LEXICAL RETRIEVAL IN SECOND LANGUAGE LEARNERS: HOW PROFICIENCY IMPACTS FIRST LANGUAGE VERBAL FLUENCY PERFORMANCE

By

LAKE L. MATHISON

A dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Graduate Program in Spanish

Written under the direction of

Liliana Sánchez

And approved by

New Brunswick, New Jersey

October 2017

ABSTRACT OF THE DISSERTATION

Lexical retrieval in second language learners:

How proficiency impacts first language verbal fluency performance

by LAKE L. MATHISON

Dissertation Director:

Liliana Sánchez

Bilingual lexical retrieval requires the ability to access, select, and produce words from the appropriate language according to the context or task requirements. One particular measure used to test the efficiency of word retrieval is the verbal fluency task. Research comparing monolingual with bilingual verbal fluency has primarily focused on heritage bilinguals and has resulted in mixed findings (Friesen et al. 2015; Gollan et al. 2002; Luo et al. 2010; Rosselli et al. 2000; Sandoval et al. 2010). Relatively little work has been done examining the verbal fluency of second language (L2) learners (Baus et al. 2013; Linck et al. 2009; Van Assche et al. 2013), especially relative to monolinguals (Ljungberg et al. 2013).

This dissertation investigates the impact of L2 proficiency on learners' ability to retrieve words in their first language (L1). To that end, 122 English monolingual and English-Spanish L2 learner/bilingual participants completed verbal fluency tasks. Analysis of their English performance reports word totals as well as word frequency, retrieval latencies, and the time-course of retrieval. Results suggest that highly proficient L2 bilinguals have an L1 retrieval advantage in task efficiency, specifically the ability to

retrieve more words than monolinguals but with similar spread of production over time. This study is significant in its addition to our knowledge of how L2 study impacts the L1 lexical retrieval process.

ACKNOWLEDGEMENTS

I would like to take the opportunity to express my gratitude to the people who contributed to the successful completion of this project through their feedback and/or support.

First thanks go to my committee. To Liliana Sánchez for being my adviser and my friend. Her guidance, conversation, and patience have been invaluable through the whole process. To José Camacho for his thought-provoking feedback on the dissertation and generous advising during my pre-qualifying work. To Jennifer Austin for her feedback and support in the dissertation and friendly conversation throughout. A special thanks to Tamar Gollan, whose work is responsible for inspiring this project.

Additional thanks are due to my professors and peers in the Spanish department. To Tom Stephens, Nydia Flores, and Nuria Sagarra for their time and advice. A warm shout out to the Bilingualism and SLA graduate students in the cohorts preceding and following me, since I had no cohort of my own. Though I got to know some of them more than others, we're all in this together. The names are too numerous to mention, but they know who they are and how valuable their friendship has been—in encouragement and commiseration.

Finally, a heartfelt thanks to my family and friends for their support of this endeavor and for striving to understand and subsequently explain to others what exactly it is that I study. Through the ups and downs, frustrations and accomplishments, they never wavered in their belief that I would reach this goal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Introduction

1.1. Background and significance of the study

Research on bilinguals has found differences relative to monolinguals on a number of linguistic and cognitive processes, sometimes resulting in advantages—such as in cognitive flexibility and executive control—and other times manifesting in disadvantages—such as in vocabulary knowledge and lexical retrieval (Bialystok et al. 2008; Gollan et al. 2002; Hilchey & Klein 2011; Hoff et al. 2012; Sadat et al. 2012). Bilingual advantages are often touted in the news and popular media as reasons to study a second language (L2), but much of the work done in these areas has focused on early, heritage, or otherwise natural bilingual populations. It is therefore unclear whether the same effects can be expected to manifest in bilinguals with late, formal, L2 learning experiences or at what level of proficiency. The current study explores the cognitive consequences of bilingualism in the L2 learner/bilingual population by comparing participant groups including English monolinguals and English-native learners of Spanish at low, mid, and high proficiency levels. Comparisons center on performance in the first language (L1) on one particular measure of lexical retrieval, verbal fluency tasks.

Verbal fluency describes the speed with which an individual can access their stored lexicon, retrieve, and produce exemplars in a short period of time based on content or sound association. Greater speed and production are interpreted as indicative of the ease with which the mind retrieves lexical information for production while maintaining a set of constraints. Verbal fluency tasks direct participants to produce as many words as possible for a given category in a determined amount of time, most often one minute (however, see Linck et al. 2009; Roberts & Le Dorze 1997 for examples of other time limits). To be counted correct, words must conform to a specific category. The most commonly used category types for measurement are semantic and letter, the latter of which this dissertation will refer to as letter/phonemic (for reasons detailed in Chapter 3). In trials of semantic fluency, participants produce words belonging to a specific category, typically concrete nouns like Animals or Fruits. In trials of letter/phonemic fluency, participants are directed to produce words that start with a given letter/sound, most commonly the set of F, A, and S. The two types of fluency category are both seen as assessing lexical knowledge, as they require producing words from the lexicon, and executive control, as they require maintaining production over time with the constraints of producing only words that conform to the given cue and not repeating previously produced exemplars. However, given the differences between the category types, semantic fluency performance is considered to be more indicative of lexical knowledge while letter/phonemic performance relies more heavily on executive control (Shao et al. 2014).

Though developed as a neuropsychological test for monolinguals, more recently verbal fluency tasks have been used to investigate the cognitive impact of knowing two languages, by comparing the performance of monolingual and heritage bilingual populations (Bialystok et al. 2008; Friesen et al. 2015; Gollan et al. 2002; Kormi-Nouri et al. 2012; Luo et al. 2010; Rosselli et al. 2000; Sandoval et al. 2010). The findings of heritage bilingual verbal fluency have been mixed, with much of the research finding that they produce fewer words relative to monolinguals (Gollan et al. 2002; Rosselli et al. 2000) while some more recent research argues that they are capable of producing *more* words as long as linguistic factors such as vocabulary knowledge are matched with monolinguals (Bialystok et al. 2008; Friesen et al. 2015; Ljungberg et al. 2013; Luo et al. 2010).

While bilinguals in general are a relatively new avenue of interest in verbal fluency research, even fewer studies investigate formal L2 learners (Linck et al. 2009; Baus et al. 2013; Ljungberg et al. 2013; Van Assche et al. 2013). Studies that do examine L2 learner performance tend to compare internally, i.e. learners against themselves or other learners, rather than compare with monolinguals. Previous work has also not investigated the effect of L2 proficiency, typically studying one ability level or collapsing participants into one group. Among the outstanding questions that remain unaddressed, we do not yet know how or if L1 verbal fluency changes when/as a person learns an L2 and what role L2 proficiency plays in any potential changes. This study therefore advances the field of second language acquisition (SLA) by contributing new findings on bilingual lexical retrieval specific to the L2 learner population. Comparing L2 bilinguals to monolinguals allows for the examination of whether adolescent/adult formal second language study results in changes to L1 lexical retrieval relative to monolingualism; if so, including a range of L2 proficiency levels helps assess where along the proficiency spectrum that impact manifests.

L2 learners/bilinguals are a productive population to study given their different L2 acquisition process compared to heritage populations. The theories that have been proposed to account for heritage verbal fluency performance include retrieval slowing from cross-linguistic interference or accumulated frequency effects, vocabulary size, and bilingual executive control (Bialystok et al. 2008; Gollan et al. 2002; Luo et al. 2010; Sandoval et al. 2010), consequences often proposed to develop as a result of lifelong, consistent exposure to and management of two languages. L2 learners, given their early monolingualism and maintenance of a largely L1 immersive environment even after the onset of L2 study, are less likely to suffer from frequency effects in their L1 exposure and use and are likely to have monolingual-like L1 vocabulary knowledge. Therefore, this research also stands to contribute to the field of SLA by helping to distinguish between factors that are typically confounded in early heritage bilingual speakers, including age of acquisition and exposure.

