monolingual peers, but did not differ from one another. However, none of Prior and MacWhinney (2010), Prior and Gollan (2011), Paap and Greenberg (2013), Hernández et al (2013), or Paap and Sawi (2014) observed such differences among college students. Additionally, Gold et al. (2013) only reported on switching costs, and Garbin et al. (2010) did not include a non-switch block so mix costs were not calculated.

Similar to the work on interference-control tasks, there have been two proposals for a bilingual advantage on task switching tasks, each of which aligns with a different dependent
variable, switch and mixing costs. Evidence for an advantage on switch costs is unreliable and evidence for advantage on mixing costs is weak.

**Attempts to synthesize the literature on bilingualism and executive function**

Many researchers have sought to synthesize these studies, with an eye toward explaining the lack of reliability. In a recent literature review, Valian (2015) argued that many cognitively challenging activities, including music and exercise, involve and strengthen EF. She speculated that bilingual advantages on EF tasks are real but, as they compete with advantages conferred by these other activities, they are difficult to detect in individual studies. If this is true, one promising approach is to quantitatively synthesize many studies, rather than consider individual studies, since, in a single study the effect of bilingualism may be overwhelmed by other factors. In addition to two older syntheses (Costa et al., 2009; Hilchey & Klein, 2011), two recent studies have quantitatively synthesized the large database of bilingual advantage studies.

(de Bruin et al. (2015) conducted two analyses to determine whether publication bias may explain the unreliable findings on the bilingual advantage. First, they collected conference abstracts reporting on comparisons between bilinguals and monolinguals on any cognitive task, and conducted a logistic regression to determine whether abstracts reporting an advantage were more likely to result in publication than those that did not. Results revealed 63% of conference abstracts reporting a bilingual advantage were published while 36% of studies reporting no bilingual advantage were published. Second, they conducted a meta-analysis on all the published literature, which yielded a medium effect size ($d = .30$), with funnel plots indicating strong evidence of publication bias.

The de Bruin et al study (2015) is a valuable contribution to the sociology of science, showing that studies reporting a bilingual advantage are more likely to get published than those
that do not. However, it does not eliminate the possibility of a bilingual advantage. First, as noted by Bialystok et al. (2015), publication bias is ubiquitous to psychology. Its existence does not mean the absence of an effect. Second, the de Bruin et al. study included any study that contained a bilingual group, a monolingual group and some sort of cognitive task. For example, they included studies with clinical populations, studies with bilinguals with very little L2 experience, studies in which the dependent variable was a mathematics task, and studies with linguistic tasks, such as verbal fluency. However, there is no reason to believe that all bilinguals differ from all monolinguals on all tasks. As discussed earlier, there are several competing predictions about which tasks bilinguals should outperform monolinguals on. There are even different predictions about which score from a particular task should exhibit a bilingual advantage (e.g., the global RT or interference cost from interference-control tasks). Moreover, as both EF and the bilingual lexicon are, arguably, dynamic systems adapted to experience, even if the appropriate task is selected, it is unlikely that all bilinguals would differ from all monolinguals on it. A consistent advantage may emerge for particular tasks or groups, even if the entire literature suffers from publication bias.

A second synthesis was conducted by Paap, Sawi and Johnson (2015). They surveyed the literature and summarized the results of all studies comparing monolinguals and bilinguals on interference-control tasks and task-switching tasks. They relied on a vote-counting procedure, in which they coded each comparison as either supporting or failing to support a bilingual advantage according to whether it yielded a significant $p$-value. Across all measures, proportions of comparisons yielding a significant bilingual advantage were low (the highest proportion of significant tests as $\sim .22$). They argued that the few significant comparisons might reflect
questionable research practices endemic to psychology, confounds and between-group differences on non-EF constructs measured by these tasks.

Paap et al.’s (2015) contribution is important, especially because of its focus on questionable research practices and interpretive methods in psychology research in general and bilingual advantage research in particular. These factors certainly account for some the discrepant findings. However, their review does not definitively disprove the existence of a bilingual advantage. First, as mentioned above, it is unlikely that the advantage exists for all individuals. Looking solely at aggregate results might obscure effects for specific groups. Second, as noted by Linck (2015), the vote-counting procedure is potentially misleading as it conflates large and small effect sizes. A more appropriate technique for synthesizing these results is meta-analysis.

A complementary approach to the two reviews above is to conduct a set of meta-analyses that include a narrower body of tasks and participants, and examine the effect of theoretically significant moderator variables. This is the approach taken in the present dissertation. The present dissertation is a meta-analysis that differs from the previous quantitative syntheses in several ways. First it contains only studies with psychologically typical bilinguals and excludes atypical populations, such as bimodal bilinguals and individuals suffering from Alzheimer’s, as well as second language learners. Second, unlike the analysis reported by de Bruin et al. (2015), it only contains computerized interference-control and task-switching tasks. These tasks are both the most well studied tasks in the cognitive psychological literature and the most well represented tasks in the bilingual advantage literature. They also represent an attractive middle ground relative to the other tasks: they are arguably non-linguistic (unlike, for example, verbal fluency), yet conceptually similar enough to putative processes in bilingual language processing.
to reveal a bilingual advantage (unlike, for example, a mathematics test). Third, this dissertation also examines the impact of several theoretically significant moderators. As argued above, a bilingual advantage may be restricted to some participants and some tasks. If larger effect sizes emerge for certain participants or tasks, and this pattern is consistent with theories about bilingual language processing and EF, this would provide strong evidence for the existence of the bilingual advantage.

The present dissertation contains two sets of meta-analyses. The first considers interference-control tasks, since these are the most ubiquitous. The second contains task-switching tasks. As outlined above, predictions for bilingual advantages on these two tasks derive from slightly different assumptions about the locus of the bilingual advantage. Hypotheses about advantages on either interference-control measure typically stem from considerations of bilingual lexical representation, and hypotheses about advantages on task-switching measures typically stem from considerations of bilingual language use. It was therefore reasoned that for simplicity’s sake the two sets of tasks should be analyzed separately.

Each set of meta-analyses considers the effects of theoretically significant moderator variables. In Chapter 2, the theoretical justification for each moderator variable is discussed and predictions are made about the pattern of moderation. All of the moderators in Chapter 2 were considered in the analyses of interference control tasks. However not all moderators could be considered in the analyses of task switching tasks; there are fewer of these studies, and, as a result, not all of the potential moderators varied across these studies. Therefore, the following section states whether the moderator was considered for the interference control tasks only or for both sets of tasks.
Chapter 2: Moderators and Predictions

Chapter 1 reviewed research on the hypothesized bilingual advantage. It argued that results were unreliable and that one way forward was a meta-analysis. The present dissertation will consist of two-meta-analyses: one on interference control tasks and one on task-switching tasks. This chapter describes each of the moderator variables to be included in these models. Since not all moderator variables could be reliably coded for the smaller tasks-switching dataset, some moderators were only included for interference-control tasks.

DV

An obvious moderator is dependent variable (DV). This is defined as interference cost or global RT for interference-control tasks and switch cost and mixing cost for task-switching tasks. As outlined in the previous section, these costs are thought to reflect separate constructs and finding a consistent advantage on one rather than another would suggest different loci for the bilingual advantage. As all of these loci seem plausible, no prediction is made about the direction of this relationship.

Age (Interference Control Only)

One moderator that has been considered extensively in the bilingual advantage literature is age. It is agreed upon that executive functions exhibit a complex developmental trajectory, developing slowly during childhood, peaking in early adulthood and declining later in life (for a discussion see Zelazo and Lee, 2010). For example, Waszak, Li and Hommel (2010) conducted a lifespan study of interference cost on the flanker task, and observed smallest costs between the ages of 16 and 42. Interference costs decreased non-linearly over childhood and increased non-linearly after the age of 42. Bialystok, Martin and Viswanathan (2005) have argued that when EF is operating at peak efficiency, bilingual advantages might not be easily detected. If bilingualism
accelerates the development of and ameliorates the decline of this system, strong advantages may be seen during childhood and older adulthood.

Consistent with this prediction, there are a large number of studies showing a bilingual advantage on interference-control tasks amongst children (Engel de Abreu et al., 2012; Kapa & Colombo, 2013; Martin-Rhee & Bialystok, 2008; Poarch & Van Hell, 2013; Poarch & Bialystok, 2015; Yang, Yang & Lust, 2012). However, a few recent, large-scale studies have failed to observe these effects (Antón et al., 2014; Duñabeitia et al., 2014; Gathercole et al., 2014). As would be predicted, findings are even more mixed among young adults, with some studies finding a bilingual advantage on interference control tasks (e.g. Costa, Hernandez & Sebastian-Galles, 2008; Coderre, Van Heuven & Conklin, 2013; Luk et al., 2010; Luk, de Sa & Bialystok, 2011), and several studies failing to do so (Bialystok, Craik, Klein & Viswanathan, 2004; Gathercole et al., 2014; Paap & Greenberg, 2013; Paap & Sawi, 2014). Results with older adults are also mixed. Bialystok et al. (2004) report several studies showing a bilingual advantage for older adults on the Simon task, and Salvatierra and Rosselli (2011) report one, too, but Kousaie and Phillips (2011), Kirk et al. (2014) and Gathercole et al. (2014) do not.

Theory and results suggest that the bilingual advantage may be moderated by age. If this is correct, one of three patterns of results will hold. First, there could be a main effect of age, with children and perhaps older adults showing larger advantages than young adults. Second, since the advantage might be specific to interference cost or global RTs, there could be an interaction between age and DV, showing that children and older adults differ from young adults on one of the two measures. Third, it is possible there will be an interaction between age and DV, with children and older adults outperforming young adults on different DVs. If this were the case, it would suggest that these measures reflect different constructs across the lifespan. While it
is possible that costs on the switching tasks would interact with age, very few task-switching studies have included children \( (n = 1) \) or older adults \( (n = 1) \). Therefore, age could not be considered as a moderator in Study 2.

**Task (Interference Control Only)**

Several interference-control tasks have been used to compare monolinguals and bilinguals. While these tasks are often viewed as interchangeable indicators of the same constructs, there are several reasons to doubt this claim. First, as discussed in Chapter 1, structural equation modeling studies have typically not included all three tasks in the same battery, nor have they used costs corresponding to interference costs or global RTs for any of these. Second, as reviewed by Paap and Sawi (2014), correlations between raw scores on these tasks are often quite low, especially the interference costs. For example, Paap and Greenberg (2013) reported non-significant correlations between interference costs in the Simon and flanker task \( (r = .01) \). Strikingly, interference costs across different versions of the same flanker task are often uncorrelated as well (Shilling, Chetwynd & Rabbitt, 2002; Salthouse, 2010). Global RTs tend to exhibit larger correlations; for example Paap and Sawi (2014) report correlations of .6 between global RTs from Simon and Flanker tasks. Given this pattern of correlations, it is sensible to question whether the Simon, Flanker and Stroop tasks measure the same constructs. If the three tasks measure separate constructs, a bilingual advantage may emerge on only a single task. It is, therefore, possible that task will moderate effect sizes or that it will interact with DV. However, there is no strong reason to make a directional prediction. For task-switching tasks, the majority has used some version of the color-shape task, so task will not be included as a moderator in Study 2.

**Age of Acquisition (Interference Control Only)**
Many researchers have argued that reliable bilingual advantages may exist only for certain bilingual populations. Indeed, variables such as proficiency and age of acquisition play important roles in models of bilingual lexicon and empirical research on language switching and L1-L2 priming. Because proficiency is not reported in a standardized way across studies, it could not be considered as a moderator in this study. However, age of L2 acquisition is often reported, and appears to play an important role in bilingual lexical development.

An influential computational model of bilingual lexical development is DevLexII (Li, Farkas, & MacWhinney, 2004). DevLexII is an unsupervised connectionist network that uses a self-organized feature map and Hebbian learning to model the emergence of the lexicon. It accounts for many important phenomena from language acquisition, including vocabulary bursts and sensitive periods (for an overview see Li, 2012). The model has been used to simulate lexical development for varying L2 Age of Acquisition (AoA). The most important finding is that when the second language is introduced much later than the first language, first language representations have become entrenched and cover the entire semantic space. As a result, L2 words do not develop independent semantic representations but are parasitic on representations from the L1 and are compressed into small areas of the semantic space. However, when the second language is introduced shortly after the first language, because first language representations are less entrenched, L2 words develop independent semantic representations, leading to two competing lexicons.

Empirical evidence is consistent with the notion that the organization of the bilingual lexicon depends on AoA and clarifies the consequences of AoA for language processing. Sabourin, Brien and Burkholder (2014) compared three bilingual groups of varying ages of acquisition in a masked-priming lexical decision experiment. Four prime types were compared:
identity primes, unrelated primes, L2 translation-equivalent primes, and L1 semantic-associate primes. For the sequential and early bilinguals, L2 translation-equivalents significantly primed target words and did not significantly differ from L1 semantic-associates. For later L2 learners, L2 translation-equivalents did not significantly prime L1 words and differed significantly from semantic-associate primes. Importantly the early and late L2 groups had been matched on proficiency, so group differences likely reflect differences in age of acquisition rather than proficiency. The authors also reported that age of acquisition, treated as an independent variable, significantly predicted L2 translation-equivalent priming effects.

The computational and empirical results above suggest two things: that the organization of the bilingual lexicon depends on AoA, and that for earlier L2 learners, L2 representations can affect L1 processing. These results suggest that bilinguals with earlier AoAs might experience competition between representations when using both the L1 and L2, whereas later learners might only experience competition when using the L1. If this is the case, then early bilinguals should exhibit larger advantages than later bilinguals.

At least three studies have compared bilinguals of varying AoAs to monolinguals on interference-control tasks. Luk, de Sa and Bialystok (2011) compared college-aged early bilinguals (average age of regular use of L2: 5) to late bilinguals (average age of regular use of L2: 15) and to monolinguals on the flanker task and found that early bilinguals exhibited significantly smaller interference costs than the other two groups, who did not significantly differ. Subsequent regressions found that AoA, treated as a continuous variable, was positively related to RT. Kapa and Colombo (2013) compared ten-year old monolinguals, early bilinguals (L2 learned before age of 3) and late (L2 learned after 3) bilinguals on the flanker task. They found that, after controlling for age and English receptive vocabulary, the early bilinguals
outperformed both the late bilinguals and the monolinguals. Tao et al (2013) compared college-aged monolinguals to early (age of exposure to L2: 3 years old) and late (age of exposure to L2: 8 years old) bilinguals on a lateralized flanker task. Early bilinguals exhibited smaller global RTs and interference costs than monolinguals, while late bilinguals only exhibited smaller interference costs.

