

Working memory and second language comprehension and production: A meta-analysis

Jared A. Linck · Peter Osthus · Joel T. Koeth · Michael F. Bunting

Published online: 24 December 2013
© Psychonomic Society, Inc. 2013

Abstract Although working memory (WM) figures centrally in many theories of second language (L2) proficiency development and processing, some have argued that the importance of WM is overstated (e.g., Juffs, *Transactions of the Philological Society*, 102, 199–225, 2004). Despite many studies over the past two decades, the literature lacks a quantitative synthesis of the extant results. In this article, we report a meta-analysis of data from 79 samples involving 3,707 participants providing 748 effect sizes. The results indicate that WM is positively associated with both L2 processing and proficiency outcomes, with an estimated population effect size (ρ) of .255. In additional analyses, we assessed whether the WM–criterion relationship was modulated by potential covariates identified in the literature search (i.e., participant characteristics, WM measure features, criterion measure factors, and publication status). The results of the covariate analyses indicated larger effect sizes for the executive control (vs. storage) component of WM, and for verbal (vs. nonverbal) measures of WM. Minimal publication bias was detected, suggesting that WM has a robust, positive relationship with L2 outcomes. We discuss the implications of these results for models of WM and theories of L2 processing and L2 proficiency development.

Keywords Working memory · L2 processing · L2 comprehension · L2 production · Second language acquisition · Meta-analysis

Electronic supplementary material The online version of this article (doi:10.3758/s13423-013-0565-2) contains supplementary material, which is available to authorized users.

J. A. Linck (✉) · P. Osthus · M. F. Bunting
Center for Advanced Study of Language, University of Maryland,
7005 52nd Avenue, College Park, MD 20742, USA
e-mail: jlinck@umd.edu

P. Osthus · J. T. Koeth
Second Language Acquisition Program, University of Maryland,
College Park, MD, USA

Introduction

Robust working memory (WM) effects have been found across a range of complex cognitive processes, including reasoning, problem solving, planning, abstraction, mental arithmetic, and first language (L1) comprehension (Engle, 2002; for a meta-analysis addressing the role of WM in L1 comprehension, see Daneman & Merikle, 1996). Although theoretical positions vary, there is little argument that these complex cognitive behaviors at least partially rely on the attentional and executive control processes that underlie WM performance. Second language (L2) processing places demands on these WM resources, too, especially for less-proficient speakers, and growing evidence indicates that executive functions support the various cognitive control mechanisms necessary for L2 use (e.g., Abutalebi & Green, 2008; Hernandez & Meschyan, 2006). It should thus come as no surprise that WM has been implicated in studies of L2 processing (see, e.g., Michael & Gollan, 2005; Tokowicz, Michael, & Kroll, 2004) and learning (e.g., Linck & Weiss, 2011; Martin & Ellis, 2012). Although it is uncontroversial that WM is related to both L2 proficiency development and use, the magnitude of the WM effects and the specific component that drives these effects (i.e., the executive control vs. short-term store component of WM) have been somewhat inconsistent across studies (for recent reviews, see Juffs & Harrington, 2011, and Williams, 2011). Moreover, it is unclear whether these inconsistencies simply reflect variation around the population effect size due to noise (e.g., sampling error or measurement error), or whether they are instead due to systematic differences that can inform theoretical models. For example, to foreshadow our results, complex span measures are stronger predictors of L2 outcomes than are simple span measures, suggesting that the executive control component of WM may play a larger role than short-term memory when using an L2. A sizeable number of relevant studies that have varied in their research design factors (e.g., sample size, language of WM assessment)

and participant characteristics (e.g., L2 proficiency level) are now available in the literature. Thus, a systematic, quantitative review seems warranted. To this end, we report a meta-analysis of the extant studies to better elucidate whether and under what conditions WM is related to performance on measures of L2 processing and proficiency.

Working memory

According to contemporary views, WM refers to the cognitive system(s) responsible for the control, regulation, and active maintenance of information in the face of distracting information (e.g., Conway, Jarrold, Kane, Miyake, & Towse, 2007). Baddeley's seminal multicomponent model divided the construct of WM into two separable systems, a storage-based system (i.e., slave systems), analogous to short-term memory (STM), and an executive, attentional system that controls information between the slave systems and long-term memory stores (Baddeley, 1986; Baddeley & Hitch, 1974).

Many modern theories of human cognition describe a single system that is dedicated to the temporary processing, maintenance, and holding of information that is relevant to current tasks—that is, the WM system. Many theoretical models exist to describe its operation (see the variety of opinions offered in Miyake & Shah, 1999), but its function remains similar across models: It orders, stores, and manages immediate sensory details until they can be properly incorporated into the cognitive process that must integrate that data. The amount of data that can be stored for immediate, accurate recall (*availability*) is limited in size, and the speed with which it can be recalled (*accessibility*) varies. Ideal WM function, then, would increase both the accuracy of recall and the rate at which information in WM can be accessed.

WM is classically discussed in terms of two different subsystems or components: *visuospatial* WM, which represents, manipulates, and briefly maintains information in the spatial domain; and *verbal* WM, which handles verbally mediated representations and processing (Baddeley & Hitch, 1974; Baddeley & Logie, 1999).¹ More recent theories of WM are process-oriented rather than structural. Probably the most influential model in this regard is Cowan's (1995, 2001, 2005)

model. Cowan proposed a two-tier structure for WM, distinguishing a zone of privileged and immediate access—the focus of attention—from activated but not immediately accessible long-term memory. Memory in the focus of attention is highly accessible and available,² but the focus of attention is capacity-limited to a fixed number of items, or *chunks*. The activated portion of long-term memory is not capacity-limited, but memory in this state is prone to forgetting due to interference and/or decay. Attentional control processes are responsible for manipulating the contents of WM. Among other things, these processes activate, focus, update, switch, and inhibit memory during information processing. Here and for the remainder of this article, we will use the terms “attentional control processes” and “executive function” synonymously, which is consistent with Engle and Kane's (2004) highly influential executive-attention theory of the variation in WM capacity.

Empirical support for the role of WM in complex cognition has come from the finding that WM capacity is a reliable predictor of performance on a wide variety of learning and high-level cognitive tasks, including tasks that tap general fluid intelligence (Engle, Tuholski, Laughlin, & Conway, 1999), reasoning ability (Kyllonen & Christal, 1990), mathematical ability (Ashcraft & Krause, 2007), and spatial ability (Kane et al., 2004). WM is an important component in many learning processes, including taking notes, following directions, or ignoring distractions (Engle, 2001; Engle, Carullo, & Collins, 1991; Piolat, Olive, & Kellogg, 2005).

Evidence suggests that WM is also an important part of language comprehension. Speakers with larger WM capacity are better able to learn vocabulary (in both first and second languages), write more proficiently, and have better L1 reading and listening comprehension (Atkins & Baddeley, 1998; Daneman & Hannon, 2007; Engle, 2001). Individual differences in WM—that is, the extent to which normal adults vary in their WM capacity—should therefore be important for understanding differences in these and other text comprehension processes. For instance, people differ in (1) the ability to remember new information encountered while reading, (2) the ability to make inferences about information encountered while reading, (3) the ability to access knowledge from long-term memory, and (4) the ability to integrate new information with knowledge from long-term memory (Daneman & Hannon, 2007). Because WM plays an important role in these broader cognitive processes and abilities, it comes as no surprise that WM is considered to be one of the most critical components of cognitive and linguistic achievement.

The measurement of WM capacity, which reflects individual differences in the efficacy with which the WM system

¹ A topic of debate in the literature involves whether verbal WM can be further divided into two subcomponents (or resource pools) that handle (1) verbal (but not syntactic) processes for cognitive tasks generally, and (2) syntactic/grammatical processes that are at work specifically during linguistically mediated tasks, such as sentence processing and comprehension (Caplan & Waters, 1999, *inter alia*). Some researchers favor this subdivision, whereas others claim that linguistic and nonlinguistic (but still verbally mediated) tasks rely on a single pool of WM resources (Fedorenko, Gibson, & Rohde, 2007; MacDonald & Christiansen, 2002, *inter alia*). Though it is crucial to theories of human information processing, this domain-general versus domain-specific argument is beyond the scope of the present discussion.

² The *accessibility* of activated memory is defined by the time needed to retrieve it, whereas *availability* is defined by the probability of accurate retrieval (McElree, 2001).

functions (see Shipstead, Harrison, et al., 2013), is often separated between tasks that measure an individual's ability to store and rehearse information—the so-called “simple” span tasks—and those that measure an individual's ability to store information while faced with additional processing tasks—often termed “complex” span tasks. Simple span tasks, such as the forward digit span, word span, or nonword span, require an individual to recall a string of nonrelated letters, words, digits, or visual objects after a brief period of presentation. Complex span tasks, on the other hand, require an individual to actively process input (e.g., a sentence or a simple mathematical equation) while remembering a string of letters, words, digits, or objects.

A meta-analysis by Daneman and Merikle (1996) revealed that complex span tasks were better predictors of L1 comprehension than simple span tasks. It is also of interest to note that their findings did not depend on the nature of the stimuli in the complex span tasks. That is, operation span tasks (a predominantly nonlinguistic task) accounted for just as much variance of the criterion measures as did reading/listening span tasks that required the processing of linguistic material. The latter findings support the widely held notion that the executive control component of WM is a domain-general system (e.g., Baddeley, 2007).

A preliminary synthesis, which examined 16 studies focused on the role of WM in second language acquisition (SLA), featured a mean correlation coefficient (r) of .18 (Watanabe & Bergsleithner, 2006), suggesting that WM is positively related to L2 proficiency outcomes. Although few would argue that WM is unimportant for L2 processing, some researchers contend that WM's importance has been overstated (e.g., Juffs & Harrington, 2011). This debate has been fueled not only by inconsistent results, but also by diverse research methodologies, which have led to difficulties in qualitatively comparing studies. In the following sections, we will review several studies that have examined the relationship between WM and L2 processing and proficiency development, identifying design factors that may have contributed to the diversity in the reported WM effects. The review below will provide a broad overview of the literature included in the meta-analysis, so as to justify the isolation of specific predictor variables. For a more in-depth discussion of specific studies, the interested reader is referred to Juffs and Harrington (2011) and Williams (2011).

Levels of proficiency

L2 processing (i.e., production and comprehension) generally requires more cognitive resources than does processing in the L1 (e.g., Green, 1998; Hernandez & Meschyan, 2006). Therefore, it is reasonable to argue that individuals with greater WM resources would perform better on processing tasks in the L2. However, a collection of studies have suggested that

proficiency level may moderate this processing advantage. For example, according to Abu-Rabia (2001), Hummel (2009), Leiser (2007), Linck, Hoshino, and Kroll (2008), and Weissheimer and Mota (2009), low-proficiency bilinguals with greater WM spans performed significantly better than those with lower span scores on tasks addressing L2 processing abilities. However, when examining highly proficient bilinguals, several studies failed to show significant L2 processing advantages for individuals with higher WM scores (Fehrer & Fry, 2007a, 2007b; Foote, 2011; Hummel, 2009).

Working memory span task variables

The literature is quite inconsistent when it comes to selecting the language in which to measure WM capacity. Research by Osaka and Osaka (1992) indicated a strong positive correlation between WM span tasks administered in the L1 and those administered in the L2, supporting views that WM is a domain-general resource. However, the L2 proficiency of the participants was a factor to consider when deciding the language of the WM span task. To resolve this issue, researchers sometimes administer the same class of WM span task in both the L1 and the L2. Previous research has concerned itself with differences in the domain (i.e., verbal vs. nonverbal) content of the WM span task. Daneman and Merikle (1996) found that the content domain of the WM span task did not contribute to substantial differences in the amounts of variance explained. Although the content of the domain may not matter to a significant degree, Daneman and Merikle's meta-analysis did show that complex span tasks were much better predictors of L1 comprehension performance than were simple span tasks. In L2 research, both simple and complex span tasks have been found to significantly predict L2 processing and proficiency outcome measures (for simple span, see, e.g., Christoffels, De Groot, & Waldorp, 2003; O'Brien, Segalowitz, Freed, & Collentine, 2007; Slevc & Miyake, 2006; for complex span, see, e.g., Abu-Rabia, 2001; Mackey & Sachs, 2012; Révész, 2012). However, it is difficult to determine which of these span tasks is a better predictor of L2 processing and proficiency tasks, due to the heterogeneity of the L2 tasks used and other research design factors.