1.2. Notes on terminology

Before transitioning to the review of the literature, it is pertinent to make various notes on terminology regarding bilingual experience, specifically the terms heritage bilingual and second language learner/bilingual, L1, L2, dominant and non-dominant language. Though the bilingual experience can and should be thought of as a continuum, for the sake of simplicity and expediency, bilinguals will be grouped into two rough categories, relevant both to the previous literature and the common language backgrounds represented in the United States: heritage and second language.

1.2.1. Heritage bilingual

Subsequently, unless otherwise specified, the term “heritage” bilingual (heretofore referred to as HB) refers to individuals who were exposed to two languages from an early point in life and in a natural setting. This definition combines the circumstances of simultaneous and sequential bilingualism. A primary assumption of this definition is that the family/home language (or one of the family/home languages, in the case of simultaneous bilingualism) was a minority language different from the majority community language.

1.2.2. Second language learner/bilingual

The term “second language” learner or bilingual refers to individuals whose family/home language was the majority language, and who were introduced to a second language in a formal setting, typically beginning in adolescence. The exact age of exposure can vary extensively depending on the academic environment, but the formality of L2 study in an immersive L1 environment remains relatively constant. There is no clearly defined or defensible line between the two states, though roughly one might be considered a “learner” while still formally studying the L2 and a “bilingual” after having reached sufficient competence to end formal study. However, it is easy to see how the exceptions might outweigh the examples, with students who reach admirable proficiency but continue to “study” the language or others who terminate formal study but continue to grow in their proficiency through informal exposure and use. The distinction might also differ between self-rating and third-party, with the resulting impossibility of determining which assessment is “correct.” Given that the current study examines L2 learners at the intermediate and advanced stage of university study as well as beyond, including graduate students and professionals, the compromise “L2 learner/bilingual” will be supplemented with the common abbreviation “L2er.”

1.2.3. First/second, dominant/non-dominant language

The terms L1 and L2 will be used to refer to the “first language” and “second language,” respectively. These terms will be used to denote chronological exposure to language, regardless of eventual adult dominance. While L1 and L2 are typically unambiguous in the context of L2ers, this distinction is more fraught in heritage populations, which constitute the bulk of the research presented in the literature review (e.g. Bialystok et al. 2008; Gollan et al. 2002; Luo et al. 2010). It is often the case, for example, that adult HBs will claim and/or demonstrate dominance in the majority community language and might therefore consider it their “primary” language, potentially confusing dominance for chronology, resulting in a perplexing summary of participants who by other descriptors fit the heritage label but provide mixed reports of the community language as L1 or L2 (e.g. Luo et al. 2010). Therefore, the words “dominant” and “non-dominant” will be used to distinguish such cases rather than L1 or L2.

1.3. Overview of research questions

In summary, the purpose of the current study was to examine the cognitive consequences of bilingualism in L2 learners/bilinguals, specifically regarding productive lexical retrieval. Broadly speaking, the research questions addressed in the study include the following:

1. Does formal, non-immersed, L2 learning/bilingualism impact L1 productive lexical retrieval, and if so, at what L2 proficiency level do effects manifest?

2. Does L1 productive lexical retrieval demonstrate evidence of retrieval slowing through competition or weaker links, reduced L1 vocabulary knowledge, or enhanced executive control?

More detailed research questions and hypotheses are found in the review of the literature in Chapter 2.

1.4. Organization of the study

The dissertation is organized in five chapters, as follows. Chapter 1 includes the background and purpose of the study, notes on terminology, and an overview of the research questions. Chapter 2 presents a review of the literature, including description of models of lexical retrieval and research on bilingual verbal fluency and the theories that have been proposed to account for bilingual production: retrieval slowing through competition or weaker links, reduced vocabulary, and executive control advantages. Each theoretical subsection concludes with the research question and hypotheses regarding the specific theory. Chapter 3 describes the methodology of the current study, presenting information regarding the participant groups, materials, data coding and analysis. Chapter 4 presents the results of the study, specifically the statistical analyses for each of the verbal fluency measures. Chapter 5 summarizes the findings of the study, discusses the findings in relation to the research questions and hypotheses, situates the findings in the context of the previous literature, considers the limitations of the study and directions for future research, and ends with concluding remarks.

Chapter 2: Literature Review

2.1. Introduction

This chapter is dedicated to presenting the theory and previous research that motivated the present study. The purpose of the current project is to examine the L1 productive lexical retrieval process of L2 learners, specifically within the confines of verbal fluency, to examine whether the distinct acquisition experience of formal L2 learning impacts lexical processing as has been demonstrated, with mixed results, in HB populations. The chapter will begin by presenting information relevant to the current study regarding models of lexical retrieval for production and from there move on to describing the theoretical accounts that have been proposed by various studies (Bialystok et al. 2008; Friesen et al. 2015; Gollan et al. 2002; Luo et al. 2010; Sandoval et al. 2010; Rosselli et al. 2000) for bilingual performance in verbal fluency tasks.

The section on lexical retrieval will present models of both monolingual and bilingual productive lexical retrieval (e.g. Green 1998; Levelt et al. 1999) as well as consider the goodness of fit of current models to the particular productive task of verbal fluency. Verbal fluency can be subdivided into its two most popular types, semantic and letter/phonemic, which have been shown to be differentially productive and illustrative of distinct processing strategies (Lezak 1983; Shao et al. 2014; Tombaugh et al. 1999), therefore, this section will also provide an account for the performance differences seen between semantic and letter/phonemic fluency, taking into account the retrieval process as described in the models.

The section on theoretical accounts for bilingual verbal fluency performance will begin by providing a general frame for the verbal fluency task as well as the various measures that can be studied within the performance data generated. The remaining sections will each present a theory that has been proposed to account for bilingual verbal fluency performance, based mainly on HBs and sometimes applied to the relatively fewer studies involving L2ers: retrieval slowing, including accounts based on competition and weaker links; reduced vocabulary; and executive control advantages. Each section will describe the theory and its explanation for differences between semantic and letter/phonemic fluency, including where appropriate the asymmetric performance gap between monolinguals and bilinguals in the two category types. The evidence in support of and/or at odds with each theory, from studies with both HBs and L2ers populations when available, will be detailed, and each theory subsection will conclude with a summary and the presentation of related research questions and multimeasure predictions. Finally, the conclusion of the chapter will summarize the findings across theories and describe how the current study attempts to address the limitations of prior research.

2.2. Lexical retrieval

Lexical retrieval is the process by which a person activates and selects an item from their lexicon. The present study utilizes verbal fluency tasks to examine the lexical retrieval process of L2ers in their L1. Their L1 performance is compared with monolinguals to examine whether and when increasing L2 proficiency impacts L1 processing and the specific effects of such an impact. Therefore, this research is informed by models of lexical retrieval, both for monolingual and bilingual processing.

2.2.1. Models of lexical retrieval

Lexical retrieval is generally modeled as a series of stages of activation from the presentation of stimulus to the selection of an appropriate lexical item, the specifics of which depend on whether the retrieval is for the purposes of comprehension or production. Therefore, separate theories exist to account for each (e.g. comprehension: Dijkstra & Van Heuven 1998; McClelland & Rumelhart 1981; production: Costa et al. 1999; Finkbeiner et al. 2006; Green 1998; Levelt et al. 1999). Given the current study's emphasis on lexical retrieval for production, models for recognition will not be discussed in detail. Models of recognition such as Interactive Activation (IA, McClelland & Rumelhart 1981) and its bilingual version, the BIA(+) (Dijkstra & Van Heuven 1998, 2002) trace word recognition from letter features to words, stopping short of conceptual activation. Conceptual activation is the foundational first stage in models of production, making the predictions of each difficult to generalize to the other. The following will summarize the relevant assumptions made by various well-known productive lexical retrieval models of both monolingual and bilingual processing.