These results suggest that AoA moderates the bilingual advantage. It is possible that there will be a main effect of age of acquisition, with early AoA samples yielding larger bilingual advantages than later AoA samples. If the bilingual advantage is specific to a particular DV, there might be an interaction between AoA and DV, with larger bilingual advantages on one of the two dependent variables. While it is logically possible that AoA will moderate effect sizes in the task-switching dataset, the range of AoAs amongst these studies is quite small, so AoA will not be included as a moderator in Study 2.

Lab

There are at least two reasons why lab could moderate effect sizes. First, it is conventional to control for lab in meta-analyses, as some research findings might be artifacts of the particular procedures of a few labs. For example, Powers et al. (2013) conducted a meta-analysis comparing videogame players and non-players on various cognitive tasks. They conducted separate meta-analyses on correlational and experimental studies, and found that lab was a significant moderator among correlational studies but not experimental studies. They interpret these finding as a Hawthorne Effect: Some research teams are well known for conducting research on the benefits of video games. Participants at these universities may be aware of these researchers’ hypotheses and act accordingly. Hawthorne effects may also arise from causes independent of reputation such as wording of consent forms and scripts, which
might differ across labs. It is also possible that such site-specific effects reflect differences in the
degree of confounding between independent and extraneous variables across sites. In other
words, frequent players of videogames at University A may differ dramatically from frequent
players of videogames at University B. Such a scenario seems very likely in the bilingual
advantage literature, as the relationship between bilingualism and SES likely varies across
countries.

Second, different labs might have access to different groups of bilinguals. As argued
earlier, it is plausible that the bilingual advantage is specific to certain groups of bilinguals. For
example, executive control skills may be related to second-language proficiency, frequency of
second-language use, or any number of other variables. If such variables are important, and these
variables on average vary across labs, strong lab effects might be present.

A main effect of lab will be consistent with either of these scenarios. However, if the
main effect from lab is due to differences in confounding variables or Hawthorne effects across
universities, it is unlikely that lab will interact with DV. If the lab effect is due to differences in
bilingual characteristics, it is more likely that lab will interact with DV, since it is plausible that
bilingual effects are restricted to a single construct. A significant interaction between lab and
DV, therefore, would be more consistent with the latter interpretation.
Chapter 3: Study 1

Study 1 considered the effects of several moderator variables in the interference-control task dataset. Given the considerations in Chapters 1 and 2, the following set of hypotheses were tested:

1) Consistent with de Bruin et al. (2015), there will be a statistically significant small-to-medium average effect size.

2) DV will significantly moderate effect sizes. As different accounts of bilingual language control make different predictions about which DV an advantage should manifest on, no directional prediction was made.

3) Task may moderate effect sizes. It also may interact with DV. However, there is no strong theoretical justification for a directional hypothesis.

4) Samples of young children and older adults should exhibit larger bilingual advantages than samples of college students. This may manifest in a main effect of age or an interaction between age and DV.

5) Samples of early AoA participants should exhibit larger advantages than samples of later AoA participants.

6) There will be significant lab effects.
   a. If bilingual advantages are due to uncontrolled subject factors (e.g., proficiency) lab should interact with DV, such that lab effects will be pronounced for one DV but not for the other.
   b. If lab effects are the result of non-substantive differences in context and correlated background variables, there may be an interaction between DV and lab, but this interaction will not exhibit a clear pattern.
Method

**Literature search.** PsycINFO was searched periodically until July 2015. Search terms included a combination of *bilingual* or *bilingualism* with *executive control, executive function, inhibition* or *interference control*. Reference sections of the relevant studies and review articles (e.g., de Bruin, Treccani & Della Salla, 2015; Paap, Sawi & Johnson, 2015) were examined to identify additional studies for inclusion. Forty-three studies were included met the inclusion criteria for the analysis:

1. Study included at least one bilingual group. Because different studies use different measures as indicators of bilingualism (e.g., AoA, frequency of use, overall proficiency), it is impossible to identify a single definition of bilingualism. Therefore, a group of participants was designated as bilingual if any of the following were true: the age at which they began learning their L2 was equal to or less than half their age at the time of testing; participants reported near equivalent proficiency in both languages; participants reported native or near-native attainment in their L2; participants reported using each of their two languages in at least 40% of their daily activities; participants reported using both languages at home; participants reported using one language at home and one language at school (with the exception of children who were recently enrolled in immersion programs).

2. Study included at least one monolingual group, defined as participants with only minimal exposure to an L2, e.g., through foreign language classes at school.

3. Participants were at least five years old and without psychological impairment. We therefore excluded studies examining potential benefits of bilingualism as a protective factor in dementia (e.g., Bialystok, Craik & Freedman, 2007).
4. If demographic measures for the two groups were reported, the bilingual and monolingual groups did not significantly differ on any non-verbal measures. It is very common for bilinguals to exhibit smaller vocabularies and slower lexical access than monolinguals (for a review see Bialystok et al., 2009). Ideally, the analysis would include L2 proficiency as a moderator. However, this was untenable as the studies vary in (a) whether proficiency is measured at all, (b) whether proficiency was measured in the L1 or L2, and (c) the type of proficiency measure employed (lexical access, reading comprehension, vocabulary sizes, educational test score, etc.). Therefore, differences in linguistic processing, reported in some studies, had to be overlooked. However, some studies included bilingual groups at different SES levels than the monolingual groups (e.g., Carlson & Meltzoff, 2008). In these studies, bilingual advantages only emerged when controlling for SES and not when comparing raw scores. Therefore, they were excluded from the analysis.

5. It contained RT data from at least one interference-control task (e.g., flanker, Simon, Stroop). Interference control tasks were defined as follows: Participants were asked to make judgments about a visually presented target stimulus, with a second, distracter cue varying systematically across trials. On incongruent trials the distracter cue elicited a response different from the target cue; on congruent trials the distracter cue elicited the same response as the target. All responses were key presses and contained no additional manipulation of executive control demands. This criteria excluded studies if they contained an unusually small number of trials of either type (e.g., Costa et al., 2009, with 75% incongruent trials); if they used an unconventional response modality, such as monitoring eye movements or requiring a verbal response (e.g., Viswanthan & Bialystok,
2006: Study 1, in which participants completed a traditional anti-saccade that involves control of eye-movements); or if they included a separate EF manipulation beyond those standard to the task. This last point requires some elaboration. In a version of the Simon task used in Bialystok et al. (2004), participants memorize four color-response mappings rather than two. As this working-memory manipulation is viewed as a manipulation of EF, this version of the task is not included in the analysis. On the other hand, many studies use the standard or lateralized Attentional Networks Task, which contain manipulations aimed at attentional orienting and engaging. As these manipulations are by-design not executive control manipulations, these tasks are included in the analysis.

Data preparation and effect size calculation. The forty-three studies were further broken down into 84 comparisons. A comparison was defined as a comparison between a bilingual group and a monolingual group on a single task. If a study reported on several separate bilingual and monolingual groups, (e.g., varying in age, etc), each of these groups was included in a separate comparison. Care was taken to minimize the number of statistically dependent comparisons while maximizing the number of available effect sizes. First, if a study contained a single monolingual group and several bilingual groups and the bilingual groups did not differ on any moderator-variables (specifically AoA), the two bilingual groups were average together. If the two bilingual groups differed on a moderator variable, they were treated as separate comparisons. The same monolingual group was included in both comparisons, but its N was split in half across the two comparisons. For example, Luk, De Sa and Bialystok (2011) compared one monolingual group to early and late bilingual groups. Since the early and late bilingual groups differ on AoA, they were included in different comparisons. The same monolingual mean was
Figure 1. Schematic of data structure, using one example study. Straight arrows indicate nesting relationships, while curved arrows indicated assumed correlations. Paap and Sawi (2014) report on two tasks, the flanker task and the Simon task. Data from each task was included as a single comparison. The two comparisons were assumed to be independent. Within each comparison, interference costs and global RTs were calculated. These were assumed to be correlated.

included in both comparisons, but the monolingual $N$ for each comparison was half the $N$ in the study. Second, if a study reported multiple blocks on the same task, only the first block was included. Third, if multiple tasks were given to the same sample, each task counted as a separate comparison. This amounts to assuming that scores on these tasks are independent. As described in the introductions, the interference costs on interference-control tasks typically are not correlated, but the global RTs are. The assumption of independence is violated for global RTs
from studies that included multiple interference control tasks. Results should be interpreted with
some caution because of this violation.

Recall that interference costs are calculated at the difference between congruent and
incongruent trials and global RTs are calculated as the average across these two trial types. In
some studies, these scores and their standard deviations were reported. However in many cases,
means and standard deviations (or standard errors) were reported for the congruent and
incongruent trials, but not interference costs or global RTs. Because congruent and incongruent
trials are correlated, calculating the standard deviation for interference costs and global RTs from
these scores requires the correlation between trial types. Two strategies were employed to
estimate correlations.

First, in Bialystok et al. (2004), mean and standard deviations were reported for
congruent trials, incongruent trials and interference costs, but not global RTs. Given the standard
deviations of these three quantities and the assumption that these variables are normally
distributed, one can solve algebraically for the correlation between congruent and incongruent
trials. This correlation coefficient was used to calculate the standard deviations of global RTs.
This was the only study for which this procedure could be used.

Second, in many cases correlation coefficients had to be simulated. In order to do so, a
dataset of correlation coefficients was created from two sources. First, authors who were emailed
to provide other information for the analysis were asked to provide correlations between
congruent and incongruent trials. These eight correlation coefficients were combined with three
correlation coefficients from an unpublished dataset. A random effects meta-analysis was
conducted on these estimates in order to estimate a mean and standard deviation for the
correlation coefficients. The meta-analysis revealed an average correlation of .849 with
significant heterogeneity, $Q(10) = 31.96, p < .001$, and a standard deviation of .047. For every comparison missing a correlation coefficient, a coefficient was simulated from a normal distribution with a mean of .849 and a standard deviation of .047. Since there was significant heterogeneity among the observed correlation coefficients, and a large number of correlations were simulated, all analyses were conducted several times, assuming different correlation coefficients, to determine how sensitive results were to a particular correlation coefficient. These results are discussed in the Results section but, in general, the choice of correlation coefficients did not affect the qualitative pattern of results.

Five comparisons, all of which used the Simon Task, did not report standard deviations for any of congruent trials, incongruent trials or either aggregated score, and authors did not respond to requests for these data. The standard deviations were estimated via linear regression in the following manner. First a dataset with all the reported means and standard deviations for the Simon task was constructed. Second, a set of linear regressions were conducted, predicting standard deviations from means. Separate regressions were fit for the monolingual congruent, monolingual incongruent, bilingual congruent and bilingual incongruent RTs. Prior to fitting the regressions, four loess lines depicting the relationship between means and standard deviations, one for each of the aforementioned RT types, were examined for evidence of non-linearity. These plots revealed the relationship between mean and SDs for each of the four RT types was approximately linear with the exception of four data points. Martin Rhee and Bialystok (2008) Study 1 & Study 2 reported standard deviations much smaller than what would be predicted based on their means. Bialystok et al. (2004) reported extremely large standard deviations for both younger and older adults. In order to preserve a linear relationship, these four data points were dropped from models estimating standard deviations. Four linear regressions were then run
and each the intercept, slope and standard deviation of residuals were saved. Third, for each comparison missing a standard deviation, the standard deviation was simulated from a normal distribution regression equation for the reported mean with random error according to the standard deviation of that model’s residuals.

After computing missing standard deviations, Cohen’s $d$ was calculated for interference costs and global RTs for each of the 84 comparisons. This lead to a total of 168 effect sizes, nested within 84 comparisons, which were assumed to be independent. Figure 1 provides a schematic overview of the nature of the data and the assumed sources of dependence within the dataset.

**Moderator Coding.**

**Task.** Task was coded as a factor variable with four levels, Flanker, Simon, Stroop and Other. Flanker tasks included the traditional and alternating position Flanker task, as well as the Attentional Network Task (ANT). Most studies report mean congruent and incongruent RTs for the ANT collapsing across trials with different cueing and alerting manipulations. In these studies, interference costs and global RTs were calculated across these trial types. In a few studies, RTs were reported for the various cueing and alerting conditions separately. In these studies, interference costs and global RTs were calculated on the no-cue trials, as these are the conditions that most closely resemble the Flanker task outside the Attentional Network Test. Simon Arrows and Simon Color were both coded as Simon Tasks. Traditional Stroop color-word interference tasks were coded as Stroop tasks. Tasks coded as other include, the Global-Local Task, the lateralized Flanker task (in which arrows were presented vertically, rather than horizontally, on either the right or left side of the screen), and non-standard tasks that met the interference-control definition described in the inclusion criteria. All 84 comparisons were coded
as one of these four comparisons. Table 1 lists each task and how it was coded for all 84 comparisons.

**Age.** Since it was predicted that the largest effect sizes would be observed in children and older adults, age was coded categorically. As discussed in the introduction section, at least one lifespan study has demonstrated that interference costs on the Flanker task reach a ceiling between ages 16 and 42 (Waszak, Li & Hommel, 2010). It was, therefore, decided to code comparisons of participants younger than 13 as children and comparisons of participants between 18 and 40 as young adults. One comparison had an average age of 15 (Gathercole et al., 2014: Teenagers). This comparison was dropped in order to keep the children and adult groups as distinct as possible. Comparisons with participants over the age of 60 were coded as older adults. Several comparisons had average ages between 40 and 60. In order to include these data, these comparisons were coded as other adults, but no specific predictions were made about this group. A total of 83 comparisons were coded according to this scheme. Table 1 lists the age category for each comparison.