Simple span tasks consistently account for significant amounts of variance in L1 vocabulary learning tasks and outcome measures (Gathercole, Willis, Emslie, & Baddeley, 1992). Similar results have been found in L2 learning studies (see Williams, 2011, for a review). Several researchers have suggested that the role of phonological STM is greater in less-proficient bilinguals (Cheung, 1996; Juffs & Harrington, 2011). Studies by Abu-Rabia (2001) and Speciale, Ellis, and Bywater (2004) showed that simple span measures correlated significantly with L2 lexical development. However, not all studies have shown relationships between simple span measures and L2 lexical development. For example, Akamatsu (2008) failed to find a significant correlation between word

span (a simple span task) and gains made on a word-recognition task following a 7-week word recognition training period. One potential explanation for these inconsistencies is that changes in the speed of lexical retrieval likely reflect different processes or mechanisms than those contributing to knowledge acquisition that have been measured in other studies (e.g., L2 vocabulary development). Akamatsu even commented on the relative lack of cognitive demand in the word recognition tasks, another potential factor that may have led to the null findings. The rationale is that tasks that consume many cognitive resources will give individuals with high WM capacities an inherent advantage over those with lower WM capacities. If a particular task is so easy that all participants can perform it without much effort, WM capacity differences are much less likely to be found.

Studies spanning the past two decades have indicated that WM plays a role in L2 processing and proficiency development (e.g., Michael & Gollan, 2005; Williams, 2011). However, since the results across studies have been inconsistent, the precise nature of this role remains unclear, and Watanabe and Bergsleithner's (2006) preliminary synthesis did not examine the potential influence of covariates on the magnitude of the population effect size. Our review of the literature identified a number of research design factors that may have contributed to the heterogeneity of the research findings. Specifically, we identified three categories of potential moderators of the relationship between WM and L2 outcomes, including characteristics of the WM measures, features of the criterion measures, and the proficiency of the participants included in the study (see Table 1). Given the interest in WM's impact on L2 processing and proficiency development, and the number of studies now available in the literature, it is time for a quantitative synthesis of the extant results.

Table 1 Categories of variables examined as potential moderators of the relationship between working memory and second language criterion outcomes

Variable	Levels
WM Task Features	
Content ^a	Verbal, math, visuospatial
Language	L1, L2
Complexity	Simple, complex
Criterion Task Features	
Focus	Proficiency, processing
Skills	Comprehension (reception), production
Test type	Standardized, nonstandardized
Participant Features	
Proficiency	Less proficient, highly proficient

^a Measures employing math and visuospatial content were collapsed into a “nonverbal” content factor for the analyses. L1 = native language. L2 = nonnative language

The present meta-analysis

The goals of our meta-analytic review were twofold. First, we wanted to estimate the population WM effect size, on the metric of the correlation coefficient. Second, we wanted to examine the potential moderating influences of relevant variables, to better understand the boundary conditions of WM effects. To our knowledge, this is the first exhaustive quantitative synthesis of studies of WM effects in the L2 literature that has taken such covariates into account, and as such, it represents a major step forward in the field's understanding of the relationship between WM and L2 processing and proficiency outcomes.

Method

Literature search

Studies were located online via keyword searches in databases (Academic Search Premier, Dissertation Abstracts International, ERIC, PsycINFO, and the Psychology and Behavioral Sciences collection) and in the Google Scholar search engine. All of the searches used variations of the following terms: *second language*, *foreign language*, *bilingual** (with the asterisk serving as a wildcard operator), *working memory*, *working memory capacity*, *WMC*, *working memory span*, *short-term memory*, *short-term memory span*, *reading span*, *listening span*, *operation span*, *digit span*, *nonword span*, *word span*, and *letter span*. Tables of contents were inspected in peer-reviewed journals that focus on SLA- and bilingualism-related topics (i.e., *Applied Psycholinguistics*, *Bilingualism: Language and Cognition*, *Language and Cognitive Processes*, *Language Learning*, *Second Language Research*, and *Studies in Second Language Acquisition*). The reference lists of publications located through these search methods were also inspected to identify studies cited therein. Finally, for each study on this interim list, a “cited by” search was conducted to identify more recent articles that have cited the target reference. Our search included published articles and book chapters, as well as unpublished masters theses and doctoral dissertations that were available in the databases as of September 19, 2012. In order to provide a comprehensive analysis and to mitigate the “file drawer problem” (i.e., publication bias; see below), unpublished studies were included in the meta-analysis (e.g., Rosenthal, 1979). All studies included in the meta-analysis are identified with an asterisk (*) in the References list.

Inclusion criteria

A set of inclusion criteria was designed to focus the meta-analysis on studies relevant to understanding the role of WM in adult L2 proficiency and processing outcomes. Each study

was examined to identify whether it satisfied the following set of criteria.

1. All participants were classified as adults (above the age of 18).
2. Participants were classified as “nonnative bilinguals.” Here, we use the term “bilingual” liberally to refer to an individual with at least minimal knowledge of an L2, and “nonnative” to refer to individuals who began learning an L2 after first becoming proficient in a primary (native) language. This criterion excluded heritage speakers and childhood bilinguals who acquired both languages simultaneously as children.
3. No participant had a known history of neurological or psychopathological problems (including learning and language impairments).
4. In each study, at least one WM measure and one L2 outcome measure (assessing an aspect of processing and proficiency) were administered. Studies using performance measures that required participants to learn non-words or artificial grammar rules were not included in the analysis, to restrict the analysis to studies of natural L2 processing and/or proficiency outcomes.
5. It was necessary that each study quantify the relationship between the WM measures and the criterion measures through either a Pearson product–moment correlation coefficient (r) or another statistic (e.g., t , Cohen’s d , or F) that could be transformed into a correlation coefficient (see the [Appendix](#) for equations). Following standard meta-analytic procedures, results from analyses of variance were included only for F statistics with one degree of freedom in the numerator (e.g., Rosenthal, 1995). When a study simply stated that an effect was nonsignificant, without reporting an actual effect size, the effect size was assigned an estimate of $r = 0$ for the main analyses reported below. However, this is known to lead to conservative, downward-biased population estimates, and therefore an alternative approach—excluding the effect size—was conducted as part of a “sensitivity analysis” (Rosenthal, 1995). Note that this alternative approach is itself known to introduce upward bias in the population estimate. The sensitivity analysis allowed for an examination of the extent to which the inferences drawn from the meta-analyses were sensitive to these decisions.

Because nearly all of the studies failed to report knowledge of languages other than the two being empirically addressed, we were unable to control for proficiency in additional languages. The criteria above led to a final data set of 748 effect sizes from 79 independent samples involving 3,707 participants. See the [online supplemental materials](#) for a table that provides the following information for each of the 79 independent samples: study reference, sample size, median

correlation coefficient, range of correlation coefficients, participant proficiency level, publication status, the coding results for the WM measures and criterion measures (see below), and the specific WM task(s) and criterion measure(s) used in a study.

Variables and coding procedures

The literature review identified several variables likely to influence the strength and/or direction of effect sizes. These variables categorized relevant characteristics of the WM tasks, the criterion tasks, the participants, and the publication status of the study. We will review the variable-coding procedures in turn.

WM span tasks The WM span tasks varied on a number of factors, both within and across studies (see Table 2). First, the language of performance was classified as either L1 or L2, with tasks requiring the processing or storage of numeric stimuli coded as L1, since numeric calculation is typically performed in the L1. Second, tasks were classified as *simple* span tasks (i.e., measuring *storage only*) or *complex* span tasks (requiring both *storage and processing*; see Daneman & Merikle, 1996; Unsworth & Engle, 2007). Finally, WM measures were classified according to the content domain of the stimuli: *verbal* (i.e., requiring processing of linguistic material such as words or sentences) or *nonverbal* (i.e., requiring processing of nonlinguistic material, including numeric digits, math equations, or visuospatial images). For complex span tasks that included both verbal and nonverbal stimuli, this variable was determined on the basis of the content of the processing component. For example, in the operation span task, participants must process (make judgments about) simple arithmetic problems while storing words or letters; because the processing component was nonverbal, the operation span was classified as a nonverbal span task (see Daneman & Merikle, 1996, for a similar classification of the task).

L2 performance measures Criterion measures of L2 performance were classified on the basis of the modality of the measure—namely, *comprehension* (e.g., lexical decision task), *production* (e.g., cloze test), or *both* (e.g., simultaneous interpretation: composite measures combining separate indicators of comprehension and production). In addition, each criterion measure was classified as focusing on language *processing* or *proficiency*. Processing measures gauged online language-processing abilities, such as those measured by gating tasks (e.g., McDonald, 2006), hesitation phenomena (e.g., Fehrer & Fry, 2007b), fluency during an oral proficiency interview (e.g., O’Brien, Segalowitz, Collentine, & Freed, 2006; O’Brien et al., 2007), speech generation tasks (e.g., Weissheimer & Mota, 2009), or lexical recognition tasks (e.g., Leiser, 2007). Proficiency measures, on the other hand,

Table 2 Classification system for coding working memory span tasks

	Verbal	Nonverbal
Simple	Word span	Digit span
	Nonword span	Counting span
	Letter span	Backward digit span
		Letter rotation
		Size judgment
Complex	Reading span	Operation span
	Listening span	Math span
	Speaking span	<i>N</i> -back
	English opposites span	AMIPB

assessed L2 knowledge or more general language abilities. These included standardized tests of proficiency, such as the Michigan Test (e.g., Juffs, 2005) and the grammar and reading sections of TOEFL (e.g., Harrington & Sawyer, 1992), as well as nonstandardized tests of vocabulary (e.g., Hummel, 2009). We also noted whether these criterion measures used a *standardized* measure (e.g., TOEFL section scores, Michigan Test scores) or a *nonstandardized* measure (e.g., cloze test, grammaticality judgment task). A list of the various criterion measures found in the studies and their classifications within this coding scheme is presented in Table 3.

Participant L2 proficiency The studies included in this meta-analysis varied greatly in how they described and/or quantified the L2 proficiency of their sample populations. For this meta-analysis, participants were categorized as either *highly proficient learners* (i.e., having extensive academic or professional exposure and/or intensive immersion experience) or *less-proficient learners*. Individuals labeled as highly proficient learners met one or more of the following criteria: (1) international students enrolled in an academic program (undergraduate or graduate) administered entirely in the participant's L2, (2) masters- or PhD-level students specializing in the foreign language of study, or (3) professionals who were functioning completely in their foreign language and had begun learning their foreign language during adulthood. Individuals not meeting any of these criteria were labeled as less-proficient learners.

Publication status Each study was coded as *published* or *unpublished*. This coding allowed for an examination of the potential for *publication bias*—the systematic underreporting of smaller effect sizes due to nonsignificant null hypothesis significance tests. If a meta-analysis focuses solely on published studies, the researcher risks inflating the estimated population effect size (see Rosenthal, 1979). Therefore, a concerted effort was made to include unpublished reports, including master's theses and dissertation studies. Their inclusion allowed us to explicitly compare the estimated effect

sizes from published versus unpublished studies, while also mitigating to some extent threats of publication bias.

Some unpublished studies that were identified during the literature search phase included subsets or supersets of participants whose data were subsequently published. The published and unpublished reports often contained nonidentical analyses (e.g., the dissertation contained the full correlation matrix for the predictors and outcomes, whereas the published article only reported targeted correlations to test specific hypotheses) and/or participant samples (e.g., the sample reported in the dissertation was supplemented with subsequently tested participants or was combined with another sample for the published article). Meta-analysis assumes and requires the independence of participants between different studies in the meta-analyzed data set (e.g., Hedges, Tipton, & Johnson, 2010). Therefore, when the samples of two studies overlapped, we included in the data set the study with the larger, more encompassing sample, and excluded the other study. In most cases, this led to the inclusion of the published article and the exclusion of the unpublished study. However, in one case (Fortkamp & Bergsleithner, 2007), the published article appeared to include a subset of the sample from the unpublished dissertation (Bergsleithner, 2007), and therefore the unpublished dissertation with the larger sample was included in place of the published article. A few studies were also reported in university or departmental bulletins (e.g., Ikeno, 2006, published in the *Bulletin of the Faculty of Education*). Although it was not easy to ascertain the extent of peer review for these venues, the bulletins were regularly produced and published by the universities, and therefore the studies reported therein were coded as published.