Levelt et al. (1999) provide one such model of speech production. The model, like others, concerns itself only with the production of isolated words as opposed to syntactic structures, which is appropriate for the current study, as the phenomenon under investigation is restricted to the word level, not the composition of sentences or phrases. In the model, which is based entirely on excitatory responses, words are produced through a series of stages of spreading activation: first, the lexical concept, i.e. the conceptual representation, meaning, or idea; second, the lexical concept's *lemma*, i.e. relevant syntactic information; and third, the morphological and phonological code. The word's phonemes are subsequently prepared for articulation. In productive lexical retrieval, once a given concept is activated, the activation spreads outward to related concepts as well as down to the corresponding lemmas. The lemmas of non-target lexical concepts receive relatively less activation, and the extent of their competitiveness with the target lemma depends on their state of activation, i.e. more highly activated alternatives provide more competition. The target lemma is selected out of the network of activation when and because it is the highest activated.

Various fine-grained assumptions of the model can be debated, such as whether activation is feed-forward or whether later stages in the selection process can feed activation back; whether excitatory responses are complemented by inhibitory responses, in which the system can actively suppress non-target competitors; whether the lemma is a separate level (Berg & Schade 1992; Harley 1999; O'Seaghdha 1999). The purpose of the current study is not to test the validity of the minute details of the models beyond what has been described above, and the nature of the methods and results do not provide evidence to that end; therefore the current dissertation will not enter into those debates. The relevant aspects of monolingual lexical production models include the assumption of conceptual activation and the separation of the retrieval process into stages in which exemplars compete for selection.

Though the similarity in monolingual and bilingual lexical organization and storage cannot be taken for granted, theories of bilingual lexical processing have generally used models of L1 processing as a starting place (de Bot & Schreuder 1993, Green 1998, Poulisse & Bongaerts 1994 from Levelt 1992, 1993; Levelt et al.1999; Dijkstra & Van Heuven 1998 from McClelland & Rummelhart 1981), maintaining a multi-stage activation process and the assumption that lemmas compete for selection (but see Finkbeiner et al. 2006; Mahon et al. 2007). A longstanding question in the field of bilingualism is whether both languages are kept together in the mind or apart. Early debates on how a bilingual's languages interact in the brain proposed that one language had to be "turned off" in order for the other language to function (Penfield & Roberts 1959; Macnamara & Kushnir 1971), necessitating separation. Current assumptions of bilingual lexical processing have rejected that notion, arguing that both of a bilingual's languages are, at least in some respect, active at any given moment. The dual activation account has been supported through highly constrained recognition tasks involving the L1 and L2 of both HBs and L2ers, including lexical or phoneme judgment (Colomé 2001; Soares & Grosjean 1984; Ransdell & Fischler 1987) and eye-tracking of visual world paradigm (Blumenfeld & Marian 2007; Marian & Spivey 2003; Spivey & Marian 1999).

Dijkstra and Van Heuven (2002) take a step further, making a point of separating the processing from the storage of a bilingual's two languages. Procedurally, the languages can be activated selectively, i.e. one at a time, or non-selectively, in which both languages are activated, regardless of the language in use. Structurally, languages can be hypothesized to be stored independently from one another or in one integrated lexicon. Dijkstra and Van Heuven characterize the distinction between independent and integrated hypotheses as one of competition: independent lexica experience competition only from the target language, while integrated lexica can experience competition from words in either/any of the languages. The Revised Hierarchical Model (RHM; Kroll and Stewart 1994) originally posited that "[w]ords in each of a bilingual's two languages are thought to be stored in separate lexical memory systems, whereas concepts are stored in an abstract memory system common to both languages" (p.150). After further evidence, they confirm nonselective parallel activation can be incorporated into the model (Kroll & Dijkstra 2002; Kroll et al. 2010), but contend that separate lexica do not necessarily preclude interaction between cross-linguistic exemplars, but rather have "parallel access and sublexical activation that creates resonance among shared lexical features" (Kroll et al. 2010, p.374). Alternatively, Costa et al. (1999) argue that languages are activated in parallel but can be selected independently, resulting in only the items from the target language entering into competition for selection.

Though competition for selection is inherent in many bilingual lexical retrieval models (but see Finkbeiner et al. 2006), inhibition is not as widely assumed. The monolingual production model proposed by Levelt (e.g. 1989, 1999, 2001), and the bilingual models based on it (de Bot & Schreuder 1993; Poulisse & Bongaerts 1994), assert that excitatory processes alone are sufficient, while influential models such as Green's (1998) Inhibitory Control (IC) argue that the combination of processes is more efficient. Green (1998) follows Levelt et al. (1999) in assuming multiple levels of activation, including conceptual representations and lemmas containing syntactic properties, but unlike Levelt et al.'s model, the IC model asserts that in bilingual processing, the lemma is also "tagged" with a particular language, L1 or L2, and that the target language item is selected through an inhibitory process in which the task management system suppresses lemmas with non-target language tags, though not until both lemmas associated with a given concept have been activated. Inhibition is useful in part to explain counterintuitive asymmetric switch costs in productive tasks like cued picture naming. In L2 learners, cued switching into the dominant language takes longer than cued switching into the non-dominant language (e.g. Costa & Santesteban 2004; Meuter & Allport 1999, though see Gollan & Ferreira 2009; Kleinman & Gollan 2016 for how switch costs change or reduce in voluntary rather than cued switch paradigms). Inhibition can explain this result by assuming that the non-target language is suppressed during target language production; as Green (1998, p.74) describes, "overcoming prior inhibition [is] a function of the prior amount of suppression;" in other words, the relative strength of the dominant language requires proportionally more suppression be applied in order to successfully inhibit its use, and the release of the suppression takes longer than the release of the relatively less suppression required to inhibit the non-dominant language.

However, Costa and colleagues argue that the inhibitory control model does not apply universally across the bilingualism spectrum, and that highly proficient bilinguals, regardless of language pair or age of acquisition no longer demonstrate evidence of asymmetrical switch costs (Costa & Santesteban 2004; Costa et al. 2006). Symmetrical cued switch costs were found even in a relatively weaker L3 of proficient bilinguals, suggesting that proficiency imbalance is not the only factor. Costa and colleagues posit that inhibitory control dominates early L2 learning, but that at some point of proficiency, when the L2 lexical items are sufficiently robust, a language-specific selection mechanism develops that "considers for selection only the lexical representations corresponding to the intended language" (Costa et al. 2006, p.1067). Language-specific selection models maintain that a bilingual’s languages are activated in parallel (Costa et al. 1999) but counter that it is possible for the words of only the target language to compete for selection. To search for the mechanism switch proposed by Costa and colleagues, Schwieter and Sunderman (2008) cross-referenced mixed picture naming performance with L2 lexical robustness (measured with semantic verbal fluency) to explore at which point of L2 development asymmetrical switch costs indicating inhibitory control transition to the symmetrical switch costs indicative of language-specific selection. Results indicate that as lexical robustness improves, the asymmetry of switch costs shrinks until disappearing. The proposed shift from inhibitory control to language-specific selection is relevant for the current study as the L2er participants consist of groups at multiple proficiency levels.

2.2.2. Modeling verbal fluency tasks

The particular task employed in the current study is verbal fluency. However, the nature of lexical retrieval in a verbal fluency task is different enough from the example tasks described in models of lexical retrieval that a preliminary description is warranted, along with an accompanying consideration of the elements that are not accounted for in the models. There are multiple reasons why models of productive lexical retrieval are unsatisfactory for describing verbal fluency. The aim of the current study is not to revise or create a model of monolingual or bilingual lexical retrieval, but it is important to recognize the limitations of the current models in regards to describing the experimental task this study employed.