**AoA.** As discussed in the introduction, Sabourin, Brien and Burkholder (2014) found significant L2-to-L1 translation-equivalent priming among bilinguals with an early AoA, but not proficiency-matched bilinguals with a later AoA. The authors reasoned that, considering research on the putative critical period for L2 acquisition, 7 years old seemed like an appropriate age for distinguishing between early and late AoA. Given that this study provided psycholinguistic evidence that the lexicons of early and late AoA bilinguals differ, their cut-offs for early and late AoA were followed, and the comparisons were split into two groups: participants who began learning their second language by age 6 (early AoA), participants who began learning both
languages after age 8 (late AoA). Comparisons that did not fall into these groups were excluded from the analysis.

Discerning AoA from publications was challenging since research groups reported different information about their participants. There was no single rule that could be applied to every single study. Therefore, a set of rules was applied in a fixed order to determine AoA. If participants self-reported any of age of L2 acquisition, age of L2 immersion (or immigration), age of regular L2 exposure, or age of L2 use, the earliest of these values was selected and coded as either before 7 or after 7 (the few cases of exactly 7 were coded as NA). If none of these quantities was reported and the number of years of exposure of the L2 was reported, this number was subtracted from the participants’ age to estimate the average AoA and this number was coded. If participants did not report AoA, but the authors estimated AoA based on their knowledge of the community, this estimate was used and categorized as before or after 7. If the authors did not estimate a specific age, but noted that the participants began learning their L2 before or during school, the comparison was coded as early AoA, and if the authors noted that participants began learning their L2 during later grades, the comparison was coded as later AoA. If authors did not state a specific age of L2 acquisition/exposure or state when the L2 was learned relative to the onset of schooling, but described the bilinguals as early or later bilinguals, this characterization was adopted. If none of the prior criteria were met, but the authors noted that the bilingual participants spoke both languages at home and the bilinguals were not immigrants, the group was coded as early AoA. Finally any sample with an average age less than or equal to six years old was coded as early AoA.

Luk, De Sa and Bialystok (2011) reported age of exposure and age of regular exposure. In this case regular exposure was used, even though it was a little later.
Because this coding scheme was complicated, all comparisons were coded twice, once by
the author and once by a trained research assistant. The two coders agreed on 79 of the 84 initial
comparisons. The author reviewed the remaining 5 comparisons, identified the reason for
disagreement and made the final coding decision. On review, the causes for each of the five
disagreements were easily determined. Coding resulted in 54 early AoA comparisons, 18 late
AoA comparisons and 12 NA comparisons, including ages of 7 and comparisons for which an
AoA could not easily be discerned. Table 1 lists the value of AoA for each comparison.

**Lab.** Lab was coded as factor variable with 7 levels, using the following scheme. First,
the number of citations in the dataset per author was calculated. Then for each comparison, the
most cited author (the author who had the highest number of citations in the dataset) was noted.
Any authors that appeared as the most common author on 4 or more comparisons was treated as
a lab. Lab was coded as the university to reflect the fact that each lab contributed multiple first
authors. The six labs were FIU (Florida International University), OC (University of Ottawa and
Concordia University), NW (Northwestern University), SFSU (San Francisco State University),
UFRGS (Universidad Federal do Rio Grande do Sol), and York (York University). Comparisons
with a most common author that contributed fewer than four comparisons were coded as Other.

**Analytic Strategy.** As global RTs and interference cost are dependent, the primary
analyses were three-level meta-analyses, which were conducted using the *metasem* package in *R*
(Cheung, 2013). Unlike traditional random effects meta-analysis, which decomposes effect-size
variance into two sources, three-level meta-analysis decomposes variance into three sources:
sampling error, within-cluster variance, and between-cluster variance. Doing so explicitly
models dependence between observations within a cluster. In this study, clusters were defined as
comparisons, thereby modeling the dependence between global RTs and interference costs.
The multi-level structure of the data permits the calculation of two measures of explained variance. $R^2_{\text{within}}$ is the proportion of variation within clusters that is explained by a moderator and can be affected by either DV or the interactions of the DVs with between-cluster variables. $R^2_{\text{between}}$ is the proportion of variation between clusters that is explained by a moderator and can be affected by the between-cluster moderators and their interactions with the DV. Both of these numbers are reported for each of the models tested.

A set of 10 models was fit to examine the effects of the five moderators and their predicted interactions. First, a null model, with no moderators was fit to the data to serve as a baseline. Second, a series of 9 models was fit to test the effects of predicted main effects and interactions (see Table 2). Model 1 included just the effect of the DV. Models 2-9 tested the effect of each of the additional moderators described above. For each additional moderator, two models were fit, one with only an additive effect of the moderator variable, and one with an interaction between DV and that moderator variable. All moderator variables were dummy coded.

Some effect sizes were exceptionally large (see Table 1). These outlier effect sizes might exert a strong influence on models, but there is no principled method for excluding them from analyses. Therefore, a sensitivity analysis was conducted for each model. Specifically each model was fit $K-1$ times, where $K$ equals the number of comparisons. On each iteration, one of the comparisons was dropped. These ranges illustrate how sensitive estimates are to the presence of particular effect sizes. In order to facilitate communication, these sensitivity ranges are reported for only significant effects.

Statistical inference was tested using two methods. First, the overall fit of each model was evaluated using a likelihood ratio test. All models were compared to the null model. Models
2-9 were also compared to model 1. All models with interactions were compared to the relevant model with just additive effects. Second, since moderators were included as factor variables, slopes for each factor were examined.

Given the evidence of publication bias documented by de Bruin et al (2015), it was necessary to test whether significant moderator effects could be due to publication bias. This could occur if publication bias was stronger at one level of a moderator variable than another. To test this, for every significant comparison, funnel plots were compared for each level of the comparison. In funnel plots, the X axis represents effect sizes and the Y axis represents standard errors. Funnel plots assume that large-\(N\), low variance studies are likely to be published regardless of whether they produce statistically significant effects and that if publication bias is present small-\(N\), high variance studies, are only likely to get published if they produce significant effects. Therefore, if publication bias is not present, effect sizes from each study should be symmetrically distributed around the estimated average effect size. However, if publication bias is present, there should be a negative relationship between variance and effect size, with low-variance effect sizes near, or below, the estimated average effect size and high-variance effect sizes only above the estimated average effect size. If funnel plots suggest publication bias at one level of a moderator, but not the other, this differential publication bias may be the cause of the significant moderator.

After all models were fit, a second sensitivity analysis was conducted to test the sensitivity of results to assumptions about the correlation between congruent and incongruent trials. Several alternative datasets were created. These datasets assumed different means and standard deviations for the distribution of simulated correlation coefficients. The entire set of
Table 1

**Effect sizes and moderators for each of the 84 comparisons for Interference Control Tasks.**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Sample Size</th>
<th>Moderators</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Global RT</td>
</tr>
<tr>
<td>Abutalebi et al 2012</td>
<td>N Bil</td>
<td>N Mon</td>
<td>Task</td>
</tr>
<tr>
<td>Anton et al 2014</td>
<td>17</td>
<td>14</td>
<td>Flanker</td>
</tr>
<tr>
<td>Bialystok et al 2004</td>
<td>180</td>
<td>180</td>
<td>Flanker</td>
</tr>
<tr>
<td>Study 1 Young</td>
<td>10</td>
<td>10</td>
<td>Simon</td>
</tr>
<tr>
<td>Study 1 Old</td>
<td>10</td>
<td>10</td>
<td>Simon</td>
</tr>
<tr>
<td>Study 2 (2 Squares) Young</td>
<td>32</td>
<td>32</td>
<td>Simon</td>
</tr>
<tr>
<td>Bialystok et al 2004</td>
<td>10</td>
<td>10</td>
<td>Simon</td>
</tr>
<tr>
<td>Study 3</td>
<td>19</td>
<td>10</td>
<td>Simon</td>
</tr>
<tr>
<td>Bialystok and Depape 2009</td>
<td>19</td>
<td>17</td>
<td>Simon</td>
</tr>
<tr>
<td>Players Arrows</td>
<td>30</td>
<td>31</td>
<td>Simon</td>
</tr>
<tr>
<td>Non-Players Arrows</td>
<td>24</td>
<td>24</td>
<td>Simon</td>
</tr>
<tr>
<td>Bialystok et al 2008</td>
<td>24</td>
<td>24</td>
<td>Stroop</td>
</tr>
<tr>
<td>Young Stroop</td>
<td>24</td>
<td>24</td>
<td>Stroop</td>
</tr>
<tr>
<td>Old Simon</td>
<td>24</td>
<td>24</td>
<td>Simon</td>
</tr>
<tr>
<td>Study</td>
<td>N1</td>
<td>N2</td>
<td>Type</td>
</tr>
<tr>
<td>------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>----------</td>
</tr>
<tr>
<td>Bialystok 2010</td>
<td></td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Study 1</td>
<td>26</td>
<td>25</td>
<td>Other</td>
</tr>
<tr>
<td>Bialystok 2010</td>
<td></td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Study 2</td>
<td>25</td>
<td>25</td>
<td>Other</td>
</tr>
<tr>
<td>Bialystok 2010</td>
<td></td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Study 3</td>
<td>25</td>
<td>25</td>
<td>Other</td>
</tr>
<tr>
<td>Billig and Scholl 2011</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Middle Aged Simon Arrows</td>
<td>20</td>
<td>21</td>
<td>Simon</td>
</tr>
<tr>
<td>Billig and Scholl 2011</td>
<td></td>
<td></td>
<td>Stroop</td>
</tr>
<tr>
<td>Middle Aged Stroop</td>
<td>20</td>
<td>21</td>
<td>Stroop</td>
</tr>
<tr>
<td>Billig and Scholl 2011</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Elderly Simon Arrows</td>
<td>21</td>
<td>21</td>
<td>Simon</td>
</tr>
<tr>
<td>Billig and Scholl 2011</td>
<td></td>
<td></td>
<td>Stroop</td>
</tr>
<tr>
<td>Elderly Stroop</td>
<td>21</td>
<td>21</td>
<td>Stroop</td>
</tr>
<tr>
<td>Blumenfeld and Marian 2011</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Stroop</td>
<td>30</td>
<td>30</td>
<td>Other</td>
</tr>
<tr>
<td>Blumenfeld and Marian 2011</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Simon</td>
<td>30</td>
<td>30</td>
<td>Other</td>
</tr>
<tr>
<td>Blumenfeld and Marian 2014</td>
<td></td>
<td></td>
<td>Stroop</td>
</tr>
<tr>
<td>Study 2 Stroop</td>
<td>60</td>
<td>60</td>
<td>Stroop</td>
</tr>
<tr>
<td>Blumenfeld and Marian 2014</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Study 2 Simon</td>
<td>60</td>
<td>60</td>
<td>Simon</td>
</tr>
<tr>
<td>Calvo and Bialystok 2014</td>
<td></td>
<td></td>
<td>Flanker</td>
</tr>
<tr>
<td>Working Class</td>
<td>44</td>
<td>20</td>
<td>Flanker</td>
</tr>
<tr>
<td>Calvo and Bialystok 2014</td>
<td></td>
<td></td>
<td>Flanker</td>
</tr>
<tr>
<td>Middle Class</td>
<td>65</td>
<td>46</td>
<td>Flanker</td>
</tr>
<tr>
<td>Coderre et al 2013</td>
<td></td>
<td></td>
<td>Flanker</td>
</tr>
<tr>
<td>English L1</td>
<td>37</td>
<td>23</td>
<td>Stroop</td>
</tr>
<tr>
<td>Coderre and van Heuven 2014</td>
<td></td>
<td></td>
<td>Stroop</td>
</tr>
<tr>
<td>Stroop</td>
<td>58</td>
<td>18</td>
<td>Stroop</td>
</tr>
<tr>
<td>Coderre and van Heuven 2014</td>
<td></td>
<td></td>
<td>Simon</td>
</tr>
<tr>
<td>Simon</td>
<td>58</td>
<td>18</td>
<td>Simon</td>
</tr>
<tr>
<td>Costa et al 2008</td>
<td>100</td>
<td>100</td>
<td>Flanker</td>
</tr>
<tr>
<td>Costa et al 2009</td>
<td>31</td>
<td>31</td>
<td>Flanker</td>
</tr>
<tr>
<td>Costa et al 2009</td>
<td></td>
<td></td>
<td>Stroop</td>
</tr>
<tr>
<td>Cottini et al 2014</td>
<td></td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Study</td>
<td>Group 1</td>
<td>Group 2</td>
<td>Task</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>---------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>Dunabeita et al 2014</td>
<td>252</td>
<td>252</td>
<td>Stroop</td>
</tr>
<tr>
<td>Dunabeita et al 2014</td>
<td>252</td>
<td>252</td>
<td>Numerical Stroop</td>
</tr>
<tr>
<td>Emmorey et al 2009</td>
<td>15</td>
<td>15</td>
<td>Flanker</td>
</tr>
<tr>
<td>Heidlmayr et al 2014</td>
<td>33</td>
<td>31</td>
<td>Stroop</td>
</tr>
<tr>
<td>Kapa and Colombo 2013</td>
<td>21</td>
<td>11</td>
<td>Flanker</td>
</tr>
<tr>
<td>Kapa and Colombo 2013</td>
<td>36</td>
<td>11</td>
<td>Flanker</td>
</tr>
<tr>
<td>Kirk et al 2014</td>
<td>16</td>
<td>16</td>
<td>Simon</td>
</tr>
<tr>
<td>Kirk et al 2014</td>
<td>16</td>
<td>16</td>
<td>Simon</td>
</tr>
<tr>
<td>Gaelic-English</td>
<td>16</td>
<td>16</td>
<td>Simon</td>
</tr>
<tr>
<td>Kousaie and Phillips 2012</td>
<td>26</td>
<td>25</td>
<td>Stroop</td>
</tr>
<tr>
<td>Kousaie and Phillips 2012</td>
<td>26</td>
<td>25</td>
<td>Simon</td>
</tr>
<tr>
<td>Kousaie and Phillips 2012</td>
<td>26</td>
<td>25</td>
<td>Flanker</td>
</tr>
<tr>
<td>Kousaie et al 2014</td>
<td>36</td>
<td>61</td>
<td>Simon</td>
</tr>
<tr>
<td>MartinRhee and Bialystok 2008</td>
<td>51</td>
<td>70</td>
<td>Simon</td>
</tr>
<tr>
<td>MartinRhee and Bialystok 2008</td>
<td>17</td>
<td>17</td>
<td>Simon</td>
</tr>
<tr>
<td>MartinRhee and Bialystok 2008</td>
<td>20</td>
<td>21</td>
<td>Simon</td>
</tr>
<tr>
<td>Marzecova et al 2013</td>
<td>18</td>
<td>17</td>
<td>Flanker</td>
</tr>
<tr>
<td>Gathercole et al 2014</td>
<td>49</td>
<td>60</td>
<td>Simon</td>
</tr>
<tr>
<td>Gathercole et al 2014</td>
<td>48</td>
<td>33</td>
<td>Simon</td>
</tr>
<tr>
<td>Gathercole et al 2014</td>
<td>23</td>
<td>20</td>
<td>Simon</td>
</tr>
<tr>
<td>Study</td>
<td>Participants</td>
<td>Task</td>
<td>Condition</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Young Adults</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gathercole et al 2014</td>
<td>17</td>
<td>20</td>
<td>Simon</td>
</tr>
<tr>
<td>Luk et al 2011</td>
<td>43</td>
<td>19</td>
<td>Flanker</td>
</tr>
<tr>
<td>Luk et al 2011</td>
<td>42</td>
<td>19</td>
<td>Flanker</td>
</tr>
<tr>
<td>Luk et al 2010</td>
<td>10</td>
<td>10</td>
<td>Flanker</td>
</tr>
<tr>
<td>Mor et al 2014</td>
<td>20</td>
<td>20</td>
<td>Other</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 1 Simon</td>
<td>34</td>
<td>45</td>
<td>Simon</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 2 Simon</td>
<td>36</td>
<td>50</td>
<td>Simon</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 3 Simon</td>
<td>51</td>
<td>55</td>
<td>Simon</td>
</tr>
<tr>
<td>Flanker</td>
<td>49</td>
<td>55</td>
<td>Flanker</td>
</tr>
<tr>
<td>Paap and Sawi 2014 Flanker</td>
<td>58</td>
<td>62</td>
<td>Flanker</td>
</tr>
<tr>
<td>Paap and Sawi 2014 Simon</td>
<td>58</td>
<td>62</td>
<td>Simon</td>
</tr>
<tr>
<td>Engel de Abreu et al 2012</td>
<td>40</td>
<td>40</td>
<td>Flanker</td>
</tr>
<tr>
<td>Poarch and Bialystok 2015</td>
<td>30</td>
<td>15</td>
<td>Flanker</td>
</tr>
<tr>
<td>Poarch and Bialystok 2015</td>
<td>99</td>
<td>60</td>
<td>Flanker</td>
</tr>
<tr>
<td>Study</td>
<td>Age Group</td>
<td>Simon</td>
<td>Condition</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-----------</td>
<td>-------</td>
<td>-----------</td>
</tr>
<tr>
<td>Poarch and VanHell 2012</td>
<td>Early</td>
<td>Simon</td>
<td>Early Children</td>
</tr>
<tr>
<td>Rubio-Fernandez and Glucksberg 2012</td>
<td>Early</td>
<td>Simon</td>
<td>Early Young Adult</td>
</tr>
<tr>
<td>Salvatierra and Rosselli 2010 Young</td>
<td>Late</td>
<td>Simon</td>
<td>Late Young Adult</td>
</tr>
<tr>
<td>Salvatierra and Rosselli 2010 Old</td>
<td>Late</td>
<td>Simon</td>
<td>Late Older Adult</td>
</tr>
<tr>
<td>Schroeder and Marian 2012</td>
<td>Late</td>
<td>Simon</td>
<td>Late Older Adult</td>
</tr>
<tr>
<td>Tao et al 2011 early</td>
<td>Early</td>
<td>Other</td>
<td>Early Young Adult</td>
</tr>
<tr>
<td>Tao et al 2011 late</td>
<td>NA</td>
<td>Other</td>
<td>NA Young Adult</td>
</tr>
<tr>
<td>Yang Yang and Lust 2011</td>
<td>Late</td>
<td>Flanker</td>
<td>Late Children</td>
</tr>
<tr>
<td>Yudes et al 2011</td>
<td>NA</td>
<td>Simon</td>
<td>NA Young Adult</td>
</tr>
</tbody>
</table>
models was then re-fit to each of these dataset to determine whether conclusions differed from those made in the original study.