Interrater agreement

The data coding was performed by the first and second authors, with approximately 70% of the data being coded independently to check for interrater agreement. Across all coded variables, agreement ranged from 86% to 100% (median = 98.9%). After the initial coding was completed, disagreements were discussed and resolved by providing further specification of our coding scheme, where needed. The lowest agreement of 86% was found for the criterion modality variable, and was driven by disagreements on the criterion outcomes from three specific studies reporting multiple outcomes.

Analytic approach

The goal of this meta-analysis was to estimate the mean of the population distribution of effect sizes and to generalize the results beyond the sample of examined studies. Therefore, random-effects models were employed (Hedges & Vevea, 1998). Most studies reported multiple effect sizes (a median of four *r*s per sample), often due to the inclusion of different

Table 3 Classification system for coding second language criterion measures

Criterion Focus and Modality	Criterion Test Type	
	Standardized	Nonstandardized
Proficiency		
Comprehension	TOEFL2 Grammar test	Inferential comprehension test
	TOEFL2 Reading test	Literal comprehension test
	Michigan test	Overall reading comprehension test
Production	Woodcock Reading Masters test—Spelling	Syntactic comprehension test
		Vocabulary test
		Grammar test
		Cloze test
		Narrative monologue test
		Word translation test
		Interpretation task
		Accuracy rate
		Complexity
		Sentence-level translation proficiency
Both	—	Global proficiency test
		Language course final exam
Processing		
Comprehension	—	Phonological discrimination task
		Orthographic discrimination task
		Visual word verification
		Lexical decision task
		Grammaticality judgment task
		Gating task
Production	Woodcock Reading Masters test—Real-word read aloud Woodcock Reading Masters test—Pseudoword read aloud	Hesitation phenomena
		Read aloud task
		Picture description
		Speech rate
		Sentence production task
		Simultaneous interpretation task
Both	—	

types of WM measures (e.g., simple and complex span tasks, or L1 and L2 administrations). This violates the assumption of independence among effect sizes in standard meta-analytic procedures (Hedges et al., 2010). Various methods have been proposed to address this violation, such as collapsing these dependent effect sizes into one “synthetic” effect size per study by computing the mean or median effect or by randomly selecting one effect size per study (see Marin-Martinez & Sanchez-Meca, 1999, for a comparison of the methods). A recently proposed alternative is to explicitly model the interdependence among effect sizes by robust variance estimation (Hedges et al., 2010), thereby eliminating the need to discard information through aggregation. Therefore, we employed robust variance estimation procedures in the R statistical software package (R

Development Core Team, 2012) using the R code provided in Hedges et al.’s [appendix](#).

As part of sensitivity analyses, we also followed an aggregation approach by first computing the median effect sizes for each study, then conducting standard random-effects meta-analyses using the “metafor” R package (Viechtbauer, 2010). Broadly speaking, the results paralleled those of the robust standard error (*SE*) method, and we report the results of both approaches below (see the “[Complete-data meta-analysis](#)” section). However, since we believe the robust *SE* method was better suited to our specific data set, for the covariate analyses we only report the results from the robust *SE* analyses.

Following Higgins and Thompson (2002), rather than focusing on a categorical significance test of effect size heterogeneity between studies, we focused on *quantifying* the degree of

heterogeneity in the analyzed effect size by reporting τ^2 . We planned a priori to examine a number of potential covariates in order to address theoretical claims that have been posed in the literature, regardless of whether a hypothesis test determined that a significant amount of heterogeneity was present between studies. Moreover, a number of these covariates were observed within studies (e.g., both simple and complex span tasks), suggesting that characterizations of the between-study variance would not provide a complete picture of the results.

Correlation coefficients are known to be nonnormally distributed, and thus the recommended effect size for meta-analysis is Fisher's z transform of r (e.g., Schafer, 1999). However, in the results below, we report effect sizes and 95% confidence intervals (CIs) on the original r metric to facilitate interpretation. See the [Appendix](#) for the equations used to convert effect sizes between r and Fisher's z .

File drawer analysis

A concern for any meta-analysis is publication bias—the potential for more extreme results to be overly represented in the literature, due to biases against publishing nonsignificant effects (Rosenthal, 1979). We assessed the presence of publication bias in our sample by multiple methods. First, we computed a fail-safe N (Orwin, 1983; Rosenthal, 1979), which computes the number of missing, unpublished, or future studies with null effects that would be required to render the probability of a Type I error for a significance test of $\hat{\rho}$ to increase above an acceptable level. Orwin suggested a variation on Rosenthal's fail-safe N that identifies the number of studies with a particular effect size (e.g., null) that would be required to alter the observed effect size to reach a designated criterion (e.g., $\hat{\rho} = .01$) that the meta-analyst believes would call into question the validity of the findings. For each analysis, we computed Orwin's fail-safe N , using a criterion effect size of $r = .01$ and an effect size of $r = 0$ for missing studies (see the [Appendix](#) for the equation and details). This approach provides a sense of the stability of the findings of the meta-analysis, with a suggested rule of thumb being that results are valid and robust against the “file drawer problem” if the fail-safe N reaches or surpasses the value of $5k + 10$, where k is the number of studies in the analysis (Rosenthal, 1979).

Covariate analyses

We examined the potential moderating influences of the categorical covariates identified in the literature review by first

³ In recent years, some have called for researchers to examine potential covariates using meta-regression techniques (e.g., Sutton & Higgins, 2008). However, in the literature being synthesized in this study, most covariates varied both between and within studies. Meta-regression techniques were designed to explain variability between studies only, and therefore were not appropriate for these covariates.

creating subsets of the data set for each level of the covariate, then separately fitting random-effects models with robust standard errors to each subset.³ We also examined theoretically motivated interactions between the covariates, where sufficient data were available in each data subset to draw reasonable inferences. We report the results of two such interactions—Language \times Complexity and Focus \times Complexity.

Results and discussion

Prior to the analysis, extreme outliers were removed in order to prevent any undue influence on the inferences. Outliers were identified as any effect size more than twice the interquartile range above or below the median effect size (see Fig. 1). This criterion was applied to Fisher's- z -transformed data, since these were the values submitted to the meta-analyses. This procedure identified 40 observed effect sizes—2/3 positive, 1/3 negative—corresponding to the following unique values (converted to the r metric): $-.97, -.62, -.57, -.53, -.48, -.42, -.39, -.37, -.36, .66, .67, .68, .69, .70, .71, .72, .73, .74, .76, .79$, and $.80$. These effect sizes were excluded prior to analysis in order to avoid any undue influence of extremely unlikely values. Note that this procedure removed twice as many positive as negative correlation values. This is not problematic, because it suggests that the most likely effect of this outlier removal procedure would be to attenuate any potential positive inflation of the population effect size estimate due to publication bias.

Complete-data meta-analysis

Descriptive statistics Table 4 provides descriptive statistics to characterize the sample of studies (e.g., number of independent samples, total number of participants) and the effect sizes included in the analyses (e.g., proportion of correlations with positive values, minimum and maximum observed values). Descriptive statistics are provided first for the complete data set, then separately for each covariate subset analysis.

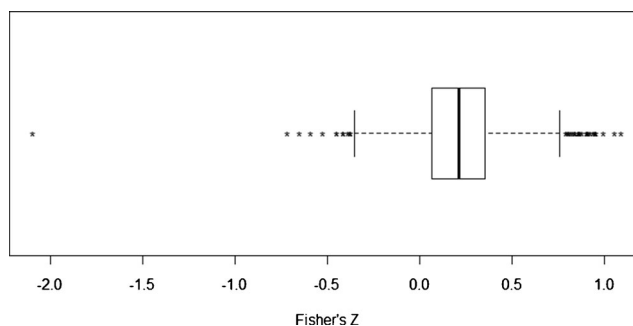


Fig. 1 Boxplot of the distribution of the analyzed effect sizes (Fisher's z values) in the full data set, with outliers identified by asterisks (see the text for the corresponding values on the r metric)

Table 4 Descriptive statistics for the main complete-data analysis, individual covariate subset analyses, and covariate interaction subset analyses

Covariate	Subset	<i>k</i>	<i>N</i>	<i>n</i>	Number of <i>r</i> s	Median No. <i>r</i> s per Study	% Positive	Min <i>r</i>	Median <i>r</i>	Max <i>r</i>
Complete Data										
–	–	79	3,707	42	748	4	84.1	–.340	.202	.640
WM Covariates										
Language	L1	53	2,701	42	393	4	87.0	–.300	.191	.630
	L2	46	2,203	35	355	4	80.8	–.340	.213	.640
Complexity	Simple	29	1,634	43	272	5	85.7	–.340	.190	.610
	Complex	70	3,161	38	476	4	83.2	–.290	.215	.640
Content	Verbal	71	3,305	40	625	4	86.7	–.340	.216	.640
	Nonverbal	22	1,283	43	123	3	70.7	–.290	.120	.560
Criterion Covariates										
Focus	Processing	32	1,194	36	271	4	76.8	–.290	.223	.640
	Proficiency	60	3,139	44	477	4	88.3	–.340	.193	.640
Modality	Comprehension	43	2,411	49	210	3	81.4	–.271	.187	.620
	Production	42	1,712	37	412	5	83.0	–.340	.223	.640
	Both/aggregate	11	588	40	126	6	92.1	–.300	.168	.610
Type	Standardized	14	862	45	49	3	91.8	–.164	.240	.615
	Nonstandardized	72	3,205	41	699	4	83.5	–.340	.199	.640
Study Covariates										
Participant proficiency	Learners	70	3,362	42	644	4	86.2	–.340	.208	.640
	Bilinguals	9	345	21	104	5	71.2	–.290	.180	.640
Publication status	Published	62	2,805	42	470	4	83.4	–.340	.240	.631
	Not published	17	902	36	278	8	85.3	–.282	.152	.640
Covariate Interactions										
Language × Complexity	L1 simple	25	1,406	43	190	5	86.8	–.300	.180	.570
	L2 simple	13	901	44	82	4	82.9	–.340	.208	.610
	L1 complex	45	2,268	40	203	2	87.2	–.160	.211	.630
	L2 complex	43	1,894	32	273	3	80.2	–.290	.220	.640
Focus × Complexity	Processing, simple	8	312	42	56	5	85.7	–.140	.264	.490
	Proficiency, simple	24	1,461	44	216	6	85.6	–.340	.158	.610
	Processing, complex	30	1,101	33	215	4	74.4	–.290	.200	.640
	Proficiency, complex	51	2,593	42	261	3	90.4	–.271	.221	.640

k = number of studies. *N* = number of independent participants. *n* = median number of participants per study. % positive = percentage of effect sizes greater than zero

Inferential statistics/results A random-effects model with robust standard errors was fit to the full data set to compute $\hat{\rho}$ —the estimate of the mean value of ρ in the population distribution of effect sizes (Hedges et al., 2010). The results are reported in Table 5, including $\hat{\rho}$, 95% CIs, and fail-safe *N*. The analysis suggests that the population distribution of ρ is centered around a value of .255 and is significantly positive, as indicated by the 95% CIs that do not overlap with zero. In fact, across all analyses reported below, none of the 95% CIs overlapped with zero, indicating a robust positive relationship between WM and L2 outcomes across the range of covariates investigated here.