Verbal fluency tasks ask participants to name as many words as possible in a given timeframe (typically one minute), according to a given category (e.g. Animals, words starting with F) and are less constrained than many other popular recognition and production tasks. Models like Levelt et al.'s (1999) and Green's (1998) detail the production process of single isolated words, consistent with the types of task used both in example descriptions of the retrieval process and to test predictions of the model, such as lexical decision, word/picture classification or naming. Such tasks consist of a series of trials, each trial initiated with a prompt and ending with a single correct response of an isolated word (e.g., in lexical decision—yes/no; in word/picture classification—a given superordinate category; in word/picture naming—the item's name). From an experimental standpoint, and to the extent that these models describe, the trial is then complete and the next starts fresh. Verbal fluency, on the other hand, provides a broad category cue, and the response consists of a series of related words, resulting in an extensive array of correct responses. Rather than starting anew after the production of each individual word, verbal fluency trials necessarily expect previously produced words to be held in memory enough that they are not repeated (to repeat a word is considered an error).

Levelt's and others' models, conceiving of the trial as a single item, are not entirely clear on what happens to the activation of a particular item once it is articulated. It seems clear that activation may not remain at a state equivalent to that during articulation, or the same word would be repeated again and again. The response to previous exemplars likely differs based on the task and would be controlled by an external system like a language task schema (Green 1998). Evidence that even in single word trials, each trial does not necessarily start "fresh" comes from various studies on picture naming. For example, given a task like picture naming, in which there is no instruction to ignore or suppress previously produced exemplars, subsequent presentations of repeat stimuli experience facilitation from residual activation (Cave 1997; Durso & Johnson 1979; Mitchell & Brown 1988). Conversely, picture-word interference tasks instruct participants to ignore distractor stimuli, and the presentation of a target word that was previously presented as a distractor results in interference and retrieval slowing relative to control, suggesting that rejected words are in fact inhibited (Fox 1996; Neuman et al. 1999).

Whether a previously sampled verbal fluency exemplar acts in the same way as a primed repetition or inhibited distractor is unknown. The task of verbal fluency, producing a word followed by related words without repetition, is more analogous to the latter, though the order of events is opposite. In the interference task described above, a word must be actively ignored and then later produced, while in verbal fluency, a word is produced and then must be rejected from further reproductions. In a system assuming only excitatory responses, the activation likely declines, back to neutral or at least to below the activation level of other lemmas, to free up other exemplars for production. In a model that includes inhibition, the word may be inhibited to below neutral to prevent it from being produced again. When semantic verbal fluency trials are repeated (Roberts & Le Dorze 1997), the number of responses often increases, with some but not complete overlap in content, suggesting limited residual activation. Importantly, however, the results reflect the impact *between* trials, not within one trial.

Another area of mismatch between models of lexical retrieval and the verbal fluency task results from the two different types of category: semantic and letter/phonemic. The above described productive models have a somewhat easier job accounting for semantic fluency, given that it should follow the same staged process from activation of lexical concepts through articulation. The semantic cue activates a superordinate category (e.g. Animals), requiring understanding of the meaning of the category to which the cue refers as well as sufficient understanding of each of the selected exemplars to verify that they are category members. Letter/phonemic fluency, meanwhile, presents an interesting challenge to models of lexical retrieval. Letter/phonemic cues (e.g. words starting with F) are based not on meaning, the supposed first step in productive retrieval, but orthographic/phonological form, the supposed last step before articulation. In fact, meaning has no bearing on whether a given word is appropriate for a letter/phonemic trial, and participants could hypothetically produce a word they had heard before but for which they had developed no sense of meaning or conceptual representation. Therefore, letter/phonemic fluency could be argued to lack conceptual mediation entirely, or at least access it far less than its semantic counterpart.

Levelt et al. (1999) assert that the "intentional production of a meaningful word always involves the activation of its lexical concept," (p.3). Though they argue that nearly all words are meaningful, they allow for tasks to differ in intent, with "intentional" taken to mean that "the word's meaning has *relevance* for the speech act [which] is often not the case in recitation, song, reading aloud, and so on" (p.37, emphasis in original). Various studies (e.g. Kroll & Stuart 1994) indicate that word naming, i.e. reading a cued word aloud, is a primarily lexical activity, capable of bypassing conceptual mediation.

Similarly, letter/phonemic fluency, in which a word's meaning is not relevant for the speech act, may lack intent and therefore not activate its lexical concept. This characteristic separates letter/phonemic fluency from the type of production being accounted for by such models. However, word naming is prompted production, with each trial again consisting of one word, while letter/phonemic verbal fluency is relatively free production (within the confines of the cue, at least) with multiple words being produced per trial. Therefore, while in a certain sense word naming and letter/phonemic fluency are analogous, the comparison is not entirely adequate.

Given letter/phonemic fluency's atypical retrieval process, it is unsurprising that the models previously presented do not specifically account for retrieval of that particular type. It is therefore instructive to consider other models, such as the mathematical random-search model (McGill 1963), which accounts for production of multiple words per trial and successfully predicts exponentially declining rates of retrieval as seen in verbal fluency trials (e.g. Crowe 1998). Applied to verbal fluency (as described by Rohrer et al. 1995):

> [T]he retrieval cue (e.g., *farm animals*) delimits a mental search set that contains the relevant items (e.g., *lamb*, *sheep, cow,* and so forth). Exemplars are randomly sampled one at a time, at a constant rate. Each item has the same probability of being sampled, and this probability holds constant throughout the recall period. Each sampled exemplar is immediately recognized as either a not-yet-sampled exemplar (and then retrieved into consciousness) or a previously sampled exemplar (and then ignored). As the number of not-yet-sampled items decreases,

> the number of items retrieved in each, say, 5-s bin correspondingly declines throughout the recall period. (p.1129, italics in original)

The original random-search model, while valuable, is simplistic, ignoring multiple key elements of retrieval from the lexicon. It does not account for the fact that retrieval of "relevant" items from the search set may also activate irrelevant search items. Its assumption that "each item has the same probability of being sampled" ignores that frequency, salience, and prior activation all affect the strength of the connection and therefore a given item's chance of being selected (subsequent models have attempted to account for variable accessibility, e.g. Vorberg & Ulrich 1987). Also, as admitted in Rohrer & Wixted (1994), the random-sample model does not account for clustering—defined as the production of multiple words in quick succession—which implies that not all words are sampled independently from each other.

Bilingual lexical retrieval is further complicated under the random-search model, because it is unclear how/whether the non-target language would be represented in the search set. Retrieval according to the model "depends on both the breadth of search and speed of processing," such that "the average time needed to retrieve the items within the search set increases when either the size of the search set increases or the duration of each random sampling increases" (Rohrer et al. 1995, p.1130). Determining how bilingualism impacts search set size depends on what is considered to be the object of retrieval from a staged retrieval model. If the object of retrieval is conceptual representations, the search set of concepts would be relatively similar regardless of bilingual status (perhaps slightly smaller, not having the amount of exposure to low-frequency vocabulary as a monolingual). However, it is unlikely to be the concept, given that as discussed above, conceptual access is only guaranteed in semantic fluency, which would make the object of retrieval different for the two types of fluency category.

If the object of retrieval is the lemma or word form, the size of the search set of a bilingual would depend on assumptions regarding whether exemplars from both languages compete for selection. If selection is language specific and it is possible for only the words of the target language to compete (as is assumed by the predictions of Luo et al. 2010 and Sandoval et al. 2010, discussed below), the bilingual search set would likely be smaller, given that bilinguals often demonstrate reduced vocabulary knowledge in each individual language relative to a monolingual (e.g. Bialystok & Feng 2009; Bialystok et al. 2008; Portocarrero et al. 2007). Retrieval would involve activating a superordinate concept and preemptively eliminating non-target language exemplars from the selection process.