Cohen’s (1992) method for interpreting effect sizes was used throughout. For $d$, .2 was considered small, .5 was considered medium and .8 was considered large. $R^2$s were converted to $f^2$ and .02 was considered small, .15 medium and .35 large.

Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Moderators</th>
<th>$R^2_{\text{within}}$</th>
<th>$R^2_{\text{between}}$</th>
<th>$P_{\text{null}}$</th>
<th>$P_{\text{model1}}$</th>
<th>$P_{\text{addon1}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td>DV</td>
<td>.00</td>
<td>.00</td>
<td>.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>DV + Task</td>
<td>.00</td>
<td>.02</td>
<td>.62</td>
<td>.47</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>DV * Task</td>
<td>.07</td>
<td>.01</td>
<td>.55</td>
<td>.46</td>
<td>.36</td>
</tr>
<tr>
<td>Model 4</td>
<td>DV + Age</td>
<td>.00</td>
<td>.28</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Model 4b</td>
<td>DV + Age</td>
<td>.00</td>
<td>.17</td>
<td>.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Model 5</td>
<td>DV * Age</td>
<td>.19</td>
<td>.25</td>
<td>.02</td>
<td></td>
<td>.02</td>
</tr>
<tr>
<td>Model 5b</td>
<td>DV * Age</td>
<td>.03</td>
<td>.17</td>
<td>.38</td>
<td>.21</td>
<td>.11</td>
</tr>
<tr>
<td>Model 6</td>
<td>DV + AoA</td>
<td>.00</td>
<td>.02</td>
<td>.38</td>
<td>.21</td>
<td></td>
</tr>
<tr>
<td>Model 7</td>
<td>DV * AoA</td>
<td>.29</td>
<td>.00</td>
<td>.003</td>
<td>.001</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Model 8</td>
<td>DV + Lab</td>
<td>.00</td>
<td>.39</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td></td>
</tr>
<tr>
<td>Model 9</td>
<td>DV * Lab</td>
<td>.34</td>
<td>.31</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Notes. Models with interactions are compared to the null model, model 1, and the relevant additive model.

Results

Prior to model-fitting, random-effects meta-analyses were conducted for interference costs and global RTs separately in order to create forest plots. The average effect size for interference costs was .24 ($CI: .12 : .36$), was statistically significant ($p < .001$), and, as evident in Figure 2, exhibited significant heterogeneity, $Q(83) = 324.58, p < .001$. The average effect size of global RTs was .34 ($CI: .14 : .53$) and was also statistically significant ($p < .001$). As is evident in Figure 3, it also exhibited significant heterogeneity, $Q(83) = 459.59, p < .001$. Both Figures 2 and 3 reveal significant heterogeneity, and outlier effect sizes, some of which are very large, necessitating the sensitivity analyses described in the last section.
The first model of interest was the null model, a three-level meta-analysis with no predictors. This model served two purposes: First it provided an aggregate estimate of the size and variability of the effect size across both interference costs and global RTs. Second, it served as the baseline model to compare with the subsequent models with additional moderators. The overall effect size was .29 (CI: .15 : .44) and was statistically significant ($p < .001$).
Sensitivity analyses yielded a range of effect sizes from .25 to .30. There was also significant variability both within comparisons ($\tau^2_{\text{within}} = .11, p = .001$) and between studies ($\tau^2_{\text{between}} = .33, p < .001$). Models 1-9 were compared to the null model. Table 2 contains global comparisons between models 1-9 and the null model. Table 3 contains regression parameters for each model.
### Table 3

**Regression Coefficients from Moderator Models**

<table>
<thead>
<tr>
<th>Task</th>
<th>Age</th>
<th>AoA</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 2</td>
<td>Model 3</td>
<td>Model 4</td>
</tr>
<tr>
<td>Intercept</td>
<td>.19</td>
<td>.11</td>
<td>.12</td>
</tr>
<tr>
<td>DV</td>
<td>-.04</td>
<td>.13</td>
<td>-.03</td>
</tr>
</tbody>
</table>

**Task**: Other Category: Baseline

- Simon: .20 | .30
- Flanker: .02 | .06
- Other: -.01 | .22
- DV*Simon: -.19
- DV*Flanker: -.08
- DV*Other: -.44*

**Age**: Baseline Category: Young Adults.

- Children: .18 | .18 | .34* | .35*
- O. Adult: 1.4** | 1.8** | .72**
- Elderly: .31 | .65** | .26
- DV*: -.32* | -.34*
- Children: -.76* | -.14
- Adult: DV*: .08

**AoA**: Baseline Category: Early

- Later AoA: -.23 | -.51*

**Lab**: Baseline Category: Other

- FIU | - .26 | -.17
- NW | - .13 | -.08
- OC | - .24 | -.30
- SFSU | - .30 | -.26
- UFRGS | .04 | .16
- York | .53** | .79**
- FIU * DV | - .18
- NW * DV | - .09
- OC * DV | .11
- SFSU * DV | .07
- UFRGS * | - .24
- York | - .53**

**Variance Components**

<table>
<thead>
<tr>
<th>Tau2 2</th>
<th>.10**</th>
<th>.09**</th>
<th>.10**</th>
<th>.12**</th>
<th>.08**</th>
<th>.12**</th>
<th>.13**</th>
<th>.09*</th>
<th>.10**</th>
<th>.07**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau2 3</td>
<td>.30**</td>
<td>.30**</td>
<td>.22**</td>
<td>.26***</td>
<td>.23**</td>
<td>.27***</td>
<td>.36**</td>
<td>.37**</td>
<td>.19**</td>
<td>.21**</td>
</tr>
</tbody>
</table>

**Note.**

- \( a = .05 < p < .10 \)
- \( * = .01 < p < .05 \)
- \( ** = p < .01 \)
Model 1 was a three-level meta-analysis and contained DV as a moderator. According to a likelihood ratio test, model 1 did not fit significantly better than the null model ($p = .85$; see also Table 2). Consistent with the likelihood ratio test, the coefficient for DV was small and non-significant ($B = -.02$, CI: $-.15 : .12$, $Z = -.28$, $p = .78$). Including DV as a moderator accounted for very little variance within comparisons, $R^2_{\text{within}} < .001$. While the main effect of DV was non-significant, it is possible that DV would interact with another moderator, and its effect might only be detectable then. Therefore, DV was included in all subsequent models.

**The effect of task.** Model 2 included main effects for Task (with Stroop as the baseline level) and DV. According to likelihood ratio tests, this model did not fit significantly better than the null model ($p = .64$) or model 1 ($p = .48$). Slopes were non-significant for the Simon task, Flanker task and other tasks (See Table 3). Including task as a moderator also accounted for very little variability across comparisons ($R^2_{\text{between}} < .001$). Meta-analyses were conducted on each task separately in order to estimate average effect sizes for each task. Because of the small number of comparisons in several task categories, three-level meta-analyses did not always converge, so estimates from traditional random-effects meta-analysis are reported instead. As a result, confidence intervals for these estimates are approximate. Estimates are reported in Table 4.

Model 3 tested for an interaction between DV and task. A likelihood ratio test revealed that it did not differ significantly from model 2 ($p = .33$), model 1 ($p = .44$) or the null model ($p = .55$). Slopes were non-significant for Simon, Flanker, Other, Simon*DV, Flanker*DV, Other*DV. To estimate effect sizes for each DV in each task, eight random effects meta-analyses were conducted. Effect sizes for each DV within task are also reported in Table 4.
Table 4

*Estimated Effect Sizes and Confidence Intervals for Each Task.*

<table>
<thead>
<tr>
<th>Task</th>
<th>N</th>
<th>Overall</th>
<th>Interference Cost</th>
<th>Global RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simon</td>
<td>44</td>
<td>.46</td>
<td>.28</td>
<td>.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23 : 0.70)</td>
<td>(0.08 : 0.48)</td>
<td>(0.09 : 0.80)</td>
</tr>
<tr>
<td>Flanker</td>
<td>21</td>
<td>.18</td>
<td>.21</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.07 : 0.28)</td>
<td>(0.06 : 0.35)</td>
<td>(-0.02 : 0.32)</td>
</tr>
<tr>
<td>Stroop</td>
<td>10</td>
<td>.13</td>
<td>.16</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-0.03 : 0.30)</td>
<td>(-0.02 : 0.34)</td>
<td>(-0.02 : 0.23)</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>.12</td>
<td>-.01</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02 : 0.21)</td>
<td>(-0.14 : 0.12)</td>
<td>(0.03 : 0.57)</td>
</tr>
</tbody>
</table>

*Confidence Intervals in parentheses. All estimates are from traditional random-effects meta-analysis.*

**Tests of age.** Model 4 included effects for age and DV. Age was coded as a factor. Because it was predicted that children and older adults would differ from young adults, young adults was the reference level. Because only 83 of the 84 comparisons were included in this analysis, the null model and model 1 were re-fit to this subset of the data. According to a likelihood ratio test, model 4 significantly improved fit relative to both the null model and model 1 (both p’s < .001). Unexpectedly, the children did not differ significantly from the young adult group (B = .16, CI = -.07 : .39, Z = 1.35, p = .18), and neither did the older adult group (B = .30, CI = -.08 : .68, Z = .47). However, the other adult group differed significantly from the young adult group (B = 1.45, CI = .88 : 2.10, Z = 5.00, p < .001, B_{range} = 1.07 : 1.78), with age accounting for a large amount of variance between comparisons ($R^2_{between} = .27$, range = .11 : .34). As seen in Table 5, this effect was driven by a very large effect size in the other adult group (d = 2.20, CI = .79 : 3.6), and smaller effect sizes for older adults (d = .44, CI = .03 : .85), children (d = .32, CI = .09 : .56), young adults (d = .08, CI = -.01 : .17). Because the effect of age was statistically significant, funnel plots were produced for each of the four age groups to detect for the presence of differential publication bias. Figure 4 shows strong evidence of
publication bias in comparisons involving other and older adults, with a few large effect sizes with very large variances. Additionally, there is moderate evidence for publication bias in comparisons involving children, but little in comparisons involving young adults.