For the complete-data analysis, the amount of between-study heterogeneity, τ^2 , was estimated as being .017. For the

results reported in Table 5, we specified the within-study correlation between effect sizes as being .80. Following Hedges et al. (2010), to check whether the results were sensitive to this value, we estimated τ^2 , $\hat{\rho}$, and the *SE* of $\hat{\rho}$ across correlation values ranging from 0 to 1 in increments of .1. Across these values, we found that τ^2 ranged from .0167 to .0168, that $\hat{\rho}$ was equivalent to the fourth decimal place, and that the *SE* of $\hat{\rho}$ was equivalent to the fifth decimal place.

As part of a sensitivity analysis, a random-effects model was also fit to the aggregated data, with a median effect size being computed for any sample with more than one effect size. Following recommendations that meta-analysts report by-study effect sizes with CIs (e.g., Moher, Liberati, Tetzlaff, Altman, & the PRISMA Group, 2009), Fig. 2 presents a forest

Table 5 Random-effects meta-analysis results for the complete-data analysis and for models examining individual covariates and the covariate interactions

Covariate	Subset	$\hat{\rho}$	95% CI		Fail-Safe <i>N</i>
			(Lower)	(Upper)	
Full Data	–	.255	.219	.291	2,007
WM Covariates					
Language	L1	.228	.184	.271	1,186
	L2	.299	.245	.354	1,398
Complexity	Simple	.175	.128	.222	487
	Complex	.272	.232	.312	1,909
Content	Verbal	.264	.223	.305	1,873
	Nonverbal	.181	.133	.229	382
Criterion Covariates					
Focus	Processing	.242	.181	.303	766
	Proficiency	.253	.212	.295	1,511
Modality	Comprehension	.242	.191	.292	1,027
	Production	.267	.221	.314	1,124
	Both/aggregate	.212	.137	.287	227
Type	Standardized	.310	.214	.406	443
	Nonstandardized	.242	.205	.279	1,727
Study Covariates					
Participant proficiency	Learners	.255	.216	.294	1,777
	Bilinguals	.259	.165	.353	233
Publication status	Published	.266	.228	.305	1,652
	Not published	.218	.129	.307	363
Covariate Interactions					
Language × Complexity	L1 simple	.168	.115	.220	400
	L2 simple	.191	.098	.284	240
	L1 complex	.238	.188	.287	1,056
	L2 complex	.327	.271	.384	1,446
Focus × Complexity	Processing, simple	.214	.069	.358	167
	Proficiency, simple	.167	.124	.210	382
	Processing, complex	.247	.183	.312	736
	Proficiency, complex	.271	.223	.319	1,387

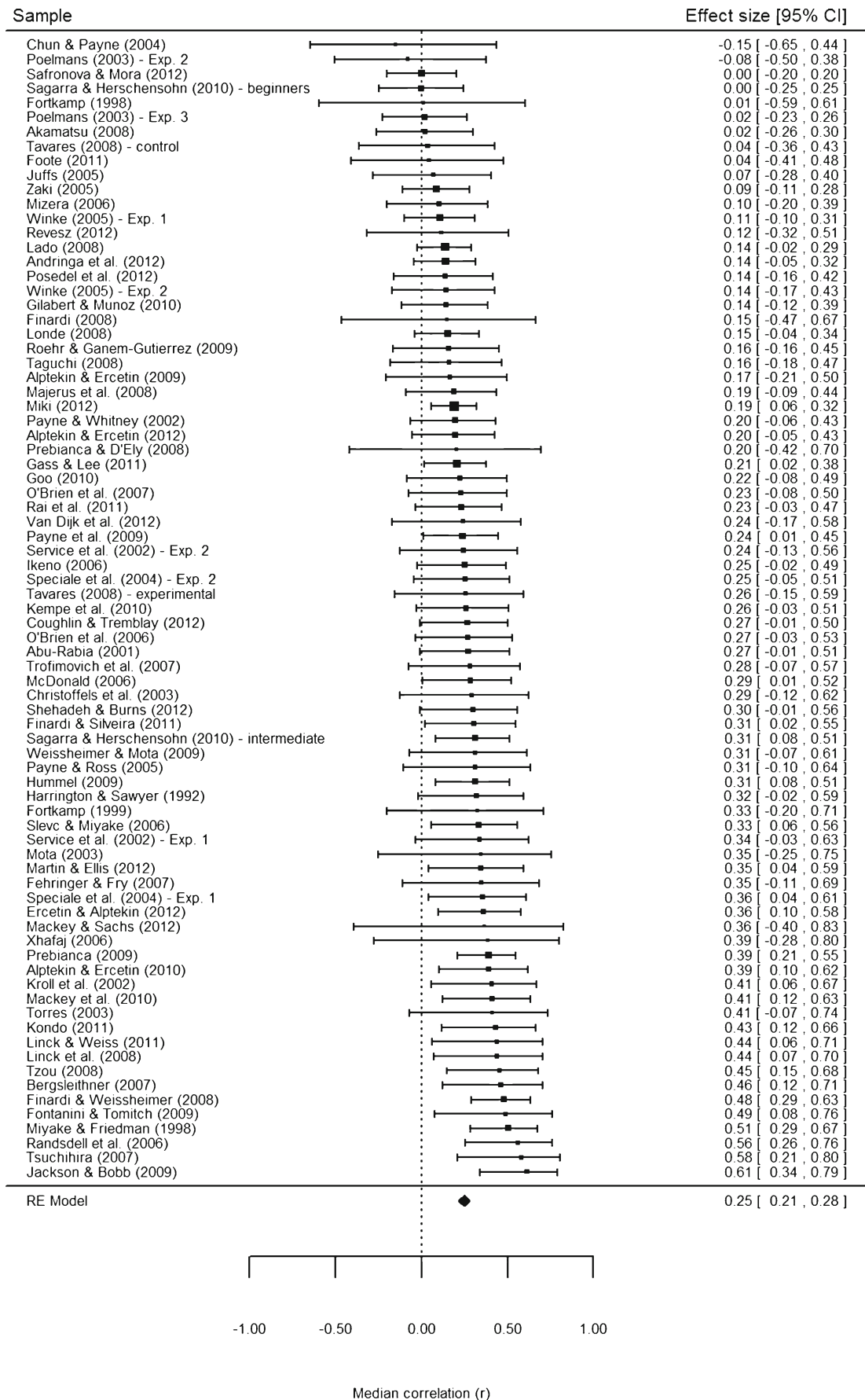
The full-data results were computed using .80 for the within-study correlation between effect sizes. 95% CI = 95% confidence interval (lower and upper limits) of $\hat{\rho}$. Fail-safe *N* was computed using Orwin's (1983) method (see the Appendix for details)

plot based on the aggregated-data random-effects model, after converting the effect sizes back to the *r* metric. Forest plots provide a visual depiction of the effect sizes estimated from a meta-analytic model, which can provide useful information regarding the nature of the distribution and precision of the effect sizes across samples.

Two patterns are worth noting. First, the population estimates were very similar in the two analyses (aggregated-data model, $r = .253$, CI = [.216, .289]; robust *SE* model, $r = .255$, CI = [.219, .291]). Second, one might be concerned that the inclusion of data from extremely small samples might bias the population effect size estimate. However, the magnitude of the

estimated effect size does not appear to be related to sample size, which is indicated by the width of the CIs (for correlation coefficients, CI widths are directly proportional to the sample

Fig. 2 Forest plot of effect sizes for each independent sample, based on a random-effects model fit to the aggregated data (median Fisher's *z*). The estimated effects were transformed back to the *r* metric prior to generating the forest plot, to facilitate interpretation. Each independent sample is shown on a separate row, and the points (and bars) represent the estimated effect size (and CIs). The bottom-most row depicts the population value and *SE*, as estimated from the aggregated-data random-effects model. The dashed vertical line indicates an effect size of zero (i.e., no relationship)



size). Indeed, the smallest samples (i.e., those effects with the widest CIs) appear across the entire range of the distribution of effect sizes, suggesting that the population estimate was not biased by the inclusion of particularly small samples.

In the aggregate-data random-effects model, approximately 12.4% of the variance in effect sizes was due to heterogeneity between the studies ($\tau^2 = .003$, $SE = .004$). A comparison with the complete-data robust *SE* results indicated that the degree of heterogeneity in the effect sizes appears to have been underestimated by the mean-aggregated random-effects model—as often happens with a samplewise aggregation procedure (Cheung & Chan, 2004). Although the random-effects meta-analysis model suggests that not much residual heterogeneity between studies needs to be explained by covariates, a number of theoretically motivated covariates were identified in the literature review, many of which may explain variability both between and *within* studies. Therefore, we now turn to a systematic examination of these covariates.

Covariate subset analyses In all of the analyses reported here, a series of random-effects models with robust standard errors were fitted separately for each covariate. The results are reported in Table 5 following the random-effects meta-analytic model results. Overall, the subset analyses corroborate the findings of the full-data meta-analysis. The central tendency of the $\hat{\rho}$ estimates from the individual and interaction covariate models is approximately centered around the full-data meta-analysis estimate (mean $\hat{\rho} = .241$, median $\hat{\rho} = .245$). Moreover, all of the covariate analysis 95% CIs cover positive values and exclude zero, indicating significant positive correlations between WM and L2 outcomes, even when accounting for a range of covariates. We now consider each set of covariates in turn.

Characteristics of WM measures The analysis of the language of the WM measure suggests that larger correlations between WM and L2 outcomes may be found when WM is measured in the L2 rather than the L1, with the L2 WM correlation estimate being .30. However, the partially overlapping 95% CIs indicate that this difference may not be robust. We argue that any difference is likely due to the confounding of L2 proficiency with WM abilities when WM tasks are administered in the L2. That is, to the extent that the WM task performance requires L2 use, the task will be an indicator of both WM abilities and L2 proficiency, and therefore will not purely measure WM. In the context of predicting L2 outcomes, this confound would inflate the WM–outcome correlation estimate. Indeed, the L2 WM covariate analysis estimated the third highest population value, and was one of only three estimates reaching .30. Therefore, these results suggest that researchers who wish to isolate the true relationship between WM and L2 proficiency should employ L1 measures, to provide a purer estimate of WM abilities.

A significantly stronger correlation was found for complex WM span tasks relative to simple span tasks (see the nonoverlapping CIs). This finding parallels the results of Daneman and Merickle's (1996) meta-analysis of WM and L1 reading comprehension, in which complex (process-plus-storage) measures of WM were stronger predictors than simple (storage-only) measures. These results indicate that, across a range of L2 processing outcome measures, better WM abilities were related to better performance. Research on L2 aptitude has also implicated WM as an important individual difference, with some arguing that WM is at the core of L2 aptitude (Miyake & Friedman, 1998). These results further corroborate claims that WM is a critical component to any successful theory of L2 aptitude (see DeKeyser & Koeth, 2011).

Note that simple span tasks that measure STM—including phonological STM—also had significant and positive relationships with L2 outcomes. Phonological STM has been identified as an important contributor to L2 aptitude (e.g., Hummel, 2009), including a recent investigation of aptitude for high-level language proficiency (Linck et al., 2013). The present results are congruent with accounts that include both executive control (i.e., WM) and (phonological) STM as abilities that account for individual differences in L2 outcomes. Any comprehensive theoretical model of L2 outcomes—both processing and proficiency—likely should include both WM and STM. It will be useful for future studies to identify the types of tasks and conditions that modulate the relative contributions of WM and STM.

Focusing on the content of the WM measures, the covariate analyses indicated that verbal WM measures were somewhat more highly correlated with L2 outcomes than nonverbal WM measures, although their 95% CIs overlapped slightly. This result replicates the patterns reported in Daneman and Merickle's (1996) meta-analysis of L1 reading comprehension.⁴ According to the multicomponent model of WM (e.g., Baddeley & Hitch, 1974), this difference would be attributed to the functioning of the phonological loop within the WM system. That is, one's facility with processing or manipulating verbal content would be driven by the domain-specific WM component specialized for verbal information. However, contemporary views posit that WM (particularly the central executive) is a domain-general ability that operates independent of language (e.g., Engle, Kane, & Tuholski, 1999). According to such accounts, this verbal/nonverbal difference would simply be due to the overlap in the content being manipulated (i.e., common method bias), despite the fact that the WM system per se is not specialized for or constrained to a specific

⁴ Daneman and Merickle (1996) compared verbal measures to math measures, but excluding visuospatial measures, whereas we included both math and visuospatial measures in our “nonverbal” measure condition.

content domain. These results are congruent with both accounts.