Alternatively, given nonselective activation and cross-linguistic competition, the bilingual search set could be argued to be substantially larger than a monolingual (while it is true that heritage bilinguals tend to have smaller vocabularies than a monolingual in each individual language, they are not smaller by half or more). The process of retrieval would involve sampling words and categorizing each as it is encountered as “previously sampled,” “not-yet-sampled-and-appropriate” or perhaps a third label like “not-yet-sampled-but-inappropriate” for all non-target language words. Another option, not considered by the original random-search model given its assumption that all items are equally likely to be selected (though variable item strength is considered in some follow-up models, e.g. Vorberg & Ulrich 1987), would involve prioritizing the activation of the target language lemmas, identified by their language node or tag (Green 1998) in order to restrict production to only target language word forms, e.g. by increasing the base activation of the target language lemmas or by inhibiting the non-target language lemma. Comparisons of language testing order in bilingual verbal fluency do suggest local and possibly global inhibition of active L1 lemmas, specifically when the L2 is tested first (Van Assche et al. 2013), though the effect was tested only with letter/phonemic cues and therefore would reflect inhibition of phonologically related lemmas rather than semantically related lemmas or translation equivalents.

The duration of sampling between eligible exemplars could therefore be argued to increase due either to encountering a larger number of items in the search set that need to be rejected/ignored or to the increased effort of inhibiting non-target language tags. This would be especially and maximally the case for balanced HBs, for whom the assumption that items have the same probability of being sampled would come the closest to accurate. The variable strength of L2 items in L2ers would lower the probability of non-target language items being selected in L1 trials or require lesser inhibition, lessening the impact (the reverse would be the case for L2 trials, with stronger L1 items increasing the probability of non-target item competition and requiring greater inhibition).

2.2.3. Semantic and letter/phonemic performance differences

The focus of the current study is not lexical storage but retrieval, more specifically retrieval for production; therefore this review of the literature focuses primarily on such models. However, in considering the performance differences between semantic and letter/phonemic fluency, it is instructive to consider how words are theorized to be *stored* in the mind. Discussions of verbal fluency have attributed production advantages in semantic fluency to its coincidence with how words are stored in memory, according to meaning and not form (Lezak 1983; Martin et al. 1994). However, this explanation likely mischaracterizes or at least oversimplifies assumptions about lexical storage.

Words in the lexicon have been theorized to be stored semantically as well as phonologically, in part because of the differing processes of language recognition and language production. Though the idea of having two separate mental "dictionaries" would appear restricted to bilinguals, i.e. one dictionary for each language, in fact dual lexical listings were also contemplated to exist in the monolingual mind (Fay & Cutler 1977) on the basis of the differing optimal retrieval strategies implied in reception versus production. Language recognition or comprehension requires a transformation of form to meaning. Meanwhile, language production requires transforming meaning into form. Maximal efficiency suggests that words should be organized by form, i.e. phonologically, by syllable structure, etc., for receptive purposes and by meaning, i.e. semantically, for productive purposes (Fay & Cutler 1977).

Speech data does in fact provide evidence of semantic relationships between words during production. Some speech errors confirm that words selected for production are connected by meaning. For instance, blends combine parts of the phonological form of two words with similar meaning, e.g. mixing "slightest" and "least" into "sleast" (Aitchison 2008, p.244) or the pop-culture reference from the film *Mean Girls* (2004): "grool" as a blend of "great" and "cool." Semantic substitutions can also occur, such as substituting "early" for "late," "good" for "bad," "yesterday" for "tomorrow," etc. Psycholinguistic research on the "tip of the tongue" (TOT) phenomenon (Brown &

McNeill 1966) indicates that subjects who cannot access the form of a target word can often provide semantic clues, such as synonyms.

However, other speech errors undermine the idea that the lexicon is dually organized, by form for comprehension and meaning for production, by indicating that at the moment of selection for production, words are also connected by form, e.g. sound, syllable structure, stress patterns. Participants experiencing TOT states, in addition to the semantic clues mentioned above, can also often provide form clues, such as initial letter or number of syllables. Additionally, as opposed to semantic substitutions, malapropisms involve substituting a word of similar sound and structure, but unrelated in meaning, e.g. "What are you *incinerating*?" rather than "What are you *insinuating*?" (Aitchison 2008, p.242, emphasis in original). In such cases, the error and the target often share an initial phoneme, as well as grammatical category (99%), syllable structure (87%), and/or stress pattern (98%) (Fay & Cutler 1977, p.508).

Though dual "dictionaries" for comprehension and production have been rejected (as have the separate language dictionaries for bilinguals), it is still the case that comprehension and production involve different processing orders, which can help to explain why letter/phonemic verbal fluency is classically more difficult than semantic fluency. Since words in the lexicon are organized with connections based on form and meaning, the issue is not so straightforward as to say that letter/phonemic verbal fluency goes against the way the human lexicon is organized, but rather that the specific act of production favors meaning-based connections, which is characteristic of semantic, but not letter/phonemic, fluency (Gollan et al. 2002). And while letter/phonemic fluency might not be against how words are stored in memory, it certainly goes against how we typically arrange thoughts for production. Humans occasionally need to vocalize lists according to semantic value (e.g. grocery lists, pointing out animals we see at the zoo), but we do not ever (unless composing alliterative poetry) need to plan utterances according to initial phoneme.

Neuropsychological literature on verbal fluency recognizes that letter/phonemic and semantic verbal fluency both recruit lexical knowledge and executive control processes, but that the emphasis differs based on different retrieval mechanisms, e.g. semantic fluency highlights vocabulary knowledge while letter/phonemic highlights executive control (Shao et al. 2014). Semantic fluency relies on the conceptual network organization of the lexicon while letter/phonemic fluency relies on strategic search and switch processes (Martin et al. 1994). Because the letter/phonemic fluency retrieval process works contrary to typical productive lexical retrieval, the brain must rely on different processing strategies, resulting in differential performance among the two category types.

2.3. Theoretical accounts for bilingual productive lexical retrieval

The following sections will explore the specific theories proposed to account for bilingual lexical retrieval in productive tasks, with an emphasis on verbal fluency, but referencing other tasks when appropriate. The theories include retrieval slowing through competition, dual-tasking, and weaker links, as well as reduced vocabulary and enhanced executive control.

2.3.1. Verbal fluency additional measures

The results of verbal fluency research have generally reported word totals, or the total number of correct words produced during the (generally one minute) timed period. A small selection of studies has investigated beyond word totals, including comparisons of mean word frequency and retrieval latencies. Word frequency is relatively self-explanatory, referring to the occurrence per million words according to corpora. Higher word frequency indicates words that are used relatively more often than words of lower frequency. Retrieval latencies involve examining the time at which a word was spoken relative to the trial and can be broken into two complementary measures (Rohrer et al. 1995, explained in more detail in the Methodology chapter). The first-response latency (FRL) measures the amount of time between the beginning of the trial and the enunciation of the first word, and represents the initiation of the retrieval process. The mean subsequent-response latency (SRL) averages the amount of time between the first word and each subsequent word, and can be used to distinguish between participants whose words are concentrated at the beginning and taper off from those whose words are more regularly spaced. Finally, the retrieval latency data can also be used to craft a time-course of retrieval, in which the minute-long trial is broken into 5-second segments. The number of words produced in each 5-second bin is then graphed to create a visual representation of the decline curve in production over time.

2.3.2. Retrieval slowing

Retrieval slowing serves as an umbrella term for multiple theories that account for production by assuming bilinguals are slower to access, select, and produce words during the retrieval process. The most common of the theories is competition, which claims the non-target language interferes with production because it supplies additional competitors for the target item. Related proposals referred to by Gollan et al. (2002) and Sandoval et al. (2010) as dual restriction and dual task analogies, respectively, compare the interference from the non-target language with parallels to monolingual processing under belabored conditions. Finally, "weaker links" is an alternative account independent of special bilingual processing, reliant instead on assumptions of frequency counts.