Table 5

Estimated Effect Sizes and Confidence Intervals for Each Age Group

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>Effect Sizes</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Overall</td>
<td>Interference Cost</td>
<td>Global RT</td>
</tr>
<tr>
<td>Children</td>
<td>22</td>
<td>.32</td>
<td>.16</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.09 : .59)</td>
<td>(-.11 : .32)</td>
<td>(.11 : .92)</td>
</tr>
<tr>
<td>Young Adults</td>
<td>43</td>
<td>.08</td>
<td>.14</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-.01 : .17)</td>
<td>(.02 : .23)</td>
<td>(-.08 : .15)</td>
</tr>
<tr>
<td>Other Adults</td>
<td>7</td>
<td>2.20</td>
<td>2.07</td>
<td>2.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.79 : 3.6)</td>
<td>(-.21 : 4.4)</td>
<td>(.43 : 4.3)</td>
</tr>
<tr>
<td>Older Adults</td>
<td>11</td>
<td>.44</td>
<td>.48</td>
<td>.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.03 : .85)</td>
<td>(.11 : .85)</td>
<td>(-.37 : 1.21)</td>
</tr>
<tr>
<td>Older + Other Adults</td>
<td>18</td>
<td>.92</td>
<td>1.03</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(.41 : 1.42)</td>
<td>(.16 : 1.91)</td>
<td>(.29 : 1.3)</td>
</tr>
</tbody>
</table>

Confidence Intervals in parentheses.
All estimates are from traditional random-effects meta-analysis.
The difference between the other adult group and young adult group was surprising. This may be due to the fact that one or two very large effect sizes were grouped with a small number of effect sizes. Because the cut-off between the older and other adult group was arbitrary, an additional model was fit in which these two groups were collapsed together. Model 4a grouped other and older adults in a single older adults category. This model fit significantly better than both the null model ($p = .001$) and model 1 ($p < .001$). This effect was driven by the older adult group ($B = -.65, Z = 3.7, p < .001$). The range of $B$ statistics was .46 to .71, suggesting that dropping one effect size reduced the coefficient by 30%. This model accounted for a medium amount of variation between comparisons ($R^2_{between} = .17$, range = .10 : .21). Table 4 also presents the effect sizes for each of the three age groups. As can be seen there was a large average effect size for the older adults group ($d = .93, CI = .42 : 1.45$). The merits of model 4a to model 4 will be discussed in the interim discussion. Both are reported here for thoroughness.

![Figure 5: Funnel Plots for Recoded Age Variable](image)

Model 5 included the interaction between DV and age. According to a likelihood ratio test, model 5 fit better than model 4 ($p < .001$). Examination of the slopes revealed that children had significantly larger global RT effect sizes than young adults ($B = .35, CI = .01 : .68, Z = 2.27, p = .02$, range $B = .23 : .35$) and that other adults had significantly larger global RT effect
sizes than young adults ($B = 1.83$, $CI = 1.17 : 2.49$, $Z = 5.50$, $p < .001$; range $B = 1.22 : 2.3$).

The interaction between the young adult factor and DV was significant ($B = -.31$, $CI = -.01 : -.61$, $Z = -2.00$, $p = .04$, $B$ range = -.36 : -.20), as was the interaction between the other adult factor and DV ($B = -.75$, $CI = -1.40 : -.01$, $Z = -2.30$, $p = .02$, $B$ range = -1.06 : -.30). Model 5 accounted for a medium amount of variance within comparisons ($R^{2}_{\text{within}} = .17$, range = .12 : .25) and a large amount between comparisons ($R^{2}_{\text{between}} = .25$, range = .11 : .31).

In order to understand the interactions, separate random effects meta-analyses were fit to the interference costs and global RTs separately. The model of the interference costs revealed that effect sizes were larger for older adults than young adults ($B = .34$, $CI = .01 : .67$, $Z = 2.02$, $p = .04$; range $B = .17 : .38$) and other adults than young adults ($B = .87$, $CI = .11 : .67$, $Z = 3.38$, $p < .001$; range $B = .64 : 1.16$) but no difference between children and young adults. The model of global RTs revealed effect sizes were significantly larger for children than young adults ($B = .43$, $CI = -.00 : .85$, $Z = 1.95$, $p = .05$; range $B = .30 : .47$) and for other adults than young adults ($B = 1.9$, $CI = 1.1 : 2.7$, $Z = 4.6$, $p < .001$; range $B = 1.20 : 2.31$), which was consistent with the coefficients in the full interaction model, but no difference between older and young adults.

Table 5 contains effect sizes for each age group. Consistent with the results from model 5, the other adult age group had extremely large effect sizes for both global RT ($d = 2.38$, $CI = .43 : 4.3$) and interference costs ($d = 2.07$, $CI = -.21 : 4.4$). Funnel plots, in Figure 6, revealed strong evidence for publication bias for both of these conditions. The children group had a small, non-significant effect size for interference cost ($d = .16$, $CI = -.11 : .32$), but a medium effect size for global RT ($d = .51$, $CI = .11 : .92$). Funnel plots revealed stronger evidence of publication bias amongst global RTs than interference costs. The young adult group had small effect sizes for both interference costs ($d = .14$, $CI = .02 : .23$) and global RTs ($d = .04$, $CI = -.08 : .15$).
Funnel plots revealed little evidence of publication bias among either global RTs or interference costs. The older adult group exhibited a small-to-medium effect size for interference costs ($d = 48, CI = .11 : .85$) and small-to-medium effect sizes for global RTs ($d = .42, CI = -.37 : 1.21$). Funnel plots revealed little evidence for publication bias amongst interference costs and some evidence of publication bias amongst global RTs.

Figure 6: Funnel Plots for Interference Costs and Global RTs for Each Age Group
As with model 4, a second analysis with other and older groups collapsed into a single other group category was conducted. This model, model 5a did not improve fit relative to the main-effects only model ($p = .11$). As was the case with model 5, examination of the coefficients revealed that the children exhibited significantly larger effect sizes on global RTs than did the older adults ($B = -.35, Z = 1.98, p = .05, B$ range = .22 : .35). The older adults exhibited significantly larger effect sizes than the young adults on global RTs ($B = .71, Z = 3.6, p < .001$). The interaction between DV and child was significant ($B = -.33, Z = -2.2, p = .03$), suggesting that the difference between children and young adults was smaller for interference costs than for global RTs.

In order to understand the interaction, separate random effects meta-analyses were fit to the interference costs and global RTs. The model of the interference costs revealed that effect sizes were larger for older adults than young adults ($B = .48, CI = .18 : .79, Z = 3.04, p = .02$), and that there was no difference between children and young adults ($B = -.03, CI = -29 : .24, Z = 3.14, p = .84$). The model of global RTs revealed a trend for significantly larger effect sizes among children than young adults ($B = .44, CI = -.02 : .89, Z = 1.88, p = .06$) and for older adults had significantly larger effect sizes than young adults ($B = .78, CI = .27 : 1.3, Z = 3.04, p = .002$). Funnel plots in Figure 6 show that for the children group, there is strong evidence of publication bias amongst the global RTs and for the older adult group there is strong evidence of publication bias amongst both global RTs and interference costs.

**Age of Acquisition.** Model 6 included effects for AoA and DV. In both model 6 and model 7, AoA was a binary variable and early AoA served as the reference level. Because only 65 of the 84 studies were coded for AoA, the null model and model 1 were re-fit to these 65 cases. A likelihood ratio test revealed that including age of acquisition did not
significantly improve fit relative to the null model \((p = .45)\) or model 1 \((p = .23)\). Consistent with this, the slope for age of acquisition was also non-significant \((B = -.20, CI = -.54 : .13, Z = -1.19, p = .24, B \text{ range} = -.24 : -.15)\). Age of acquisition accounted for a very small amount of variation between comparisons \((R^2_{\text{between}} = .02)\). Table 6 contains effect sizes for the early and late age of acquisition groups separately. The early AoA group had a medium to large effect size \((d = .39, CI = .22 : .57)\) and the later AoA group had a small effect size \((d = .13, CI = .03 : .23)\).

Model 7 included the interaction between AoA and DV. Likelihood ratio tests revealed that the model 7 fit significantly better than model 6 \((p = .001)\), model 1 \((p = .015)\) and the null model \((p = .035)\). Examination of the slopes revealed that the later AoA group had significantly smaller effect sizes for global RTs than the early AoA group \((B = -.51, CI = -.92 : -.11, Z = -2.49, p = .01; B \text{ range} = -.53 : -.42)\). Additionally, the difference between interference costs and global RTs was larger for the later AoA group than early AoA group \((B = .55, CI = .27 : .90, Z = 2.7, p < .01; B \text{ range} = .49 : .59)\). Including the interaction accounted for a moderate amount of variability within comparisons \((R^2_{\text{within}} = .27)\). In order to understand the interaction, random effects meta-analyses were conducted for the interference costs and global RTs separately.

**Figure 7: Funnel Plots for Recoded Age Variables**
Amongst interference costs, there was no significant difference between early and later AoA comparisons ($B = -.58, CI = -1.1 : -.06$, $Z = -2.2, p = .03$). Amongst global RTs there was no difference between early and late AoA comparisons ($B = -.57, CI = -1.1 -.06, Z = -2.20 , p = .03$).

Table 6 contains effect sizes for the interference costs and global RTs for the early and late AoA studies. The early AoA studies reported a medium effect size for interference costs ($d = .25, CI = .07 : .43$) and a large effect size for global RTs ($d = .56, CI = .25 : .88$), with examination of funnel plots revealing strong evidence of publication bias among both the interference costs and global RTs. The late AoA studies reported a medium effect size for interference costs ($d = .36, CI = .20 : .53$) and a small negative effect size for global RTs ($d = -.05, CI = -.17 : .07$), with examination of the funnel plots revealing little evidence for publication bias among either the interference costs or global RTs.

![Figure 7: Funnel Plots for Interference Costs and Global RTs by Age of Acquisition](image)
Table 6

*Estimated Effect Sizes and Confidence Intervals for Each AoA Group.*

<table>
<thead>
<tr>
<th>AoA</th>
<th>N</th>
<th>Overall</th>
<th>Interference Cost</th>
<th>Global RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Age of</td>
<td>55</td>
<td>.39</td>
<td>.25</td>
<td>.56</td>
</tr>
<tr>
<td>Acquisition</td>
<td></td>
<td>(.22 : .57)</td>
<td>(.07 : .43)</td>
<td>(.25 : .88)</td>
</tr>
<tr>
<td>Later Age of</td>
<td>18</td>
<td>.13</td>
<td>.36</td>
<td>-.05</td>
</tr>
<tr>
<td>Acquisition</td>
<td></td>
<td>(.03 : .23)</td>
<td>(.20 : .53)</td>
<td>(-.17 : .07)</td>
</tr>
</tbody>
</table>

*Confidence Intervals in parentheses.*

*All estimates are from traditional random-effects meta-analysis.*

The effect of lab. Model 8 included lab as a moderator variable. Lab was included as a factor variable with the other category serving as the baseline. For terminological clarity, this group will be referred to as the baseline group. According to a likelihood ratio test, model 8 fit significantly better than the null model (p < .001) and model 1 (p < .001). Examination of the coefficients revealed that effect sizes for the York lab differed significantly from those in the baseline group ($B = .60, CI = .31 : .90, Z = 4.00, p < .001; \text{range } B = .45 : .67$). None of the other labs differed significantly from the baseline group. Including lab accounted for a substantial amount of the variation between comparisons ($R^2_{between} = .39; \text{range } B = .36 : .42$).

Table 7 lists the overall effect sizes for each of the 7 lab conditions. The York lab reported a large effect size overall ($d = .98, CI = .55 : 1.41$). The baseline labs reported a smaller effect size with confidence intervals that did not contain 0 ($d = .17, CI = .09 : .26$). All the other effect sizes were small with confidence intervals that contained zero.

Model 8 included the interaction between lab and DV. According to a likelihood ratio test, model 8 fit significantly better than model 7 ($p = .01$). Examination of the coefficients revealed that, amongst the other group effect sizes, interference costs were marginally significantly larger than effect sizes for global RTs ($B = .17, CI = -.01 : .32, Z = 1.9, p = .05$). Amongst global RTs, the York lab differed from the other lab ($B = .93, CI = .60 : 1.27, Z = 5.44$, Z = .0001).
and that the difference between global RTs and interference costs in the York lab differed significantly from the difference between global RTs and interference costs in the other labs ($B = -.65$, $CI = -.97 : -.43$, $Z = -4.43, p < .001$). Including the interaction accounted for substantial variability both across and within comparisons ($R^2_{\text{between}} = .32$; range $R^2_{\text{between}} = .29 : .36$; $R^2_{\text{within}} = .32$; range $R^2_{\text{within}} = .29 : .39$).