Characteristics of criterion measures The studies examining L2 processing outcomes and L2 proficiency outcomes showed similar magnitudes of correlations, around .25. Note that the processing outcomes were associated with a wider confidence interval and smaller fail-safe N , likely reflecting the smaller number of studies available (less than half that of the proficiency outcomes). Nonetheless, this result highlights the need for WM to be incorporated into comprehensive models of L2 processing as well as theories of SLA. Future research should examine the conditions under which WM affects various aspects of L2 processing and proficiency development, which could help elucidate the role of executive functions to specific L2 processes (for an example, see Robinson, 1995).

With respect to proficiency outcomes, a rich literature has focused on L2 aptitude effects, examining the role of individual differences in cognitive and perceptual abilities (e.g., Carroll, 1985; Grigorenko, Sternberg, & Ehrman, 2000). Contrary to some who have questioned the importance of WM (vs. phonological STM) in theories of aptitude (Juffs & Harrington, 2011), this meta-analysis corroborates claims that WM is correlated with L2 proficiency outcomes and, therefore, is an important component of any theoretical model of such outcomes—including models of L2 aptitude. To the extent that different components of aptitude are relevant to predicting the rate of success at earlier stages of learning versus the attainment of high-level proficiency (e.g., Linck et al., 2013), more research will be needed to contrast the contributions of WM for these two different proficiency outcomes.

Similar correlations were also found for comprehension and production outcomes, as well as for aggregate outcomes tapping into both skills (.24, .27, and .21, respectively), suggesting that WM is relevant to understanding both receptive and productive L2 abilities. Future research could compare and contrast the roles of WM on a specific process, such as lexical access, across different skills (e.g., during reading vs. speech production). Such an approach would further enhance theories of L2 processing and L2 proficiency by increasing the specificity of the role(s) of WM at various levels of analysis and across the various subskills.

No reliable differences in effect sizes were found between standardized criterion measures, such as the TOEFL subtests, and nonstandardized criterion measures, such as grammaticality judgment tasks (see Table 3), with moderate-sized correlations for both outcome types. The estimate, numerically, was slightly higher for standardized criterion measures, although it was also more uncertain (as indicated by the larger CI),

likely due to the much smaller number of studies in our sample employing standardized criterion measures.

Characteristics of participants Similar correlations were found with high- and low-proficiency bilinguals, suggesting that WM is related to L2 outcomes for both less- and more-proficient adult learners. It remains to be determined by future research whether the precise role of WM varies as a function of L2 proficiency. For example, studies on the executive function of inhibitory control have been interpreted as suggesting that the reliance on inhibitory control to support bilingual lexical selection changes as L2 proficiency increases (e.g., Costa & Santesteban, 2004; Schwieter & Sunderman, 2008).

WM Language \times Complexity interaction Numerically larger effects were found with complex WM measures than with simple WM measures, regardless of the language of administration, although these differences were only marginally significant, and particularly for L2-administered WM measures (see the partially overlapping CIs). This pattern replicates the findings of the complexity covariate analysis reported above. Focusing on the complex WM tasks, effect sizes were marginally stronger for L2 than for L1 measures. Again, as we discussed above, we suggest that such effects are driven primarily by the confounding of L2 proficiency and WM abilities.

Criterion Focus \times WM Complexity interaction For L2 processing outcomes, similar effect sizes were found with simple and complex WM measures. However, for L2 proficiency outcomes, the effect sizes for complex WM measures were significantly larger than the effect sizes for simple WM measures (.27 vs. .17, respectively), as indicated by the nonoverlapping CIs. This pattern suggests that the executive control component and the STM component of WM are similarly important for understanding differences in L2 processing, whereas the executive control component may be more critical when examining L2 proficiency outcomes. However, given the relatively small number of studies examining the relationship between simple WM measures and processing outcomes ($k = 8$), this result should be considered with caution until it is replicated in future studies.

File drawer analysis

Publication bias is a major issue that must be addressed in any successful meta-analysis (Rosenthal, 1979). To mitigate the risk of overestimating the population effect size due to such bias, we took considerable effort to locate unpublished studies, including many masters theses and doctoral dissertations. Indeed, unpublished studies comprised over 20% of the studies in our sample, contributing over 36% of the analyzed effect

sizes. We also performed various analyses to assess the extent to which our effect size estimates were inflated due to publication bias. First, we examined the effect of publication status as a covariate. Although the effect size estimate for published studies was numerically larger than that for unpublished studies, the CIs overlapped almost entirely, indicating that the effect sizes did not differ significantly with publication status. Then, we computed the fail-safe N (Rosenthal, 1979) for each effect size estimate from the overall analysis and each covariate analysis (see rightmost column of Table 5). Across all analyses, the fail-safe N values were at least three times greater than the rule-of-thumb limit of $5k + 10$, providing further evidence that publication bias does not threaten the validity of these results. For example, for the primary analysis, the fail-safe N estimates that over 2,000 studies reporting a correlation of near zero would be required, in addition to the observed 79 studies, to eliminate our confidence that a true effect existed in the population. Taken together, the publication-status analysis results and the collection of fail-safe N findings indicate that the inferences drawn from this meta-analysis likely were not significantly affected by publication bias, and that WM effects are robust and positive.

Sensitivity analysis

To examine whether particular assumptions of the reported analytic methods could have impacted the results, we conducted a series of alternative analyses in which particular assumptions were relaxed or further constrained, to assess the degree of variability in the estimated population effect size. Specifically, we repeated the analyses after modifying the data set in the following ways: (1) nine effect sizes reported as being “nonsignificant” were dropped from the analysis (recall that for the primary analyses, these effect sizes were set to zero and included in the analysis); (2) outliers were included; and (3) nonsignificant effect sizes were dropped *and* outliers were included. As expected, the effect size estimates were similar or slightly higher across these additional analyses, and the inferences drawn from the analyses were identical, with one exception: Although the verbal-versus-nonverbal contrast was not significantly different in the primary analysis (with partially overlapping 95% CIs), the effect sizes were significantly larger for verbal WM measures than for nonverbal measures in two of the three supplemental analyses, as indicated by nonoverlapping 95% CIs. This pattern suggests that WM measures involving the processing of verbal content may be more strongly associated with L2 outcomes. Taken together, these results indicate that the effect size estimates from the present meta-analysis are relatively robust to the specific inclusion/exclusion criteria.

General discussion

Since Baddeley and Hitch’s (1974) seminal article, WM has become a topic of ever-increasing interest, and has reached such a level of import that WM is regarded as a central construct in cognitive psychology (Conway et al., 2005). Over the past two decades, this interest has expanded to include the study of bilingualism, with the multicomponent model (see Baddeley, 2000) inspiring the integration of WM into theoretical accounts of L2 processing and SLA. Some researchers have even gone so far as to argue that WM is L2 aptitude (e.g., Miyake & Friedman, 1998). However, others have recently questioned the growing emphasis on WM, arguing that the empirical evidence is too inconsistent to justify a central role for WM in theories of SLA (e.g., Juffs, 2005).

This meta-analysis was conducted to provide a quantitative synthesis of the effect sizes reported in studies examining WM and L2 processing and proficiency outcomes. A series of analyses revealed a robust, positive correlation between WM and L2 outcomes, with a population effect size estimated at .255. We examined a set of covariates that were identified in the review of the literature. The covariate-analysis results indicated that the executive control component of WM (measured with complex span tasks) is more strongly related to L2 outcomes than is the storage component (measured by simple span tasks), which showed attenuated but still significantly positive effect sizes. Verbal WM measures also demonstrated slightly stronger correlations with L2 outcomes, likely due to domain overlap in the task stimuli.

Implications for theories of WM

From the framework of the multicomponent model of WM (see Baddeley, 2000), the stronger contribution of complex span measures can be interpreted as indicating differential importance for the component subsystems. Specifically, the responsibilities of the executive control system (e.g., managing conflict and preventing interference from distracting information) may be more important to L2 processing and proficiency than is simply maintaining an active representation in the phonological store. Indeed, evidence is growing that bilinguals and L2 learners must manage conflict between potentially competing representations from both languages, even when only using one language in a monolingual context (i.e., lexical access is “language nonselective”; see Dijkstra, 2005, for a review), suggesting a critical role for the executive control component of WM for successful L2 use (see also Hernandez & Meschyan, 2006).

More contemporary views of WM that do not posit slave subsystems might accommodate these results by pointing to the greater need for executive control—that is, simultaneous

processing, attentional control, and coordination of multiple cognitive tasks—in the complex span tasks than in the simple span tasks. Given that bilinguals likely must engage these executive functions to support language use, the prediction is that WM tasks requiring greater executive control (i.e., complex span tasks) should be better predictors of L2 outcomes—as was borne out in the analysis. Although our data and analysis cannot adjudicate between these two competing views of WM, these results clearly indicate an important role of executive control processes for a range of L2 outcomes. More work will be needed to better specify the precise contributions of executive functions, as we discuss below.

Most current models of WM assume it to be domain-general. The stronger effect sizes found with L2 measures of WM (relative to L1 measures) could be interpreted as suggesting the need for a further fractionation of the multicomponent model into language-specific components (e.g., Alptekin & Erçetin, 2012). For example, one could argue for L1- and L2-specific phonological loops. However, this additional complexity is not warranted. The multicomponent model, as well as the other two more contemporary views of WM reviewed earlier, can easily address these content-specific differences, with the simplifying assumption that these effects are driven by the overlap in content between the measure and the criterion—not by the architecture of the WM construct itself. Similarly, the differences found between WM measures administered in the L1 versus the L2 likely reflect overlap in the content of the predictor and criterion. Moreover, empirical evidence has indicated that L1 and L2 measures of WM are highly related (e.g., Osaka & Osaka, 1992). No existing theoretical model posits separate WM components for each language, and the present results can be accommodated without positing any further fractionation of the WM construct.

Process decomposition

On the relation between individual differences in WM and executive functions, we have taken the position throughout this article that executive attention control processes of the WM system drive individual differences in WM; that is, the executive attention processes that are tapped by WM tasks are responsible for the covariation between WM and language processes. This view is consistent with Engle and Kane's (2004) executive-attention theory of the variation in WM capacity. According to this view, the predictive power of WM capacity tasks (i.e., complex span tasks) comes from the fact that they tap executive attention processes—namely, the ability to maintain access to information and goals in the face of distraction, and despite interference and attentional shifts. Engle and Kane's theory assumes that executive functions are components of the WM system, but they are

otherwise agnostic with respect to the number of executive functions and how they relate to one another, or to how executive functions relate to individual differences in WM.

With respect to the number and nature of executive functions, one of the most influential frameworks to date is that of Miyake et al. (2000), which offers data showing that three highly prominent functions—updating, shifting, and inhibition—are related but separable processes. But, beyond this framework, we know very little about the unity and diversity of WM and executive function, how these constructs correlate, and how these abilities operate in the L1 and L2 domains. In short, we do not know which executive functions are the most important for L2 comprehension and production.

Two forthcoming papers may soon shed some new light on this topic, but only incompletely so. Shipstead, Harrison, et al. (2013) used structural equation modeling to test the unity and diversity of WM and four executive functions, memory updating, attention control, prospective memory, and verbal fluency. Their central theorem was that the relationship between WM and general fluid intelligence, which has been well documented elsewhere (e.g., Conway, Cowan, Bunting, Theriault, & Minkoff, 2002), can be explained by these several individual executive functions. Surprisingly, what they found was that the executive functions most highly related with individual differences in WM (memory updating and attention control) did not mediate the relationship between WM and general fluid intelligence, but the variance common to all the executive functions did partially mediate that relationship. These findings underscored the fact that variance in WM is more than just variance in executive function.