2.3.2.1. Competition

It is natural to assume competition, and potentially inhibition, would be a factor in verbal fluency production, and much of the disadvantaged HB production in verbal fluency tasks relative to monolinguals is attributed (e.g. Rosselli et al. 2000; Ivanova & Costa 2008; Sandoval et al. 2010) to the increased cognitive processing requirements of knowing two languages. As described above, research has found that bilinguals cannot simply "turn off" one of their languages, and the non-target language remains active to an extent even when not in use (e.g. Blumenfeld & Marian 2007; Dijkstra & van Heuven 2002).

A competition account of bilingual verbal fluency necessarily assumes nonselective language access and likely an integrated lexicon (though Kroll et al. 2010 argue that nonselective access does not necessitate integration). In order for the non-target language in general or the translation equivalent to compete for selection and thus slow retrieval of the intended exemplar, the non-target language must be available for activation. If it is thus assumed that a bilingual's two languages are active and interdependent, activation of a given concept spreads activation to its semantic relatives, which in turn spread activation to their corresponding lemmas in the target language and also to their translation equivalents. The resulting diffusion of activation creates multiple competitors to the intended target, slowing selection of the winning lemma.

Competition accounts do not directly address the relative difficulty of semantic and letter/phonemic fluency in general, though to account for the asymmetrical bilingual performance gaps, it has been posited that the concrete nouns solicited by semantic fluency "may share more elements of their representations across languages," (Rosselli et al. 2000, p.23) resulting in greater interference. They conclude that semantic and letter/phonemic verbal fluency recruit different processing, the former more based in lexical knowledge, the latter reliant more on executive function, making semantic fluency more prone to interference from cross-linguistic competitors.

The assumption that concrete words share larger portions of their conceptual representations than nonconcrete words is based in part on the findings of Van Hell and De Groot's (1998) word association tests. They found that retrieving an associate of concrete words and nouns was easier than for abstract words and verbs, both within a language and across languages. Semantic fluency cues typically (though not always, see e.g. Portocarrero et al. 2007; Sandoval et al. 2010) involve superordinate categories that solicit concrete nouns, e.g. Animals, Fruits, Vehicles, etc., while letter/phonemic cues accept words of both open and closed class, with no regard to abstractness. Therefore, each word produced in semantic fluency is likely competing with its translation equivalent while a larger proportion of the words produced in letter/phonemic fluency are of the type (nonconcrete, non-noun) with weaker conceptual connections to their translation.

An alternative or addition to the hypothesis that concrete words share more features across languages relative to abstract words relates to the difference (discussed above) in task between semantic and letter/phonemic fluency, rather than in the words solicited. Finkbeiner et al. (2006) argue for shifting focus from lexical access to response selection, in which mismatches between responses and response selection criteria are used to reject items. In either semantic or letter/phonemic verbal fluency, translation equivalents will mismatch with the implicit target language restriction. However, during semantic fluency, cross-linguistic translation equivalents cannot be as quickly rejected as potential items, because they fit the explicit restriction of category, while during letter/phonemic fluency, most words produced in a trial will not have translation equivalents that fit the category restriction. The double mismatch of response to criteria therefore allows cross-linguistic letter/phonemic competitors to be rejected more quickly, resulting in reduced or null performance impact relative to monolinguals.

Multi-measure predictions based on the competition account predict bilinguals will produce fewer words than monolinguals. Beyond word totals, Sandoval et al. (2010) argue that competition would result in lower average word frequency and longer retrieval latencies. High-frequency words are by definition more used and therefore more likely to be active or likely to be more active than low-frequency words. Since retrieval is relative to the amount of activation of competing items, high-frequency words are more likely to compete more strongly, leading bilinguals to produce words of lower average frequency. Retrieval slowing would also delay the initiation of production, resulting in a longer FRL, as well as increase the time between words, resulting in longer mean SRLs and a time-course graph with lower initiation (fewer words produced at the beginning of the trial) and more gradual decline curve (Friesen et al. 2015; Luo et al. 2010).

2.3.2.1.1. Heritage bilingual findings attributed to competition

Given the acceptance of dual language activation, various studies finding reduced bilingual productivity in verbal fluency tasks have proposed cross-linguistic competition as the source of the disadvantage (e.g. Rosselli et al. 2000; Sandoval et al. 2010). Rosselli et al. (2000) compared elderly monolinguals (English and Spanish) and bilinguals (Spanish-English) in both semantic and letter/phonemic categories. The bilingual participants produced fewer total words in semantic fluency than the monolinguals of either language, but performed equivalent to monolinguals of either language in letter categories. The authors propose that language interference was a factor in performance, citing the representation overlap described above to account for the discrepancy between category types.

Portocarrero et al. (2007), like Rosselli et al. (2000), found disadvantaged semantic fluency in their comparison of English monolinguals and self-professed balanced bilinguals of a variety of L1s. To test Rosselli et al.'s postulation, this study included Actions (i.e. things people do, a.k.a. verbs) as a semantic cue to assess whether concrete nouns (e.g. Animals, Kitchen items) were more disadvantaged than nonconcrete words. One of few studies to disaggregate results within a category, they found that bilinguals produced fewer words than monolinguals in overall semantic fluency as well as within each individual cue. They also found an asymmetric performance gap between monolinguals and bilinguals when comparing Animals and Actions, in that the difference between groups was larger for the concrete Animals than nonconcrete Actions, in line with predictions based on Rosselli et al.'s proposed explanation. However, the same did not hold true with the Kitchen cue, leaving the hypothesis only partially supported. While the authors conclude that language interference contributed to verbal fluency performance, important to the interpretation of these results is that the bilinguals scored lower than the monolinguals on measures of receptive and expressive vocabulary, leaving open the interpretation (advanced by Bialystok et al. 2008), mentioned only in passing, that results may reflect instead a disguised deficit in vocabulary knowledge, an account that is presented below.

2.3.2.1.2. L2 learner findings attributed to competition

One of few verbal fluency studies to include L2ers rather than HBs, Linck et al. (2009) examined the impact of L2 immersion, comparing English-Spanish intermediate L2 classroom learners with students studying abroad. Participants completed semantic fluency trials in both the L1 and L2, though unlike other studies in which language order was counterbalanced, the classroom group performed the tasks in the L1 first, while the immersion group performed in the L2 first. The relevant finding to the current study is that the immersed L2ers produced fewer L1 words than their classroom counterparts. The authors surmise that immersion increased the amount of interference from the L2, requiring increased L1 inhibition and resulting in diminished L1 lexical access. Interestingly, though, the effect was only temporary and had disappeared by six months reintegration into an English-speaking environment.

Though not currently published, Linck et al.'s (2009) data were later coded and analyzed for retrieval latencies (Gerfen, Tam, McClain, Linck & Kroll, in preparation; as cited in Kroll et al. 2012 and Misra et al. 2012). Immersed learners reportedly had longer FRLs and longer inter-response latencies than classroom learners, which combined with fewer total words, supports accounts of retrieval slowing due to immersion. The six-month delayed posttest does not appear to have been subject to similar coding and analyses, so it is not possible to confirm whether the latencies rebounded similar to the word totals after reintroduction to the dominant L1 environment.