Table 7

**Estimated Effect Sizes and Confidence Intervals for Lab Group.**

<table>
<thead>
<tr>
<th>Age Group</th>
<th>N</th>
<th>Overall</th>
<th>Interference Cost</th>
<th>Global RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>30</td>
<td>.14</td>
<td>.12</td>
<td>.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.05 : .22)</td>
<td>(-.09 : .10)</td>
</tr>
<tr>
<td>FIU</td>
<td>4</td>
<td>-.11</td>
<td>-.13</td>
<td>-.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-.28 : .07)</td>
<td>(-.40 : .22)</td>
</tr>
<tr>
<td>NW</td>
<td>5</td>
<td>-.04</td>
<td>.00</td>
<td>-.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-.20 : .11)</td>
<td>(-.28 : .16)</td>
</tr>
<tr>
<td>OC</td>
<td>5</td>
<td>-.07</td>
<td>.03</td>
<td>-.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-.22 : .08)</td>
<td>(-.39 : .03)</td>
</tr>
<tr>
<td>SFSU</td>
<td>6</td>
<td>-.16</td>
<td>-.12</td>
<td>-.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-.30 : .02)</td>
<td>(-.38 : -.02)</td>
</tr>
<tr>
<td>UFRSG</td>
<td>4</td>
<td>.19</td>
<td>.15</td>
<td>.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(-.03 : .41)</td>
<td>(-.09 : .55)</td>
</tr>
<tr>
<td>York</td>
<td>30</td>
<td>.82</td>
<td>.49</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.48 : 1.15)</td>
<td>(.57 : 1.69)</td>
</tr>
</tbody>
</table>

Confidence Intervals in parentheses.
All estimates are from traditional random-effects meta-analysis

In order to determine the nature of the interaction, separate analyses were conducted on the interference costs and global RTs. For interference costs, none of the labs differed significantly from the baseline lab. However, the SFSU lab differed marginally significantly from the baseline labs ($B = -.37$, $CI = -.77 : .02$, $Z = -1.84, p = .06$). For global RTs, the only significant slope indicated that effect sizes were significantly larger for the York lab than for the baseline labs ($B = 1.03$, $CI : .61 : 1.44$, $Z = 4.8, p < .001$).
Consistent with these results, Table 7 reveals that the York lab reported very large effect sizes for global RTs ($d = 1.35$, $CI = .65 : 2.04$) and medium-to-large effect sizes for interference costs ($d = .58$, $CI = .13 : 1.04$). Only one other average effect size had a confidence interval that did not contain zero. The baseline labs had a medium effect size with confidence intervals that did not contain 0 ($d = .24$, $CI = .13 : .36$).

**Sensitivity Analysis.** As described in the Method section, correlations between congruent and incongruent trials were imputed so that standard deviations of interference costs and global RTs could be calculated. Given the large number of effect sizes that contained imputed correlation coefficients, it was important to determine how sensitive the conclusions were to the specific value of the correlation. To do this, a sensitivity analysis was conducted. Three additional datasets were generated (from here on, Set B, Set C, and Set D). Each assumed a different average correlation between congruent and incongruent trials. Set B assumed that the mean correlation coefficient for congruent and incongruent trials was .5 with the same standard deviation as main data set. This was much lower than the correlation of .83 that was estimated from the meta-analysis described in the method section, but still a plausible seeming value. Set C assumed the correlation between congruent and incongruent trials was 0 with the same standard deviation was the original dataset. A correlation of 0 is an implausible situation, but showing that results held up under these conditions would suggest the choice of correlation coefficient had very little effect on the output. Set D assumed a correlation of .90 with a reduced standard deviation so that no correlations greater than 1 are simulated. Each of the 9 models described in the previous section was then re-fit to the new datasets and inferences were compared to those made in the original data set.
Each of these three sets produced slightly different parameter estimates than those produced earlier, but every coefficient that was significant in the original analysis was significant across all three datasets and no coefficients that were non-significant in the original analysis were significant across any of the three datasets. These analyses provide strong evidence that the results above are not very sensitive to assumed correlations between congruent and incongruent trials. It is extremely unlikely that the true unobserved correlations are greater than .90 or less than .0 and changing the average imputed correlation to either of these values had no impact on statistical inference.

Summary of results

Prediction 1 was that there would be a significant, small-to-moderate overall effect size. This prediction was supported ($d = .29$, $CI = .15 : .44$). However, given the clear evidence of publication bias in the funnel plots, and the aims of this study, this should not taken as evidence for a bilingual advantage until the effect of theoretically significant moderator variables is established. Additionally there was strong evidence of heterogeneity, justifying the examination of moderator variables. Contrary to prediction 2, there was no main effect of DV. This suggests that either bilingual advantage is not restricted to one of the two DVs, or that the advantage is restricted to one of the two DVs for certain groups or tasks.

Contrary to prediction 3, there was no main effect of task, and no interaction between task and DV. This effect was surprising, given the low convergent validity among tasks. If the Simon, Flanker and Stroop tasks are unrelated, and there is a bilingual advantage on one task, it would seem likely that a task the advantage would be restricted to a single task. However, if these tasks do reflect the same construct, this finding would be consistent with the existence of a bilingual advantage.
Prediction 4 was that there would be a main effect of age, with older adults and children exhibiting larger advantages than young adults. Evidence on this prediction was mixed. Two sets of models were included in this analysis. In the first set, adults between the ages of 40 and 60, and those above 60 were grouped into separate categories. This was done because while research suggests that flanker costs begin increasing around age 40 (Waszak et al., 2010), it was desirable to make the older adult group as distinct as possible from the younger adults. Unexpectedly, the other adult group differed significantly from the young adult group, but neither of the children or older adults did. As this effect may have been driven by a few large effect sizes in a sample of 7, the older and other adult groups were collapsed into a single older adults category. Since this grouping is, arguably, theoretically more justified and reveals a more parsimonious interpretation, focus will be on the recoded models. This group exhibited significantly larger effect sizes than did the young adult group, but children did not differ from young adults.

While including the interaction between age and DV did not significantly improve the fit of the model, examination of the coefficients revealed that children exhibited significantly larger effect sizes on global RTs than did young adults. The coefficient for children*DV was also significant. Older adults exhibited significantly larger effect sizes than young adults on global RTs and the older adult coefficient did not interact with DV. These results might suggest that the bilingual advantage is more pronounced amongst children and older adults, and that the constructs reflected by the two DVs changes across the lifespan (more on this in Chapter 5). However, examination of funnel plots revealed stronger publication bias among the child comparisons than among the young adult comparisons, which could also account for the effect. Moreover, the interaction between children and DV was not strong enough to significantly improve the fit of the overall model. A sensitivity analysis also revealed that the coefficient for
the difference between the older adult and younger adult group was greatly influenced by a few effect sizes. The minimum value of the sensitivity range was $1/3$ of the estimated value, meaning that dropping one effect size reduced this effect size by 30%. Taken with the evidence of publication bias, this suggests that this overall difference might be driven by a small number of outlier effects.

Evidence for prediction 5 is also mixed. There was no significant main effect for age of acquisition, but there was a significant interaction between AoA and DV. Subsequent analyses revealed that on interference costs there was no significant difference between early and late AoA comparisons, but that on global RTs, effect sizes for early AoA comparisons were significantly larger than those for late AoA comparisons. This interaction suggests that a bilingual advantage might be restricted to early AoA bilinguals on global RTs. However, as was the case with the significant interaction between age and DV, differential publication bias between early and later AoA comparisons could also be responsible for this effect.

Prediction 6 was supported. There was a significant effect of lab, which accounted for substantial variation between and within comparisons. This effect was primarily driven by the York lab, which had significantly larger effect sizes than the baseline lab. No other lab differed significantly from the baseline labs. There was a significant interaction between lab and DV as well. This interaction was primarily driven by a difference between the effect of DV in the York lab and the baseline labs. Analyses of interference costs and global RTs revealed that the York lab differed significantly from the baseline lab only on global RTs but not interference costs. These results do not conclusively support prediction 6a or 6b. On the one hand, it seems like the lab effect was much more obvious on global RTs than interference costs. On the other hand, this lab effect was primarily driven by a single lab.
Chapter 4: Study 2

Study 2 synthesized effect sizes comparing monolinguals and bilinguals on task-switching tasks. Unfortunately, the participant- and task-characteristics do not vary much across these studies, so the systematic examination of many moderators was not possible. The following predictions were tested:

1) Prediction 1: Consistent with de Bruin et al. (2015) there will be a small-to-medium effect size.

2) Prediction 2: Effect size will be moderated by DV. Since it is theoretically possible that a bilingual advantage is restricted to either of these DVs, no directional prediction is made.

3) Prediction 3: Lab will significantly moderate effect sizes.
   a. If lab effects are due to methodological factors, there will either be no interaction between lab and DV or the interaction will not follow a coherent pattern.
   b. If the lab effects are due to difference in bilingual populations, there will be an interaction between lab and DV with a coherent interaction.

Method

Literature search. PsycINFO was searched periodically until July 2015. Search terms included a combination of bilingual or bilingualism with executive control, executive function, switching, shifting and cognitive flexibility. Reference sections of the relevant studies and review articles (e.g., de Bruin et al., 2015, Paap et al., 2015) were examined to identify additional studies for inclusion. Ten studies were included met the inclusion criteria for the analysis:
1. Study included at least one bilingual group. Because different studies use different measures as indicators of bilingualism (e.g., AoA, frequency of use, overall proficiency), it is impossible to identify a single definition of bilingualism. Therefore, a group of participants was designated as bilingual if any of the following were true: the age at which they began learning their L2 was equal to or less than half their age at the time of testing; participants reported near equivalent proficiency in both languages; participants reported native or near-native attainment in their L2; participants reported using each of their two languages in at least 40% of their daily activities; participants reported using both languages at home; participants reported using one language at home and one language at school (with the exception of children who were recently enrolled in immersion programs).

2. Study included at least one monolingual group, defined as participants with only minimal exposure to an L2, e.g., through foreign language classes at school.

3. Participants were at least five years old and without psychological impairment. We therefore excluded studies examining potential benefits of bilingualism on residual cognitive functioning in dementia.

4. It contained RT data from at least one computerized, non-linguistic task-switching task. To be included the task had to meet the following criteria: (A) Each trial contains a stimulus that varied on two dimensions, and participants must attend to one of the two on each trial; (B) The dimension to be attended to varies across trials, thereby permitting the calculation of switch costs. (C) The response is a non-verbal response, meaning that language-switching tasks were excluded; (D)
The task contained no additional executive control manipulations. Two studies (Garbin et al., 2010 and Gold et al., 2013) only contained switch blocks. These studies were included, in order to maximize the available data, but because they lacked single-task blocks it was not possible to calculate mixing costs for these trials.

**Data preparation and effect size calculation.** The 10 studies were further broken down into 15 comparisons. Consistent with Study 1, a comparison was defined contrasting a bilingual group and a monolingual group on a single task. If a study reported on several separate bilingual and monolingual groups, (e.g., varying in age), each of these groups was included in a separate comparison. Care was taken to minimize the number of statistically dependent comparisons while maximizing the number of available effect sizes. First, when a study contained a single monolingual group and several bilingual groups, RTs from the bilingual groups were averaged together. This was done because none of the bilingual groups differed on any of the moderator variables. Second, in one study (Prior & Gollan, 2013), the same task was administered on three separate occasions to test for practice effects. Only the first administration of the task was included in the data set since multiple instances of the same task are certainly correlated. Unlike the studies included in Study 1, none of the ten task-switching studies administered more than one task to the same set of participants.

Recall that that the measure of mixing costs is the difference in RTs between the repeat trials in a switch block and those of a non-switch block and the switch cost is the difference between switch and repeat trials in the switch blocks. As was the case for interference costs and global RTs in Study 1, in some studies, these scores and their standard deviations were reported. However, in many cases, means and standard deviations (or standard errors) were reported for
the non-switch blocks, repeat trials and switch trials, but not the costs. Since the standard deviations of these costs require the correlation between these trial types, these were estimated using a similar procedure to the one used in Study 1.

Correlation coefficients were simulated for both global and local switch costs. To do so, authors who were emailed to provide other information for the analysis were asked to provide correlations between non-switch block trials and repeat trials in the switch block (to calculate mixing cost), and non-switch and switch trials from the switch block (to calculate the switch cost). Two pairs of correlation coefficients were received. These two pairs of correlation coefficients were combined with three pairs of correlation coefficients from an unpublished dataset. A random effects meta-analysis was conducted on these estimates in order to estimate a mean and standard deviation for the correlation coefficients. For the mixing costs, the meta-analysis yielded an average correlation of .64, with significant heterogeneity, $Q(4) = 62.23, p < .001$ and a standard deviation of .28. For the switch costs, the meta-analysis yielded an average correlation of .88, with significant heterogeneity $Q(4) = 43.38, p = < .001$, and a standard deviation of .11. Seven correlations were then simulated from a normal distribution with a mean of .64 and a standard deviation of .28, with a restriction that the coefficient could not be greater than 1. These coefficients were used to calculate standard deviations for the mixing costs. Nine correlations were simulated from a normal distribution with a correlation of .88 and standard deviation of .11 with a restriction that the coefficient could not be greater than 1. These coefficients were used to calculate standard deviations for switch costs. Because of the heterogeneity of these correlation coefficients, sensitivity analyses were conducted to assess how sensitive results were to the assumed correlations. Cohen’s $d$ was
### Table 8

**Effect sizes and moderators for each of the 84 comparisons for Interference Control Tasks.**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Sample Size</th>
<th>Modifiers</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N Bil</td>
<td>N Bil</td>
<td>Task</td>
</tr>
<tr>
<td>Barac and Bialystok 2012</td>
<td>78</td>
<td>26</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Garbin et al 2010</td>
<td>19</td>
<td>21</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Gold et al 2013 Study 2 Younger</td>
<td>20</td>
<td>20</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Gold et al 2013 Study 2 Older</td>
<td>20</td>
<td>20</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Hernandez et al 2013 Study 3</td>
<td>38</td>
<td>39</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Moradzadeh et al 2015 Musician</td>
<td>36</td>
<td>45</td>
<td>Number-Quantity</td>
</tr>
<tr>
<td>Moradzadeh et al 2015 Non-Musicians</td>
<td>36</td>
<td>36</td>
<td>Number-Quantity</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 1</td>
<td>30</td>
<td>44</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 2</td>
<td>31</td>
<td>49</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Paap and Greenberg 2013 Study 3</td>
<td>48</td>
<td>51</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Paap and Sawi 2014</td>
<td>58</td>
<td>62</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Prior and Gollan 2011</td>
<td>84</td>
<td>47</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Prior and Gollan 2013</td>
<td>27</td>
<td>52</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Prior and MacWhinney 2010</td>
<td>44</td>
<td>44</td>
<td>Color-Shape</td>
</tr>
<tr>
<td>Rodriguez Pujadas et al 2013</td>
<td>18</td>
<td>18</td>
<td>Color-Shape</td>
</tr>
</tbody>
</table>
calculated for switch costs and mixing costs in each of the 15 comparisons. Table 8 contains Cohen’s $d$ and its variability for each comparison.