Shipstead, Trani, et al. (2013) extended these results into the L1 domain. They used structural equation modeling to relate these same four executive functions to verbal reasoning and multiple types of reading comprehension, including ordinary comprehension and comprehension when the reader is misled (e.g., garden-path sentences). The executive function memory updating fully mediated the relationship between WM and ordinary paragraph comprehension, and the attention control and verbal fluency functions were essential for comprehending the more ambiguous garden-path material. But, again, they did not find evidence of executive function mediating the relationship between WM and verbal reasoning ability, which is consistent with Shipstead, Harrison, et al.'s (2013) findings for WM and general fluid intelligence (i.e., reasoning) ability.

These new results have implications for Engle and Kane's (2004) theory, but also for how we interpret our results here. First, we must assume that variance in WM is more than just variance in executive function, but considerably more research will be needed to specify which variance in WM is due to executive function and which is due to other aspects of WM (e.g., on the size of the focus of attention, see Cowan, 2001; on retrieval from secondary memory, see Unsworth &

Spillers, 2010). Second, we do not know which executive functions are most important for L2 proficiency. What is needed is a study of the kind reported by Shipstead, Trani, et al. (2013), but with L2 materials, including tests of L2 proficiency, L2 comprehension, and verbal reasoning in the L2.

Although data on this topic is lacking, we can speculate on the kinds of executive functions that are important for L2 outcomes. For example, in a lexical decision task, participants are presented with a letter string and must decide whether or not the stimulus is a word. This task requires cognitive processes ranging from perceptual identification of the presented stimulus (e.g., a nonword letter string) to initiating a task-relevant response (e.g., pressing a button to indicate that the stimulus is not a word). To further our understanding of how WM supports the performance of L2 tasks, researchers could employ a process decomposition approach, whereby specific cognitive subprocesses are identified at a more fine-grained level (e.g., for updating: monitoring, item deletion, and active maintenance; see Miyake & Friedman, 2012, note 1). These subprocesses could then be linked to specific linguistic processes to increase the specificity of our understanding of when and how WM contributes to L2 outcomes.

Another executive function that could be important in the language domain is the need to resolve conflict between competing representations (e.g., attention control as defined by Shipstead, Harrison, et al., 2013, or inhibition as defined by Miyake et al., 2000). This ability is also relevant to L2 processing and proficiency, given that a bilingual's two languages are both active and available in most circumstances (for reviews, see Kroll, Bobb, & Wodnieka, 2006; Kroll, Sumutka, & Schwartz, 2005). WM tasks can be manipulated to place more or less focus on this “conflict resolution” aspect of performance. The prediction would be that performance in conditions that require conflict resolution should be more highly correlated with L2 outcomes that tap into some facet of linguistic conflict resolution. Consider the *N*-back task, which requires participants to decide whether each stimulus in a sequence matches the one that appeared *n* items ago. In a low-conflict version, the list of memoranda for a given trial could be selected to have minimal repetition in nontarget locations, so that there would be little need to overcome proactive interference when making a judgment. To increase the conflict resolution demands, the task could be modified to include lures—memoranda that are repeated just prior to or following the target location (e.g., on a three-back trial, a lure would appear in Position 2 or 4). We might then predict that performance in the high conflict (lures) condition should better predict performance on L2 tasks that specifically rely on this kind of conflict resolution, such as the reading of garden-path sentences that require syntactic reinterpretation at the point of disambiguation (e.g., for evidence that *n*-back training improved L1 sentence processing, see Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, 2013).

It will be important for future work to consider how specific tasks and conditions—in both WM and linguistic tasks—call upon specific executive control processes to better elucidate the contributions of specific control mechanisms to L2 processing and proficiency development.

WM and L2 outcomes: A (bi)directional relationship?

It is important to keep in mind that the effect size analyzed in this meta-analysis was the correlation coefficient, and therefore we cannot draw any inferences regarding causality. However, on the basis of these results, it is tempting to infer a directional relationship in which greater WM resources *cause* better performance on the L2 criterion measures. Such an account would be consistent with research in other domains, in which WM has been identified as a mechanism underlying individual differences in performance across a wide range of outcomes, such as analogical reasoning and reading comprehension (Cowan, 2005; Daneman & Merikle, 1996; Engle, 2001). Some evidence suggests that systematic training of executive control processes can lead to improvements not only in performance on similar WM tasks (i.e., *near transfer*; see Harrison et al., 2013; Sprenger et al., 2013; von Bastian & Oberauer, 2013), but also on language processing tasks that place similar demands on executive control (i.e., *far transfer*; e.g., Novick et al., 2013).⁵ If WM training can lead to improvements in L2 processing tasks requiring executive control, then this could suggest a causal relationship going in the direction of WM to L2 outcomes.

Evidence from another body of research suggests the opposite direction of causality. A growing literature is demonstrating that so-called “crib bilinguals” (individuals who have spoken multiple languages from birth) show enhanced executive functions relative to monolinguals. This has been demonstrated on tasks involving conflict, with evidence coming from behavioral methods (for a review, see Bialystok, 2010) as well as from neural measures of the efficiency of cognitive control (Gold, Kim, Johnson, Kryscio, & Smith, 2013). Initially, these results were interpreted as suggesting a benefit to inhibitory control processes in particular. However, more recent research has suggested that the benefits are not limited to contexts requiring inhibition, but rather extend to task conditions that place demands on executive control functions more generally.⁶ The assumption in this research is that the

⁵ There is debate concerning the strength and stability of WM training effects, but a thorough review of that literature is outside the scope of this article. The interested reader is referred to Harrison et al. (2013), Novick et al. (2013), Redick et al. (2012), Shipstead, Redick, and Engle (2012), and Sprenger et al. (2013) for critical reviews and discussion of the available evidence.

⁶ This line of research would also benefit from the process decomposition approach described above, in order to further specify the subprocesses of executive control that are enhanced by bilingual experiences.

lifetime of experience managing multiple language systems within a single mind confers benefits to the domain-general executive control abilities of bilinguals, and that these benefits extend to other domains and tasks.

So, the directionality of the relationship between WM and L2 outcomes remains unclear. On the one hand, WM has been suggested as a (causal) mechanism underlying performance in a range of domains, including L2 processing and proficiency outcomes. But the bilingual-advantage literature suggests that the repeated, intensive performance of multilingual language tasks can impact executive functioning, and hence that the causation is in the opposite direction. The currently available data from these different literatures are unable to disentangle these possibilities. One main goal of future work could be to design experiments to clarify the direction of causality in the relationship between WM and L2 outcomes. It is entirely possible that the relationship is bidirectional, or that the directionality of the relationship depends on other factors, such as the level or time course of an analysis. These possibilities should be explored in order to identify the conditions in which WM impacts L2 outcomes, as well as the specific types and durations of L2 experience that can lead to improvements in executive control. This research will further our understanding of the complex interplay between language and cognition.

Implications for models of bilingualism

As we stated in the introduction, some have argued that the role of WM in L2 processing has been overstated (Juffs & Harrington, 2011, and Williams, 2011). To the contrary, the results of this meta-analysis suggest the need to revise existing models of bilingual comprehension and production to address individual differences in WM. For example, consider the contributions of WM to Green's (1998) inhibitory control model of bilingual speech production, which was motivated in part by Norman and Shallice's (1986) model of action. When bilinguals speak in one language, they are unable to completely "turn off" their other language (see Kroll et al., 2006), suggesting the need for control mechanisms to resolve any potential cross-language competition. According to Green's model, domain-general inhibitory control is the main mechanism for resolving this lexical competition. This control mechanism is exerted by the supervisory attentional system from outside the language system by activating schemas, which then prioritize task-relevant responses and inhibit inappropriate responses. The supervisory attentional system activates schemas on the basis of the current goals of the speaker. Individual differences in WM could be accounted for at the level of the supervisory attentional system, which is responsible for maintaining task goals and prioritizing task schemas. More efficient management of task schemas would allow

individuals with greater WM to more quickly resolve interference between competing representations. That is, the relationship between WM and L2 outcomes (particularly those involving conflict) could be driven by better top-down control at the level of the task schemas.

Following our recommendation above, to consider more fine-grained subprocesses of the WM system, we might go a step further and speculate on the different roles of various executive functions. Considering the three functions from Miyake et al.'s (2000) framework, this model has clear connections to conflict resolution ability (or the inhibition executive function), which, according to Green's model, would be the primary mechanism behind the linguistic inhibitory control, and might be represented at the level of the task schemas (which are responsible for inhibiting nontarget representations within the language system). The inhibitory control model was developed to account for a range of findings suggesting that representations in the nontarget language are suppressed in order to allow successful communication in the target language (Green, 1998; also see Kroll, Bobb, Misra, & Guo, 2008, for a recent review of evidence in favor of bilingual inhibitory processes). Moreover, some evidence has directly linked better domain-general inhibitory control abilities to reduced cross-language competition, as reflected by smaller switch costs in a language-switching task (Linck, Schwieter, & Sunderman, 2012).

Less is known about the precise roles for shifting and updating. We suggest that shifting ability might be represented at the level of the supervisory attentional system, where control is exerted over the language system by prioritizing different task schemas on the basis of the current goals of the speaker. Although little direct evidence has indicated this link, one useful data point comes from evidence suggesting that bilinguals who switch between their languages frequently during naturalistic conversations (i.e., frequent code switchers) show better performance on a domain-general, nonlinguistic task-switching task (Prior & Gollan, 2011), suggesting that shifting ability may contribute to bilingual language control by supporting shifts between task schemas. In contrast, updating may best be represented at the level of goal setting and maintenance in the face of distraction. The goal of the speaker provides top-down guidance over the language system, and the current goal must be maintained in the face of distracting information that can inappropriately activate other goals. In summary, we speculate that Green's (1998) inhibitory control model could incorporate Miyake et al.'s (2000) three related but separable executive functions, such that the current goal of the speaker (*updating*) directly informs the supervisory attentional system's functioning (*shifting*), which then translates that goal into a specific task schema that exerts control over the language system (*inhibition*). This discussion provides one possible direction that could be pursued to develop models of bilingual language processing. But what

is clear is that the construct of WM—and a more nuanced fractionation of executive functions—can and should inform these developments.

Some models of language aptitude already account for differences in WM. Indeed, Miyake and Friedman (1998) argued that WM essentially underlies the components of some models of language aptitude. For example, Skehan (1989) hypothesized that language aptitude is composed of language analytic capacity, memory ability, and phonetic coding ability—all three of which may be driven by WM and STM. Similarly, Linck et al. (2013) proposed the inclusion of both WM and STM as key components of a model of aptitude for higher-level proficiency attainment. Their study was motivated by a theoretical model of aptitude focusing on the cognitive and perceptual abilities that underlie the skills required to attain high-level foreign language proficiency. With theories of language aptitude taking a more cognitive view of SLA in recent years (e.g., Dörnyei & Skehan, 2003), WM will clearly remain a core component of successful models of language aptitude.

Strengths and limitations of the present meta-analysis

The studies included in our sample cover a range of disciplines—including psycholinguistics, cognitive psychology, and SLA—and reflect a diverse set of sources (journals, proceedings, and unpublished studies). Consequently, the inferences from this meta-analysis are not biased by undue influence from a particular theoretical perspective. As we discussed above, these results have implications for models of bilingual language processing, theories of SLA, and research on the contributions of WM to performance more broadly construed.

To better understand why the effect sizes reported in the literature are so variable, we examined the studies that reported the most extreme negative effect sizes ($r_s < -.20$), as well as studies that reported nonsignificant correlations without providing a specific correlation estimate. The sample size for many of these studies was in the range of 50–100 participants—above the median across the meta-analysis sample—suggesting that the results do not necessarily stem from low power. However, many of these studies employed global outcome measures (e.g., fluency or complexity), which may be susceptible to extra measurement error, relative to specific measures of language processing or proficiency. Alternatively, with global criterion measures, perhaps learners with less WM have more of an opportunity to employ compensatory strategies, thereby reducing the potential for WM to account for variability in these outcomes. Moreover, some studies included a participant sample that was heterogeneous with respect to education, language background, and degree of acculturation

into the local society (e.g., Andringa, Olsthoorn, van Beuningen, Schoonen, & Hulstijn, 2012). The variability in education, L1 abilities, and length of L2 exposure may have introduced additional noise into the data that attenuated any detectable relationship between WM and the outcomes.