Though arguably the most relevant study to the current dissertation, Linck et al. (2009) present multiple limitations that the current study attempts to address by: comparing L2 learners to monolinguals, including L2er participants of multiple proficiency levels, examining both semantic and letter/phonemic fluency performance in full 1-minute trials as opposed to abbreviated 30 s trials, counterbalancing the order of language presentation, and analyzing production data beyond word totals (apart from the unpublished retrieval latency results of Gerfen et al. in preparation) by including word frequency and time-course analyses. Finally, Linck et al.'s study focuses on the impact of immersion, while the current study seeks to examine how L2 proficiency impacts L1 lexical retrieval in a non-immersed environment.

Van Assche et al. (2013) examine verbal fluency performance in L2 Dutch-English bilinguals. This study differed from other L2er verbal fluency studies (Baus et al. 2013; Linck et al. 2009) in multiple important ways: bilinguals were tested in an L1 immersive environment rather than during L2 study abroad; participants were tested in letter/phonemic rather than semantic categories; language testing order was counterbalanced; and importantly, given the aim of that study to explore local vs. global inhibition, trial cues only partially overlapped between languages, in that participants received some cues twice (once in each language) and other cues once. In line with competition and particularly inhibition accounts, participants who were tested in the non-dominant L2 (English) first produced fewer words on the repeated cues, but not the novel cues, suggesting local inhibition of previously activated non-target language lemmas. In support of inhibitory control models, the effect was asymmetrical, in that it was not replicated by participants tested in the dominant L1 (Dutch) first, similar to asymmetric switch costs found in numeral and picture naming (Costa & Santesteban 2004; Costa et al. 2006; Meuter & Allport 1999).

Similar to the prior L2er studies, Van Assche et al.'s (2013) results do not address how L2ers compare to monolinguals and do not examine the impact of L2 proficiency, as like in Baus et al. (2013), participants were not subdivided according to L2 knowledge, though scores on a multiple choice vocabulary test were reported.

2.3.2.1.3. Findings at odds with competition

Luo et al. (2010) discount retrieval slowing in favor of executive control advantages (discussed below) based on their predictions regarding the combined interpretation of word totals and retrieval latencies. They argue that the finding that vocabulary-matched, or "high vocabulary" (HV), HBs produced more words than monolinguals conflicts with retrieval slowing accounts, because "[i]f HV bilinguals accessed lexical items more slowly than monolinguals, then they would generate fewer words than monolinguals in the same amount of time" (p.39). A time-course could theoretically exist in which retrieval were somewhat slowed but, offset by the greater consistency of enhanced task control, resulted in a more gradual decline and more total words; however, Luo et al.'s time-course analysis did not align with that possibility. Both the FRL and initiation of retrieval in the time-course, where retrieval slowing would be the easiest to detect, were equivalent between vocabulary-matched HBs and monolinguals, indicating that the former were retrieving words at a similar rate to the latter.

2.3.2.1.4. Evidence of cross-linguistic inhibition

Ancillary to the general theory of competition, there is the question of cross-linguistic inhibition, which can be examined through analysis of language order effects. Few of the bilingual verbal fluency studies have examined the impact of language testing order. This is due in part because many studies test participants in only one language (Gollan et al. 2002, though they also included dual-language trials in which exemplars from either language could be used; González et al. 2005; Sandoval et al. 2010 Experiment 1), often because the bilinguals represent a mixed collection of L1s (Bialystok et al. 2008; Kormi-Nouri et al. 2012; Luo et al. 2010; Friesen et al. 2015). Those studies that do examine production in both languages (Baus et al. 2013; de Picciotto & Friedland 2001; Linck et al. 2009; Roberts & Le Dorze 1997; Rosselli et al. 2000; Sandoval et al. 2010 Experiment 2; Snodgrass & Tsivkin 1995; Van Assche et al. 2013) do not always counterbalance the testing order (Linck et al. 2009), or if order was counterbalanced, analyses did not compare participants according to language order

(Baus et al. 2013; de Picciotto & Friedland 2001; Roberts & Le Dorze 1997; Rosselli et al. 2000; Snodgrass & Tsivkin 1995).

Of the remaining candidates, Sandoval et al. (2010, Experiment 2) found a numerical trend, in which HBs produced slightly fewer words in the dominant L2 (English) when tested in the non-dominant L1 (Spanish) first, but the difference was not significant, and neither were comparisons between word frequency and retrieval latencies. Van Assche et al. (2013), on the other hand, did find significant language testing order effects, in line with the asymmetrical switch costs found for lower proficiency bilinguals in other tasks (Meuter & Allport 1994; Costa & Santesteban 2004; Costa et al. 2006). In one experiment, HBs produced fewer words in the dominant L2 (English) when tested in the non-dominant L1 (Chinese) first, both in cues that were repeated in both languages as well as cues that were not. In the other experiment, L2ers produced fewer words in the dominant L1 (Dutch) if tested first in the non-dominant L2 (English), though for this population, the effect was only present in cues that were repeated between languages.

2.3.2.1.5. Dual-restriction and dual-task analogies

Dual-restriction and dual-task proposals are related to competition in that they assume cross-linguistic interference, but emphasize the inhibitory task itself, instead of the translation equivalent, as the source of cognitive burden. Both accounts derive from processes shown to impact verbal fluency in monolingual studies. Gollan et al. (2002) compare bilingual processing to that of a monolingual retrieving words with two levels of restriction, e.g. a semantic category with an additional letter restriction (names that start with M, Azuma et al. 1997), which results in reduced production. Gollan et al., in turn, argue that bilingualism itself may act as a permanent dual restriction, as the bilingual brain maintains the task of assuring that the production is in the appropriate language. Thus, a monolingual needs only ensure that the exemplars pertain to the given category, while a bilingual given the same task must ensure that words pertain to the given category *as well as* come from the target language. A similar account (Sandoval et al. 2010) compares bilingual lexical retrieval to that of a monolingual retrieving words while performing a secondary task (e.g. producing exemplars while keeping track of sequences of dots on a computer screen; Rohrer et al. 1995). The additional task decreases overall production by slowing retrieval.

As discussed above, according to the random-sampling model (as described in Rohrer et al. 1995), mean retrieval latency can be slowed by increasing the amount of time between each random sampling. Bilingualism could increase the duration of sampling if, as interpreted by Sandoval et al. (2010), the need to monitor the output language serves as a distraction to the task-management system, slowing the incidence of each sample.

2.3.2.1.5.1. Dual-restriction and dual-task analogy evidence

To investigate whether bilingualism acts as a second restriction, Gollan et al. (2002) compared monolingual and heritage bilingual performance in semantic and letter/phonemic categories. While bilingual participants did produce fewer words than monolinguals in both semantic and letter/phonemic fluency, the disadvantage was disproportionate to semantic fluency, which the authors argued contradicted dual-restriction predictions, since language as a secondary restriction should impact both fluency types equally.

In an attempt to add a secondary restriction to monolingual production that did not simultaneously add an additional restriction to bilingual production, participants also completed name+letter trials (e.g. names starting with M). Contrary to the predictions, the name+letter trial did not even out the difficulty between the two populations. Monolinguals produced fewer words than in semantic and letter/phonemic trials, but the heritage bilinguals did as well. This result may be explained by the fact that certain names tend to belong to one culture more than another (e.g. English: Michael, Mitch, Mandy; Spanish: Maricela, Miguel, Moises), or differ phonologically based on the language (e.g. Mary – María); in this way, language can still play an important role in the names generated, maintaining the restriction, in this case a *third* for the bilinguals. Given the asymmetric impact of bilingualism on verbal fluency, the authors reject the dual-restriction account of cross-language interference in favor of the more traditional competition between translation equivalents.

The secondary task analogy was investigated by Sandoval et al. (2010), to examine whether the act of inhibiting the non-target language works as a distractor task, likewise slowing retrieval and negatively impacting production. The heritage bilinguals in the study produced fewer words than monolinguals. Unlike Gollan et al. (2002), both semantic and letter/phonemic fluency were impacted equally. Importantly, this study was the first to examine bilingual verbal fluency beyond word totals, examining both word frequency and retrieval latency. Bilinguals produced words of lower average frequency, in line with their predictions from a competition perspective, as explained above.