**Moderator Coding.**

*Lab.* The only moderator considered in this study was lab. Similar to the procedure in Study 1, except that papers with a most common author shared by three studies were grouped into a lab. This lead to four lab groups: SFSU (San Francisco State University), UoH (University of Haifa), UPF (Universitat Pompeu Fabra) and Other labs.

**Analytic strategy.** Because switch and mix costs are correlated, the same analytic strategy employed in Study 1 was employed here. Effect sizes were analyzed in three-level meta-analyses, with comparisons serving as clusters. A null model was fit and served as a baseline. Model 1 contained DV as a moderator. Model 2 contained DV and lab, and model 3 contained their interaction (All models are described in Table 9). Inferences were based on likelihood ratio tests and examination of slopes. For every model, a sensitivity analysis was conducted to determine whether results were unduly influenced by a single comparison. After all the models were fit, a second sensitivity analysis was conducted to determine whether results were sensitive to the assumed correlations between trial types.

**Results**

Prior to fitting three-level meta-analyses, random effects meta-analyses were run on mixing costs and switch costs separately in order to produce forest plots. The effect size for mixing cost was small and non-significant ($d = -.06$, $CI = -.32 : .20$, $p = .65$) and exhibited significant heterogeneity ($Q(12) = 45.79$, $p < .001$). This heterogeneity can be seen in the forest plot in Figure 7. The average effect size for switch cost was also small and non-significant ($d = -.02$, $CI$
Figure 8: Forest Plot of Switch Costs

Table 9

Overall Fit Statistics for Moderator Models Switching Tasks

<table>
<thead>
<tr>
<th>Model</th>
<th>Moderators</th>
<th>$R^2_{\text{within}}$</th>
<th>$R^2_{\text{between}}$</th>
<th>$P_{\text{null}}$</th>
<th>$P_{\text{model}}$</th>
<th>$P_{\text{addonly}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>DV</td>
<td>.02</td>
<td>.02</td>
<td>.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>DV + Lab</td>
<td>.09</td>
<td>.14</td>
<td>.73</td>
<td>.60</td>
<td></td>
</tr>
<tr>
<td>Model 3</td>
<td>DV * Lab</td>
<td>.66</td>
<td>.03</td>
<td>.61</td>
<td>.51</td>
<td>.34</td>
</tr>
</tbody>
</table>

Notes.
Models with interactions are compared to the null model, model 1, and the relevant additive model.
= -.16 : .13, p = .80) and exhibited marginally significant heterogeneity ($Q(14) = 22.56, p = .07$), which can be seen in Figure 8.

The null model yielded a small, negative, non-significant overall effect size, ($d = -.03, CI = -.19 : .14, Z = -.30, p = .77, range d = -.07 : .02$). As can be seen in Table 9, neither the within study variance components ($\tau^2_{\text{within}} = .03, p = .04, \text{range } \tau^2_{\text{within}} = .00 : .04$) or the between study variance ($\tau^2_{\text{between}} = .06, p = .23, \text{range } \tau^2_{\text{between}} = .01 : .07$) differed significantly from 0; however, the Q statistic revealed significant heterogeneity, $Q(27) = 68.64, p < .001$. Model 1 included DV as a factor with mixing cost as the baseline level. According to a likelihood ratio test, model 1 did not fit significantly better than the null model ($p = .75$). Consistent with this, the slope for DV was small and non-significant ($B = .04, CI = -.18 : .26, p = .75, \text{range } B = -.01 : .12$). DV accounted for little variability within comparisons ($R^2_{\text{within}} = .02, R^2_{\text{within}} \text{range } = .00, .02$).

Figure 9: Forest Plots of Mixing Costs
Table 10

**Regression Coefficients from Moderator Models Switch Tasks**

<table>
<thead>
<tr>
<th></th>
<th>Null</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-.03</td>
<td>-.05</td>
<td>-.17</td>
<td>-.07</td>
</tr>
<tr>
<td>DV</td>
<td></td>
<td>.05</td>
<td>.04</td>
<td>-.16</td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td>.31</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>Paap</td>
<td></td>
<td>.05</td>
<td>.02</td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td></td>
<td>.21</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>DV*Costa</td>
<td></td>
<td></td>
<td>.41</td>
<td></td>
</tr>
<tr>
<td>DV*Paap</td>
<td></td>
<td></td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>DV*Prior</td>
<td></td>
<td></td>
<td></td>
<td>.41&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Tau2 2</td>
<td>.04</td>
<td>.04</td>
<td>.03</td>
<td>.01**</td>
</tr>
<tr>
<td>Tau2 3</td>
<td>.06</td>
<td>.06</td>
<td>.05</td>
<td>.06**</td>
</tr>
</tbody>
</table>

*Note.*

<sup>a</sup> = .05 < <sup>l</sup>p < .10

* = .01 < <sup>l</sup>p < .05

** = <sup>l</sup>p < .01

**Testing the effect of lab.** Model 2 included lab as a factor, with other labs serving as the baseline level (for clarity, this group will be referred to as baseline). According to likelihood ratio tests, model 2 did not improve fit relative to either the null model (\(p = .74\)) or model 1 (\(p = .60\)). Consistent with this, Table 9 shows that none of the lab coefficients are significantly different from 0. As can be seen in Table 10, all effect sizes were non-significant.

Model 4 included an interaction between DV and lab. According to likelihood ratio tests, model 4 did not improve fit relative to the null model (\(p = .70\)), model 1 (\(p = .61\)) or the model with additive effects only (\(p = .54\)). Consistent with this, Table 10 shows that none of the coefficients were significantly different from 0. As can be seen in Table 11, effect sizes were medium to small and all confidence intervals contained 0.

**Sensitivity Analysis.** As described in the Method section, correlations between the trials in non-switch blocks and non-switch trials of switch blocks as well as correlations between non-switch and switch trials of switch blocks had to be imputed. Given the large number of effect
Table 11

Estimated Effect Sizes and Confidence Intervals for Each AoA Group.

<table>
<thead>
<tr>
<th>Lab</th>
<th>N</th>
<th>Overall</th>
<th>Switch Cost</th>
<th>Mix Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>5</td>
<td>- .14</td>
<td>-.25</td>
<td>-.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-.50 : .25)</td>
<td>(-.48 : -.02)</td>
<td>(-.82 : .73)</td>
</tr>
<tr>
<td>SFSU</td>
<td>4</td>
<td>- .10</td>
<td>-.08</td>
<td>-.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-.25 : .04)</td>
<td>(-.29 : .13)</td>
<td>(-.33 : .08)</td>
</tr>
<tr>
<td>UoH</td>
<td>3</td>
<td>.11</td>
<td>.21</td>
<td>.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-.06 : .28)</td>
<td>(-.03 : .44)</td>
<td>(-.22 : .26)</td>
</tr>
<tr>
<td>UPF</td>
<td>3</td>
<td>.14</td>
<td>.28</td>
<td>-.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-.36 : .65)</td>
<td>(-.41 : .97)</td>
<td>(-.63 : .28)</td>
</tr>
</tbody>
</table>

Confidence Intervals in parentheses.
All estimates are from traditional random-effects meta-analysis

sizes that contained imputed correlation coefficients, it was important to determine how sensitive
the conclusions were to the specific value of the correlation. To do this, a sensitivity analysis was
conducted. Three datasets (from here on, Set B, Set C, and Set D. The originally dataset will be
referred to as Set A) were created assuming different mean correlations between congruent and
incongruent trials. Set B assumed that the mean correlations for one half the size of those in Set
A. These correlations are plausible but much smaller than those imputed. Set C assumed the
mean correlations were both 0 and the same standard deviation was the original dataset. A
correlation of 0 is an implausible situation, but showing that results held up under these
conditions would suggest the choice of correlation coefficient had very little effect on the output.
Set D assumed mean correlations of .90, with a reduced standard deviation so that no
correlations greater than 1 are simulated. Each of the 4 models described in the previous section
was then re-fit to the new datasets and inferences were compared to those made in the original
data set.
Each of these three sets produced slightly different parameter estimates than those produced earlier, but, as was the case in the main analysis, no effects were statistically significant across any of the analyses. These analyses provide strong evidence that these results are not very sensitive to assumed correlations between the three trial types. It is extremely unlikely that the true unobserved correlations are greater than .90 or less than .0 and changing the average imputed correlation to either of these values had no impact on statistical inference.

Summary of Results
Prediction 1, that there would be a significant overall effect size, was not supported. The overall effect size was slightly negative and not significantly different than 0. Prediction 2, that effect sizes would vary according to DV, was also not supported. There was no significant difference between switch costs and mix costs. Prediction 3 was also not supported. Lab did not significantly moderate effect sizes and did not interact with DV. There appears to be no evidence of a bilingual advantage on task-switching tasks.
Chapter 5: Discussion

The purpose of this dissertation was to examine the effects of a set of theoretically significant moderator variables on the bilingual advantage within the published literature. To this end, two studies were conducted: Study 1 considered interference-control tasks and Study 2 considered domain-general task-switching tasks. Study 1 revealed a complicated pattern of results, in which the overall effect size and some moderators were significant but these effects were obscured by the presence of differential publication bias across levels of the moderator variables. The results of Study 2 were relatively straightforward. The overall effect size was non-significant and there was no effect of the moderator variables.

Discussion and Theoretical Implications of Study 1

Study 1 revealed a significant overall effect size, similar to that reported in de Bruin et al. (2015). However, given the evidence of publication bias reported herein and in de Bruin et al. (2015), and the arguments in Chapter 1, understanding the overall effect size was not the purpose of this study. There was no significant effect of DV; however, given that DV interacted with some other moderator variables, it will be discussed in the context of these results. There was a non-significant main effect of Task, and it did not interact with DV. There was a significant effect of age, driven by larger effect sizes for the other adult group than the young adult group. This effect interacted with DV, such that older adults had larger effect sizes than young adults on interference costs, whereas children had marginally significantly larger effect sizes than young adults on global RTs, and other adults had larger effect sizes than younger adults on both interference costs and global RTs. There was no significant effect of AoA, but there was a significant interaction between AoA and DV, driven by larger differences for early than late AoA participants on global RTs but no difference on interference costs. There was a main effect
of lab, and an interaction between lab and DV. However, this lab effect was driven by one lab, and the interaction was driven by smaller effect sizes on interference costs than global RTs within that lab.

The non-significant effect of task was surprising. It is widely acknowledged that correlations between interference costs on variations of Simon, Flanker and Stroop tasks are quite low (see Paap & Sawi, 2014, for a discussion). If interference costs on the four sets of tasks included in this analysis represent distinct constructs, and only one is systematically related to bilingualism, there should be a consistent advantage on only one of these tasks. There are two plausible explanations for why the effect of task was not significant. First, if the bilingual advantage is purely an artifact of publication bias, confounding variables and questionable research practices (as suggested by Paap et al., 2015), one would not expect to see an effect of task as there is no reason why these factors would affect one task and not others. So this finding could be seen as evidence against a bilingual advantage at all. A second possibility is that these costs do in fact represent similar constructs and this is not captured by the correlations reported by other authors. Consistent with this possibility, Willoughby et al. (2015) have argued that EF tasks given during the same session should not be correlated. They note that neural activity in the prefrontal cortex is related to task difficulty in an inverted-U shaped manner: as task demands increase, prefrontal cortex activity increases until a point of optimal difficulty and decreases after that point. Importantly, this optimal difficulty level is subject to individual differences. Moreover, as the amount of time spent at this difficulty level increases, neurocognitive resources available for the next task decrease. As a result, if task 1 is optimally difficult for participant A, they will have fewer EF resources available for task 2 than would participant B, for whom task 1 is too difficult. This dynamic may diminish any correlation between tasks and Willoughby and
colleagues offer it as an explanation for the low correlations between tasks within the same battery. This would be especially likely for tasks that purportedly measure the same construct, such as the Simon, Flanker and Stroop tasks. Either of these interpretations is plausible: The lack of a lab effect could suggest that there is no consistent bilingual advantage, or that the tasks may recruit more common resources than is suggested by the low correlations, and these resources are affected by bilingualism.

There was a significant effect of age, which was primarily driven by significant differences between the older and young adults groups. Examination of the interactions revealed an unexpected but interesting pattern of results. Recall the prediction that if age interacted with DV, there would be larger effect sizes for children and older adults than younger adults on either interference costs or global RTs. The older adult group exhibited significantly larger effect sizes than the young adult group on both interference costs and global RTs; moreover, the difference between effect sizes on interference costs and global RTs was non-significant. However, the children differed significantly from the young adults on only the global RTs. This effect is consistent with the prediction that the bilingual advantage is most easily detectable among older adults and children if we assume that the interference cost and global RTs do not represent the same constructs across the lifespan. Structural equation modeling studies suggest this is possible. While studies conducted on college students find that variance on batteries of EF tasks can be decomposed into distinct components (Miyake et al., 2012; Friedman, & Miyake, 2012), studies with children typically extract only a single factor. For example, Willoughby, Pek and Blair (2013) administered seven EF tasks to three-to-five year old children at three time points and found that the tasks could be described by a single factor. Similar results have been reported in many studies of children (see Willoughby, Holochwost, Blanton & Blair, 2014, for a review).
These results suggest that the nature of EF, and therefore, the interpretation of tasks which purportedly measure it, may change dramatically across the lifespan. This might explain why a larger advantage would emerge on global RTs for children and both DVs for older adults.

An alternative explanation for the interaction between age and DV is that it reflects methodological artifacts. First, remember that the difference between children and young adults on global RTs was marginally significant ($p = .05$). Moreover, as Figure 5 illustrates, there is stronger evidence of publication bias amongst the global RTs for children, than among the global RTs for younger adults. Differential publication bias between samples with children and young adults could be responsible for this significant difference. Second, as can be seen in Figure 5, there was clearly publication bias amongst the older adult group when the re-coded age variable was used. Furthermore, the sensitivity analysis revealed that dropping a single effect size from the model reduced the slope by 30%. It is, therefore, possible that the significant difference between younger and older adults is due to a few very large effect sizes.