Our survey of the literature discovered few studies of highly proficient adult learners that were relevant to the present analyses. As additional studies of WM and L2 outcomes are conducted with highly proficient learners, further synthesis of the extant results will enhance our understanding of whether and how WM's role(s) may change across the proficiency spectrum. In addition, we excluded from our meta-analysis any studies involving bilinguals who had been exposed to the L2 during childhood. Thus, it remains to be determined whether our findings would generalize to other participant populations, such as simultaneous, balanced bilinguals who have continually used both languages throughout their lives. Given that recent research has suggested that lifelong bilingualism incurs cognitive benefits including enhanced attention control (e.g., Bialystok, 2010), future studies and meta-analyses will be needed to determine whether WM's relationship with L2 outcomes differs for this population, relative to adult L2 learners.

It is also important to note that this meta-analysis focused on bivariate correlations, and therefore necessarily ignored the potential explanatory power of other relevant factors, such as general intelligence. As we mentioned previously, WM and general intelligence are correlated (e.g., Conway, Cowan, Bunting, Theriault, & Minkoff, 2002); therefore, when accounting for L2 outcomes, variance is likely to be shared between WM and general intelligence. Taking the process decomposition approach advocated above, room is certainly available to further slice up the variance in L2 outcomes and to provide incremental explanations by investigating other relevant constructs. This approach fits with the results of ongoing work on L2 aptitude, in which WM has been identified as one component of aptitude, along with other relevant cognitive abilities, including associative learning and implicit learning (e.g., Linck et al., 2013). To move the field past simply stating that WM (broadly construed) is related to L2 outcomes, it is time to focus future efforts on further specifying the subprocesses within the WM system that drive the relationships between WM and L2 outcomes, and then examining the contributions of these subprocesses and other relevant factors, like general intelligence.

Conclusions

In summary, the present meta-analysis was conducted to provide a quantitative synthesis of findings regarding the relationship between WM and a range of L2 outcomes, and to

identify moderators of this relationship. The results are congruent with claims that WM is an important component of the cognitive processes underlying bilingual language processing and performance on measures of L2 proficiency. Nonetheless, significant work still remains to be done to link specific executive functions to specific language processes, in order to advance theoretical models and further our understanding of the contributions of domain-general cognitive control mechanisms to L2 outcomes.

Author note The authors thank Matt Goldrick, Erica Michael, Jared Novick, and three anonymous reviewers for their helpful comments on a previous draft of the manuscript. Portions of these results were presented at the 53rd Annual Meeting of the Psychonomic Society and the 31st Annual Second Language Research Forum. This material is based on work supported in part by funding from the United States Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Maryland–College Park and/or any agency or entity of the United States Government.

Appendix

Effect size conversions

For studies reporting differences between two participant groups via a t test (e.g., lower vs. higher WM capacity), the t statistic was converted to a correlation coefficient by the following formula (see, e.g., DeCoster, 2009):

$$r = \sqrt{t^2 / (t^2 + n - 2)}, \quad (1)$$

where n is the sample size. For F ratios, conversion to the r metric in meta-analysis is only possible for analyses with 1 in the numerator (e.g., Schafer, 1999), using the following formula:

$$r = \sqrt{F / (F + df_{err})}, \quad (2)$$

where df_{err} is the degrees of freedom in the error term.

Prior to the analysis, all correlation coefficients were first converted to Fisher's z (Hedges & Olkin, 1985), where

$$z = .5 * \log((1 + r) / (1 - r)) \quad (3)$$

and the variance of z is computed as $1 / \sqrt{n - 3}$. The effect size estimates and confidence intervals produced by the analyses were then converted back to the original r metric:

$$r = \left(e^{(2z)} - 1 \right) / \left(e^{(2z)} + 1 \right). \quad (4)$$

Confidence intervals were estimated by first computing $1.96 * SE$ above and below the Fisher's- z effect size estimate, then converting the resulting CI values to the r metric.

Fail-safe N

Fail-safe N was computed using Orwin's (1983) method, which relies on effect size estimates on Cohen's d metric. Therefore, all inputs into the calculation were first converted from r to Cohen's d with the equation

$$d = (2 * r) / \sqrt{1 - r^2}. \quad (5)$$

Fail-safe N was then computed as

$$\text{Fail-safe } N = k \frac{d_o - d_c}{d_c - d_{fs}}, \quad (6)$$

where k = number of independent samples contributing to the effect size estimate; d_o = observed effect size (i.e., correlation coefficient), converted to Cohen's d ; d_c = value at which the validity of the findings would be called into question, converted to Cohen's d ; and d_{fs} = assumed effect size value for missing or unobserved studies, converted to Cohen's d .

References

*studies selected for inclusion in the meta-analysis

- *Abu-Rabia, S. (2001). Testing the interdependence hypothesis among native adult bilingual Russian-English students. *Journal of Psycholinguistic Research*, 30, 437–455.
- Abutalebi, J., & Green, D. (2008). Control mechanisms in bilingual language production: Neural evidence from language switching studies. *Language and Cognitive Processes*, 23, 557–582.
- *Ahmadian, M. J. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, 46, 165–175.
- *Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics*, 29, 175–193.
- *Alptekin, C., & Erçetin, G. (2009). Assessing the relationship of working memory to L2 reading: Does the nature of comprehension process and reading span task make a difference. *System*, 37, 627–639.
- *Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading*, 33, 206–219.
- *Alptekin, C., & Erçetin, G. (2012). Effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly*, 45, 235–266.
- *Andringa, S., Olsthoorn, N., van Beuningen, C., Schoonen, R., & Hulstijn, J. (2012). Determinants of success in native and non-native listening comprehension: An individual differences approach. *Language Learning*, 62(Suppl. 2), 49–78.
- Ashcraft, M. H., & Krause, J. A. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, 14, 243–248. doi:10.3758/BF03194059
- Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, 19, 537–552.
- Baddeley, A. (1986). *Working memory*. Oxford, UK: Oxford University Press, Clarendon Press.

- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423. doi:10.1016/S1364-6613(00)01538-2
- Baddeley, A. (2007). *Working memory, thought, and action*. Oxford, UK: Oxford University Press.
- Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.
- Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). New York, NY: Cambridge University Press.
- *Bergsleithner, J. M. (2007). *Working memory capacity, noticing, and L2 speech production* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis, Brasil.
- Bialystok, E. (2010). Bilingualism. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1, 559–572.
- Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, 22, 77–94.
- Carroll, J. B. (1985). Second-language abilities. In R. J. Sternberg (Ed.), *Human abilities: An information-processing approach* (pp. 83–103). New York, NY: W. H. Freeman.
- Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary learning. *Developmental Psychology*, 32, 867–873.
- Cheung, S. F., & Chan, D. K.-S. (2004). Dependent effect sizes in meta-analysis: Incorporating the degree of interdependence. *Journal of Applied Psychology*, 89, 780–791.
- *Christoffels, I. K., De Groot, A. M. B., & Waldorp, L. J. (2003). Basic skills in a complex task: A graphical model relating memory and lexical retrieval to simultaneous interpreting. *Bilingualism: Language and Cognition*, 6, 201–211.
- *Chun, D. M., & Payne, J. S. (2004). What makes students click: Working memory and look-up behavior. *System*, 32, 481–503.
- Conway, A. R. A., Cowan, N., Bunting, M. F., Theriault, D., & Minkoff, S. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30, 163–184.
- Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. N. (Eds.). (2007). *Variation in working memory*. Oxford: New York, NY.
- Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, 12, 769–786. doi:10.3758/BF03196772
- Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, 50, 491–511.
- *Coughlin, C. E., & Tremblay, A. (2013). Proficiency and working memory based explanations for nonnative speakers' sensitivity to agreement in sentence processing. *Applied Psycholinguistics*, 34, 615–646. doi:10.1017/S0142716411000890
- Cowan, N. (1995). *Attention and memory: An integrated framework (Oxford Psychology Series, No. 26)*. New York, NY: Oxford University Press.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–114. doi:10.1017/S0140525X01003922. disc. 114–185.
- Cowan, N. (2005). *Working memory capacity*. Hove, UK: Psychology Press.
- Daneman, M., & Hannon, B. (2007). What do working memory span tasks like reading span really measure? In N. Osaka, R. H. Logie, & M. D'Esposito (Eds.), *The cognitive neuroscience of working memory* (pp. 21–42). New York, NY: Oxford University Press.
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 3, 422–433. doi:10.3758/BF03214546
- DeCoster, J. (2009). *Meta-analysis notes*. Retrieved September 6, 2012, from www.stat-help.com/notes.html
- DeKeyser, R., & Koeth, J. (2011). Cognitive aptitudes for L2 learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. II, pp. 395–406). New York, NY: Routledge.
- Dijkstra, T. (2005). Bilingual visual word recognition and lexical access. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179–201). New York, NY: Oxford University Press.
- Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Oxford, UK: Blackwell.
- Engle, R. W. (2001). What is working memory capacity? In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297–314). Washington, DC: American Psychological Association.
- Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, 11, 19–23. doi:10.1111/1467-8721.00160
- Engle, R. W., Carullo, J. J., & Collins, K. W. (1991). Individual differences in working memory for comprehension and following directions. *Journal of Educational Research*, 84, 253–262.
- Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). New York, NY: Academic Press.
- Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999a). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). Cambridge, UK: Cambridge University Press.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999b). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128, 309–331. doi:10.1037/0096-3445.128.3.309
- *Erçetin, G., & Alptekin, C. (2013). The explicit/implicit knowledge distinction and working memory: Implications for second-language reading comprehension. *Applied Psycholinguistics*, 34, 727–753. doi:10.1017/s0142716411000932
- Fedorenko, E., Gibson, E., & Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language*, 56, 246–269. doi:10.1016/j.jml.2006.06.007
- *Fehrer, C., & Fry, C. (2007a). Frills, furbelows and activated memory: syntactically optional elements in the spontaneous language production of bilingual speakers. *Language Sciences*, 29, 497–511.
- *Fehrer, C., & Fry, C. (2007b). Hesitation phenomena in the language production of bilingual speakers: The role of working memory. *Folia Linguistica*, 41, 37–72.
- *Finardi, K. R. (2008). Effects of task repetition on L2 oral performance. *Trabalhos em Linguística Aplicada*, 47, 31–43.
- *Finardi, K. R., & Silveira, R. (2011). Working memory capacity in the production and acquisition of a syntactic rule in L2 speech. *Revista Brasileira de Linguística Aplicada*, 11, 199–221.
- *Finardi, K. R. and Weissheimer, J. (2008). On the Relationship between working memory capacity and L2 speech development. *Revista Signótica*, 20, 365–389.
- *Fontanini, I., & Tomitch, L. M. B. (2009). Working memory capacity and L2 university students' comprehension of linear texts and hypertexts. *International Journal of English Studies*, 9, 1–18.