There was no significant main effect for AoA. However, there was a significant interaction between AoA and DV. This interaction was due to the fact that early AoA comparisons had significantly larger effect sizes than later AoA comparisons on global RTs but not interference costs. This finding is consistent with modeling and priming studies suggesting that for early AoA bilinguals, representations of the two languages are strong enough to interact bi-directionally (Li, 2012; Sabourin, Brien & Burkholder, 2014). Furthermore, if the conventional interpretations of global RTs and interference costs (described in the introduction section) are followed, these results would suggest that conflict between competing representations in the L1 and L2 is strengthened, not necessarily through inhibition of the non-target language, but through goal maintenance (e.g., Colzato et al., 2008). This conclusion would
have the benefit of being theoretically coherent and suggesting future directions for work in bilingual lexical processing. However, these effects can also be explained in terms of differential publication bias. According to Figure 6, there is strong evidence of publication bias for global RTs in bilinguals with early AoA, but not late AoA.

There was a strong, significant lab effect, primarily driven by a difference between the York lab and the baseline lab category. The interaction between lab and DV revealed that this effect was more pronounced among global RTs than among interference costs. This effect seems more consistent with the possibility that methodological factors specific to labs might drive the effect rather than differences in bilingual subject populations. If it were due to characteristics of bilingual populations, it is strange that it would be restricted to only a single lab. There are possible methodological factors that could explain this, including differential demand characteristics (Powers et al., 2013), differences in confounding variables and differential treatment of outliers amongst the labs.

Differential treatment of outliers across labs is a particularly interesting option to consider. It is well known that experimental manipulations to cognitive tasks affect both means and tails or RT distributions (Luce, 1986). If the bilingual advantage is restricted to the tail of RT distributions, it is possible that the lab effect might reflect this. Two pieces of evidence support this claim. First, Calabria, Hernández, Martin and Costa (2011) re-analyzed data from Costa et al. (2008) and Costa et al. (2009), using the exponentially modified Gaussian distribution (a convolution of the exponential and Gaussian distributions which allows for separate estimation of RT means and tails). They observed a bilingual advantage for global RTs in both the mean and tail and a bilingual advantage for interference costs in the tail only. If the bilingual advantage is most evident in the tail of RT distributions, it may be eliminated or reduced by removing large
numbers of outliers. Consistent with this possibility, Zhou and Krott (2015) analyzed 32 published papers and categorized RT trimming methods as short allowance, medium allowance or long allowance, and found that the longer the allowed RTs, the more likely the study was report a bilingual advantage. This effect held up even when one lab, which accounted for a large number of effect sizes was removed from the dataset, suggesting it was not due to the large influence of that lab. The present analysis did not consider RT trimming procedure because it could not be reliably coded. However, given that the York lab allowed longer RT responses than many of the others, this is certainly a plausible explanation for the lab effect.

In sum, the evidence for the bilingual advantage is equivocal. There were no differences between tasks, despite the fact that they exhibit relatively low convergent validity. While the predictions for age and AoA were partially supported, these effects could also be explained in terms of differential publication bias. Finally, there was strong evidence of a lab effect, which was primarily driven by a single lab. However, this effect is consistent with research showing that different RT trimming methods influence the presence of bilingual advantages. Given these facts, there is no strong reason to assume a bilingual advantage amongst the current set of studies. However, limitations of this meta-analysis and the existing studies could be responsible for the lack of a bilingual advantage. These limitations are discussed in the Limitations and Future Directions section.

**Discussion and Theoretical Implications of Study 2**

The overall effect size for domain-general task-switching tasks was not significantly different from 0. These effect sizes were not moderated by DV, suggesting no significant difference between switch and mixing costs. Moreover, the effect sizes were not moderated by lab and there was no interaction between lab and DV. These results suggest that if a bilingual advantage exists,
it likely is not due to the experience of switching between languages and is not detectable on
domain-general task-switching tasks under standard task conditions.

Limitations and Future Directions

This dissertation suggests there is ambiguous evidence for bilingual advantages on interference
control tasks and no evidence for bilingual advantages on switching tasks. However, this does
not mean there is no bilingual advantage on executive control tasks. Limitations of the present
analysis and the literature in general may obscure such effects. For example, the present analysis
did not consider potentially important participant-level moderators or more complicated
executive control tasks. Moreover, the literature, more generally, has arguably focused on the
wrong constructs and applied statistical methods that are not sufficiently sensitive for this
context.

While the purpose of this meta-analysis was to consider whether bilingual advantages are
moderated by theoretically significant variables, several potentially important variables were not
included. The most obvious example is proficiency. Models of bilingual lexical representation
such as the Revised Hierarchical Model (Kroll, Van Hell, Tokowicz & Green, 2010) suggest that
L2 proficiency impacts the organization of the bilingual lexicon. If this is the case, it is possible
that bilingual advantages, especially on interference control tasks, are limited to high proficiency
bilinguals. The effect of proficiency was not examined in this study because it was not assessed
or reported in several publications. Furthermore, the studies that reported on proficiency
measured it in very different ways, and, as a result, it is not clear how to characterize some
comparisons as high proficiency and some bilinguals as low proficiency.

An ideal analysis would have included proficiency, as either a continuous or categorical
variable. However, evidence for AoA effects was weak to moderate. While AoA is a different
variable than proficiency, and effects of the two variables are often compared in computational modeling studies and some empirical studies (e.g., Sarbouin et al., 2014), in most studies they are highly correlated and their results are often difficult to distinguish. Outside the context of studies with sampling method aimed at pitting AoA against proficiency, it seems unlikely that the two could be reliably distinguished. So, while this effect is plausible, it seems unlikely.

A second important moderator that was not included in the present study is the frequency with which a bilingual switches between languages. If bilingual advantages emerge because of frequent language switching, rather than as a consequence of lexical representation, then it is likely that bilinguals who switch between languages more frequently would exhibit smaller switch costs than those who switch less frequently, or monolinguals. Indeed, Prior and Gollan (2011) found that Spanish-English bilinguals, who regularly switch between languages, exhibited smaller switch costs than either Chinese-English bilinguals who reported switching between languages less often as well as monolinguals. It is possible that if frequency of L2 switching had been included as a moderator, it would have significantly moderated effect sizes in Study 2.

Relatedly, it may be that a bilingual advantage exists only for bilinguals from socio-linguistic environments that exhibit dense code switching (Costa et al., 2009; Green & Wei, 2014). For example, Costa et al., (2009) speculated that in socio-linguistic contexts characterized by a high degree of code-switching, bilinguals must monitor for cues for when to switch languages; such cues may include the linguistic background of interlocutors or the broader linguistic context of the conversation. This might suggest that advantages on global RTs, or perhaps mixing costs, might only be evident for bilinguals from socio-linguistic contexts in which code-switching is the norm. More recently, Green and Wei (2014) proposed that the type of control procedures necessary for lexicalization might be adapted to the density of code-
switching within a socio-linguistic community. If this is true, one might expect that bilinguals of different socio-linguistic backgrounds might exhibit advantages on different tasks.

An additional limitation of the present analysis is its narrow focus on two sets of tasks. Interference control tasks and task-switching tasks were included in this analysis because they have conventional interpretations and because they have been widely used by researchers aiming to understand the bilingual advantage. However, some researchers have found bilingual advantages in complex tasks that involve the coordination of several executive function processes (Bialystok, 2011). These tasks were not included because they do not have conventional interpretations or a clear, theoretically motivated connection to bilingual language processing.

Moreover, some researchers have included modified versions of interference control or switching tasks with increased working memory demands. For example, Bialystok et al. (2004) included a version of the Simon task in which participants must remember four color-response mappings instead of two and found larger bilingual advantages on this task than on a version of the task with just two color-response mappings. However, they did find significant advantages on both. Moreover, Hernández et al. (2013) found that on a modified task-switching task that contained four rules, and therefore four different sets of stimulus-response mappings, bilinguals outperformed monolinguals on all trial types, even though in a second experiment, bilingual did not outperform monolinguals on traditional task-switching. If such effects are reliable, they could be explained in terms of Conflict Monitoring Theory (Botvinick, Cohen & Carter, 2004), which claims that the Anterior Cingulate Cortex detects conflict and, then, up-regulates cognitive control. Botvinick, Cohen and Carter (2004) argue that one source of conflict in cognitive tasks is errors; when errors are regularly committed, they compete with subsequent correct responses,
thereby increasing conflict. It may be that, consistent with the predictions of many authors (e.g., Costa et al., 2009; Hilchey & Klein, 2011), bilinguals more efficiently detect conflict, requiring a lower threshold for the up-regulation of cognitive control, but that the tasks summarized in this meta-analysis were not consistently challenging enough for this difference to emerge. If this is so, bilingual advantages may emerge on tasks for which the participant is likely to make errors. Indeed if one is willing to discount the evidence of publication bias amongst children on global RTs for this analysis, the larger effect sizes for global RTs amongst children is consistent with this account, as the probability of committing an error could be highest for this group.

While this account is plausible, it was not possible to include non-standard versions of these tasks in the present analysis. Scores on these tasks, inarguably, reflect additional EF processes beyond those required for conventional versions of these tasks. They, therefore, could not be averaged with effect sizes from the conventional tasks without including a moderator variable of some sort. However, there were far too few of these tasks to justify including a moderator variable.

It is also possible that bilingual advantages exist, but the existing literature has been using an inappropriate conceptualization of EF. Much of the theorizing about the bilingual advantage assumes that the demands of bilingual language processing are managed by recruitment of some sort of domain-general cognitive mechanism. Constant exercise of this mechanism strengthens it, leading to aggregate differences between populations of bilinguals and monolinguals.

Candidate mechanisms are drawn often from structural equation modeling studies of EF tasks, such as those by Miyake and colleagues (Miyake et al., 2000; Friedman et al., 2008). These modeling studies consistently find that batteries of EF tasks can be decomposed into unique factors, which are named inhibition, shifting and updating (these authors acknowledge the
likely existence of other components as well). However, more recent studies show that these
tasks can be more parsimoniously decomposed into a hierarchical structure, in which one
common EF factor (similar to goal maintenance) predicts variance in switching- and updating-
specific latent variables (see Miyake et al., 2012, for a review).

However, while these factors might usefully reflect behavioral traits amongst large
groups of people, they do not necessarily reflect discrete or distinct cognitive mechanisms.
Consider, as a thought experiment, a study in which participants completed three academic tasks
and three athletic tasks. Confirmatory factor analysis would very likely reveal that the tasks
reflect two latent constructs, academic and athletic ability, as tasks loading on factor A would be
more similar to each other than they would be to tasks loading on factor B. Such traits might be
usefully employed in areas like behavioral genetics and educational psychology. But one would
not view these factors as representing discrete or distinct modules or processes. Likewise, while
inhibition tasks load on a single factor and switching tasks load on a single factor, there is no
reason to assume that the mind has a single inhibition mechanism or a single switching
mechanism. These constructs, while perhaps usefully defined for larger-scale studies, may not be
sufficiently precise for understanding how the mind accommodates the information-processing
demands from two languages. A more precise analysis of the mechanisms involved in controlled
information processing will go a long way toward clarifying if, when and where bilingual
advantages emerge. Some studies on bilingual language control during language processing have
sought to do this, by comparing different sorts of inhibition processes (Colzato et al., 2008).

A related limitation of the bilingual advantage literature is the use of mean RTs and
difference scores as the primary measures of EF. It has long been acknowledged that task-
manipulations affect components of RT distributions beyond just the mean, particularly the tail
of the distribution (Luce, 1986). Indeed, as described above, some researchers have begun using the exponentially modified Gaussian distribution, which allows for estimates of the mean, variance and tail of an RT distribution (Calabria et al., 2011). This technique is a step forward in that it could be more sensitive to experimental effects than a simple mean would be. However, one weakness of this approach is that parameters of this distribution, like means in ANOVA, do not have distinct theoretical interpretations (Matzke & Wagenmakers, 2009).

A more precise analysis of RT data comes from cognitive process models, such as the Ratcliff drift diffusion model of two-choice forced decision tasks (Ratcliff, 1978). The Ratcliff diffusion model models RT and accuracy distributions for two-choice forced decision tasks using four parameters with distinct theoretical interpretations (for example, the rate of evidence accumulation, and an initial response bias). This model provides excellent fit to RT distributions, and has the advantage of interpretable parameters (Ratcliff & McKoon, 2008). However, these parameters cannot be identified using the normal or exponentially modified Gaussian (ex-Gaussian) distribution. Matzke and Wagenmakers (2009) simulated from the diffusion model and fit the ex-Gaussian and shifted Wald distributions to the datasets. They found that diffusion model parameters did not distinctly map to parameters of either distribution; in other words, parameters of the ex-Gaussian and shifted Wald distributions corresponding to multiple cognitive processes.

A generalization of the diffusion model has been developed for interference-control tasks, such as the Simon, Flanker and Stroop (White, Ratcliff & Sterns, 2012). This shrinking-spotlight diffusion model contains additional parameters corresponding to the shrinking of the attentional window within a trial and provides excellent fit to Simon, Flanker and Stroop data (Servant, Montagnini, & Burle, 2014). However, like the simpler diffusion model, its components combine
in unintuitive ways, meaning that means, mean differences and skewnesses from RT distributions will reflect a combination of cognitive processes. Moreover, to the best of my knowledge, no equivalent model has been developed for task-switching tasks.

**Conclusion**

While a great deal of research has sought to identify if and under what conditions bilinguals outperform monolinguals on EF tasks, results have proven unreliable. The meta-analyses reported on in this dissertation reveal weak to moderate evidence for several predictions about when bilingual advantages should emerge. Moving forward it is important to get more specific about the nature of the control processes employed in bilingual language processes and to develop explicit models of how those processes are reflected in RT distributions. Additionally, the field might benefit from a large multi-site study, with consistent methods. Such an approach would allow researchers to test for the effects of many of the bilingual characteristics described above.
References


Barac, R., & Bialystok, E. (2012). Bilingual effects on cognitive and linguistic development: Role of language, cultural background, and education. *Child Development, 83*(2), 413-422.


Humphrey, A. D., & Valian, V. V. (2012). Multilingualism and cognitive control: Simon and Flanker task performance in monolingual and multilingual young adults. In 53rd Annual meeting of the psychonomic society, Minneapolis, MN.


Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press