- *Foote, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Applied Psycholinguistics*, 32, 187–220.
- *Fortkamp, M. B. M. (1998). Measures of working memory capacity and L2 oral fluency. *Ilha do Desterro*, 35, 201–238.
- *Fortkamp, M. B. M. (1999). Working memory capacity and aspects of L2 speech production. *Communication and Cognition*, 32, 259–296.
- Fortkamp, M. B. M., & Bergsleithner, J. M. (2007). Relationship among individual differences in working memory capacity, noticing, and L2 speech production. *Signo*, 32, 40–53.
- *Gass, S., & Lee, J. (2011). Working memory capacity, Stroop interference, and proficiency in a second language. In M. Schmid & W. Lowie (Eds.), *From structure to chaos: Twenty years of modeling bilingualism* (pp. 59–84). Amsterdam, The Netherlands: Benjamins.
- Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology*, 28, 887–898.
- *Gilbert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: The role of working memory capacity. *International Journal of English Studies*, 10, 19–42.
- Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J., & Smith, C. D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *Journal of Neuroscience*, 33, 387–396. doi:10.1523/JNEUROSCI.3837-12.2013
- Goo, J. (2010). Working memory and reactivity. *Language Learning*, 60, 712–752.
- Green, D. J. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, 1, 67–81.
- Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The CANAL-F theory and test. *Modern Language Journal*, 84, 390–405.
- *Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, 14, 25–38.
- Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science*. doi:10.1177/0956797613492984. Advance online publication.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504. doi:10.1037/1082-989X.3.4.486
- Hernandez, A. E., & Meschyan, G. (2006). Executive function is necessary to enhance lexical processing in a less proficient L2: Evidence from fMRI during picture naming. *Bilingualism: Language and Cognition*, 9, 177–188.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539–1558.
- Hummel, K. M. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics*, 30, 225–249.
- *Ikeno, O. (2006). L1 and L2 working memory: An investigation into the domain specificity and processing efficiency issues. *Bulletin of the Faculty of Education at Ehime University*, 53, 113–121.
- *Jackson, C. N., & Bobb, S. C. (2009). The processing and comprehension of wh-questions among second language speakers of German. *Applied Psycholinguistics*, 30, 603–636.
- Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society*, 102, 199–225.
- *Juffs, A. (2005). The influence of first language on the processing of wh-movement in English as a second language. *Second Language Research*, 21, 121–151.
- Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, 44, 137–166.
- Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, 133, 189–217. doi:10.1037/0096-3445.133.2.189
- *Kempe, V., Brooks, P. J., & Kharkhurin, A. (2010). Cognitive predictors of generalization of Russian grammatical gender categories. *Language Learning*, 60, 127–153.
- *Kondo, A. (2012). Phonological memory and L2 pronunciation skills. In A. Stewart & N. Sonda (Eds.), *JALT2011 Conference Proceedings*. Tokyo, Japan: JALT.
- Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, 128, 416–430.
- Kroll, J. F., Bobb, S. C., & Wodnicka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition*, 9, 119–135.
- *Kroll, J. F., Michael, E., Tokowicz, N., & Dufour, R. (2002). The development of lexical fluency in a second language. *Second Language Research*, 18, 137–171.
- Kroll, J. F., Sumutka, B. M., & Schwartz, A. I. (2005). A cognitive view of the bilingual lexicon: Reading and speaking in two languages. *International Journal of Bilingualism*, 9, 27–48.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity! *Intelligence*, 14, 389–433.
- *Lado, B. (2008). *The role of bilingualism, type of feedback, and cognitive capacity in the acquisition of non-primary languages: A computer-based study* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.
- Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning*, 57, 229–270.
- *Linck, J. A., Hoshino, N., & Kroll, J. F. (2008). Cross-language lexical processes and inhibitory control. *Mental Lexicon*, 3, 349–374.
- Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., & Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, 63, 530–566. doi:10.1111/lang.12011
- Linck, J. A., Schwieter, J. W., & Sunderman, J. (2012). Inhibitory control predicts language switching performance in trilingual speech production. *Bilingualism: Language and Cognition*, 15, 651–662.
- *Linck, J. A., & Weiss, D. J. (2011). Working memory predicts the acquisition of explicit L2 knowledge. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 101–114). Washington, DC: Georgetown University Press.
- *Londe, Z. (2008). *Working memory and English as a Second Language listening comprehension tests: A latent variable approach* (Unpublished doctoral dissertation). University of California, Los Angeles, CA.
- MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109, 35–54. doi:10.1037/0033-295X.109.1.35
- *Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, 60, 501–533.
- *Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*, 62, 704–740.

- *Majerus, S., Poncelet, M., Van der Linden, M., & Weekes, B. S. (2008). Lexical learning in bilingual adults: The relative importance of short-term memory for serial order and phonological knowledge. *Cognition*, *107*, 395–419. doi:10.1016/j.cognition.2007.10.003
- Marin-Martinez, F., & Sanchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *Spanish Journal of Psychology*, *2*, 32–38.
- *Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*, 379–413.
- *McDonald, J. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*, 381–401.
- McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 817–835. doi:10.1037/0278-7393.27.3.817
- Michael, E. B., & Gollan, T. H. (2005). Being and becoming bilingual: Individual differences and consequences for language production. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 389–408). New York, NY: Oxford University Press.
- *Miki, S. (2012). Working memory as a factor affecting L2 listening comprehension sub-skills. *Kumamoto University Departmental Bulletin Paper*, *10*, 119–128.
- *Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Erlbaum.
- Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. doi:10.1006/cogp.1999.0734
- Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. New York, NY: Cambridge University Press.
- *Mizera, G. J. (2006). *Working memory and L2 oral fluency* (Unpublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *Annals of Internal Medicine*, *151*, 264–269.
- *Mota, M. B. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos*, *24*, 69–104.
- Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation* (Vol. 4, pp. 1–18). New York, NY: Plenum Press.
- Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2013). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language and Cognitive Processes*. doi:10.1080/01690965.2012.758297. Advance online publication.
- *O’Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*, 377–402.
- *O’Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*, 557–581.
- Orwin, R. G. (1983). A fail-safe *N* for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159.
- Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society*, *30*, 287–289.
- *Payne, T. W., Kalibatseva, Z., & Jungers, M. K. (2009). Does domain experience compensate for working memory capacity in second language reading comprehension? *Learning and Individual Differences*, *19*, 119–123.
- *Payne, J. S., & Ross, B. M. (2005). Synchronous CMC, working memory, and L2 oral proficiency development. *Language Learning and Technology*, *9*, 35–54.
- *Payne, J. S., & Whitney, P. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal*, *20*, 7–32.
- Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, *19*, 291–312.
- *Poelmans, P. (2003). *Developing second-language listening comprehension: Effects of training lower-order skills versus higher-order strategy* (Unpublished doctoral dissertation). Universiteit van Amsterdam, Amsterdam, The Netherlands.
- *Posedel, J., Emery, L., Souza, B., & Fountain, C. (2012). Pitch perception, working memory, and second-language phonological production. *Psychology of Music*, *40*, 508–517.
- *Prebianca, G. V. V. (2009). *Working memory capacity, lexical access and proficiency level in L2 speech production* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
- *Prebianca, G. V. V., & D’Ely, R. (2008). EFL speaking and individual differences in working memory capacity: Grammatical complexity and weighted lexical density in the oral production of beginners. *Signótica*, *20*, 335–366.
- Prior, A., & Gollan, T. H. (2011). Good language-switchers are good task-switchers: Evidence from Spanish–English and Mandarin–English bilinguals. *Journal of the International Neuropsychological Society*, *17*, 682–691.
- R Development Core Team. (2012). *R: A language and environment for statistical computing* (ISBN 3-900051-07-0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org
- *Rai, M. K., Loschky, L. C., Harris, R. J., Peck, N. R., & Cook, L. G. (2011). Effects of stress and working memory capacity on foreign language readers’ inferential processing during comprehension. *Language Learning*, *61*, 187–218.
- *Ransdell, S., Barbier, M.-L., & Niit, T. (2006). Metacognitions about language skill and working memory among monolingual and bilingual college students: When does multilingualism matter? *International Journal of Bilingual Education and Bilingualism*, *9*, 728–741.
- Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164–171.
- *Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*, 93–132.
- Robinson, P. (1995). Attention, memory and the “noticing” hypothesis. *Language Learning*, *45*, 285–331.
- *Roehr, K., & Gánem-Gutiérrez, G. A. (2009). The status of metalinguistic knowledge in instructed adult L2 learning. *Language Awareness*, *18*, 165–181.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

- Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, 118, 183–192. doi:10.1037/0033-2909.118.2.183
- *Safranová, E., & Mora, J. C. (2012). Acoustic and phonological memory in L2 perception. In S. Martín Alegre, M. Moyer, E. Pladevall & S. Tubau (Eds.), *At a time of crisis: English and American studies in Spain* (pp. 384–390). Barcelona, Spain: Universitat Autònoma de Barcelona/AEDEAN, Departament de Filologia Anglesa i de Germanística.
- *Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua*, 120, 2022–2039.
- Schafer, W. D. (1999). An overview of meta-analysis. *Measurement and Evaluation in Counseling and Development (American Counseling Association)*, 32, 43. Retrieved from EBSCOhost.
- Schwieter, J. W., & Sunderman, G. (2008). Language switching in bilingual speech production: In search of the language-specific selection mechanism. *Mental Lexicon*, 3, 214–238.
- *Service, E., Simola, M., Metsaenheimo, O., & Mauray, S. (2002). Bilingual working memory span is affected by language skill. *European Journal of Cognitive Psychology*, 14, 383–407.
- Shipstead, Z., Harrison, T. L., Trani, A. N., Redick, T. S., Sloan, P., Bunting, M. F., ... Engle, R. W. (2013). *Working memory capacity and executive functions, Part 1: General fluid intelligence*. Manuscript submitted for publication.
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, 138, 628–654. doi:10.1037/a0027473
- Shipstead, Z., Trani, A. N., Harrison, T. L., Redick, T. S., Sloan, P., Bunting, M. F., ... Engle, R. W. (2013). *Working memory capacity and executive functions, Part 2: Language comprehension*. Manuscript submitted for publication.
- Skehan, P. (1989). *Individual differences in second-language learning*. London, UK: Edward Arnold.
- *Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter? *Psychological Science*, 17, 675–681. doi:10.1111/j.1467-9280.2006.01765.x
- *Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25, 293–321.
- Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., ... Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, 41, 638–663. doi:10.1016/j.intell.2013.07.013
- Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, 27, 625–650.
- *Taguchi, N. (2008). The effect of working memory, semantic access, and listening abilities on the comprehension of conversational implicatures in L2 English. *Pragmatic and Cognition*, 16, 517–539.
- *Tavares, M. da G. G. (2008). *Pre-task planning, working memory capacity, and L2 speech performance* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
- Tokowicz, N., Michael, E. B., & Kroll, J. F. (2004). The roles of study-abroad experience and working-memory capacity in the types of errors made during translation. *Bilingualism: Language and Cognition*, 7, 255–272.
- *Torres, A. C. G. (2003). *Capacidade de memória de trabalho e desempenho de leitores na construção de idéias principais em L1 e L2* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
- *Trofimovich, P., Ammar, A., & Gatlinton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 171–195). Oxford, UK: Oxford University Press.
- *Tsuchihira, T. (2007). L2 working memory capacity and L2 listening test scores of Japanese junior college students. *Bunkyo Gakuin Foreign Language Department of Bunkyo Gakuin Junior College*, 7, 159–175.
- *Tzou, Y.-Z. (2008). *The roles of working memory, language proficiency, and training in simultaneous interpretation performance: Evidence from Chinese–English bilinguals* (Unpublished doctoral dissertation). Texas A&M University, College Station, Texas.
- Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, 133, 1038–1066. doi:10.1037/0033-2909.133.6.1038
- Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, 62, 392–406.
- *Van Dijk, R., Christoffels, I., Postma, A., & Hermans, D. (2012). The relation between the working memory skills of sign language interpreters and the quality of their interpretations. *Bilingualism: Language and Cognition*, 15, 340–350.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. Retrieved from www.jstatsoft.org/v36/i03/
- von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, 69, 36–58.
- Watanabe, Y., & Bergsleithner, J. M. (2006). A systematic research synthesis of L2 WM measurements. In Z. Madden-Wood & K. Oeki (Eds.), *Proceedings 2006: Selected papers from the tenth college-wide conference for students in languages, linguistics, and literature* (pp. 47–60). Manoa, HI: University of Hawaii, Manoa, College of Languages, Linguistics, and Literature.
- *Weissheimer, J., & Mota, M. B. (2009). Individual differences in working memory capacity and the development of L2 speech production. *Issues in Applied Linguistics*, 17, 93–112.
- Williams, J. N. (2011). Working memory and SLA. In S. M. Gass & A. Mackey (Eds.), *The handbook of second language acquisition* (pp. 427–441). New York, NY: Routledge.
- *Winke, P. (2005). *Individual differences in adult Chinese second language acquisition: The relationships among aptitude, memory, and strategies for learning* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.
- *Xhafaj, D. C. P. (2006). *Pause distribution and working memory capacity in L2 speech production*. (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina, Florianópolis, Brazil.
- *Zaki, H. M. (2005). *Language and working memory capacity in early adulthood: Contributions from first and second language proficiency* (Unpublished doctoral dissertation). Virginia Polytechnic Institute and State University, Blacksburg, VA.