Importantly, bilinguals' FRLs were longer, though in semantic fluency only, and their mean SRLs were longer for both semantic and letter/phonemic fluency. The authors therefore conclude that the evidence supports a competition account, with bilingual production analogous to monolingual production with a secondary task, though they noted that the impact on the bilinguals of the dual language monitoring was much smaller than for the monolinguals completing the counting/key tapping task in Rohrer et al. (1995). However, this discrepancy is easily accounted for by the artificial and hence more distracting nature of the secondary task utilized by Rohrer et al., as opposed to the highly practiced and natural state of language selection and inhibition in the bilingual brain.

2.3.2.1.6. Competition and inhibition summary

In summary, a bilingual's two languages are always active to a certain extent, and one language cannot be shut off from activation. One result of this dual language management is a disadvantage in lexical processing relative to monolinguals, found in vocabulary size and lexical retrieval in a variety of productive tasks like picture naming. Initial verbal fluency research finding disadvantaged production likewise turned to competition as an explanation. In subsequent and more in-depth studies, the results of bilingual verbal fluency research have provided mixed results concerning the impact of cross-linguistic interference on verbal fluency performance relative to monolinguals. Crucial to those analyses are measures of verbal fluency beyond word totals, including mean word frequency and retrieval latencies. The more recent research explicitly testing the predictions of competition against other alternatives, discussed below, has alternately favored (Sandoval et al. 2010) or conflicted (Luo et al. 2010) with competition as an explanatory factor.

Synthesizing the three experiments comparing bilingual verbal fluency by language testing order, there is mixed evidence of cross-linguistic inhibition and asymmetric switch costs in verbal fluency performance. The lack of testing order effect in HBs could be attributed to the development of a language-specific selection mechanism for proficient bilinguals, as proposed by Costa and colleagues (Costa & Santesteban 2004; Costa et al. 2006). However, one of two HB comparisons found language order effects, both global and local. One difference between the HB experiments is the language pair. Though Costa et al. (2006) argue that language similarity does not impact the development of the language-specific selection mechanism as long as participants are highly proficient bilinguals, the language pair forming the basis of that claim was Spanish-Basque, while the pair analyzed by Van Assche et al. (2013) is Chinese-English, leaving open the possibility that differences in script (which necessitated adjusted instructions for the Chinese trials) or other linguistic factors impact the likelihood of HBs making use of inhibitory mechanisms.

Meanwhile, the one available experiment with L2ers suggests local inhibition and asymmetric switch cost can be detected during verbal fluency production. One important qualification of the above results from Van Assche et al. (2013) is that participants were tested in letter/phonemic categories only. The current study replicates some elements of the Van Assche et al. experiment by testing participants in both the dominant L1 and non-dominant L2. L2 lexical retrieval is not the focus of this dissertation, and therefore L2 production data is not reported; however, it is possible within the results presented to examine language testing order on precisely the circumstances Van Assche et al. found to result in local inhibition, namely: 1) L1 production, in 2) repeated cues, 3) after testing in L2 first. The current study tested participants in the same cues in both languages and can therefore not confirm or contrast the finding regarding global inhibition.

2.3.2.2. Weaker links

While dual-restriction and dual-task accounts fit under the umbrella of cross-linguistic competition, the "weaker links" account was proposed by Gollan and colleagues (Gollan et al. 2002; Gollan et al. 2008) as an alternative to the competition account of HB lexical disadvantage based on long-term frequency effects. Gollan et al. (2008) distinguish the weaker links theory from competition-based accounts in that it is language-general. Rather than being a result of competition between languages, an issue specific to bilingual speakers, frequency effects pertain to general cognition and effect processing, regardless of language status.

The weaker links account proposes that bilinguals retrieve exemplars from their lexicon independent of non-target language interference. Frequency of vocabulary exposure and usage in either of an HB's languages is presumed to be lower than in corresponding monolinguals due to reduced time per language (see section below on input frequency assumptions for discussion of one challenge to this claim). Less practice retrieving the word forms of each language results in relatively less well-worn conceptual "links" between the semantic representations and the specific phonological forms. These "weaker links" have the result of slowing retrieval, which results in disadvantaged production relative to monolinguals.

This theory offers a separate account for why semantic fluency seems to be disproportionately disadvantaged in bilinguals. Rather than comparing the relative conceptual overlap between concrete and nonconcrete words, as described above, the weaker links account refers back to the semantic/phonological (S-to-P) connections. Hypothetically weaker in bilinguals and therefore resulting in reduced semantic fluency production, S-to-P connections are irrelevant or at the least less necessary in letter/phonemic fluency production, since the latter does not *require* conceptual mediation (though Gollan et al. 2002 admit that it may still inadvertently involve some conceptual activation, as evidenced by strings of semantically related words).

Multi-measure predictions based on the weaker links account, like competition, also predict bilinguals will produce fewer words than monolinguals, longer retrieval latencies, and a similarly low and flat time-course curve, though due to general retrieval slowing from the weaker semantic/phonological links as opposed to cross-linguistic interference. As opposed to competition, Sandoval et al. (2010) argue that weaker links would result in higher average word frequency, because accumulated frequency effects especially impact lower frequency words, making them relatively less likely to be produced.

2.3.2.2.1. Weaker links bilingual input frequency assumptions

Before continuing on to the presentation of evidence from weaker links, it is first important to consider whether the fundamental assumption of weaker links is likely to be valid. The weaker links account assumes that an early heritage bilingual has accumulated lower lifetime frequency of input in either of their languages than a respective monolingual. This assumption seems logical at first blush. After all, a bilingual is allotted 24 hours in a day, same as a monolingual, and any time spent interacting in one language is theoretically time deducted from the total exposure to the other language. However, this assumption relies on the further assumption that all people are exposed to the same amount of linguistic input, a finding which has been challenged, at least for exposure at the beginning stages of development (De Houwer 2013).

De Houwer (2013) measured maternal input, comparing monolingual Dutch and bilingual Dutch-French families with toddlers. Analysis of recordings found extensive within-group variation and no evidence of reduced Dutch input for bilingual children, at least in maternal dyadic interaction. Importantly, this challenge has its limitations, given that De Houwer (2013) only measured maternal input, not total caregiver input. Additionally, while maternal input may vary dramatically between individual mothers, the variety of input sources (e.g. friends, teachers, media) increases as children age and once they enter schooling. Therefore, the lifetime accumulated frequency difference presumed by the weaker links account is still likely.

2.3.2.2.2. Heritage bilingual findings attributed to weaker links

Findings supporting the weaker links theory in early and heritage bilinguals (Gollan et al. 2005; Gollan et al. 2008; Ivanova & Costa 2008) primarily rely on studies of picture naming. Gollan et al. (2005) compared English monolinguals and English-dominant Spanish-English HBs in picture naming in English, with some of the stimuli showing repeatedly, with the justification that if retrieval slowing in bilinguals is due to lower lifetime frequency of exposure to words, bilinguals should benefit from retrieving the same item several times in a short duration. Results showed that bilinguals named pictures slower than monolinguals on the first through third presentations, but by the fourth presentation were naming items as fast as the monolinguals.

Gollan et al. (2008) also compared English monolingual and English-dominant Spanish-English bilingual picture naming, with stimuli divided into high- and low-frequency according to their CELEX (Baayen, Piepenbrock, & van Rijn 1995) frequency per million, which also highly correlated with their LEXESP (Sebastián-Gallés, Martí, Cuetos & Carreiras 2000) Spanish frequency, reasoning that words that are considered low-frequency for a monolingual would have been encountered even less by a bilingual and should therefore show disproportionately long reaction times. The results showed that while bilinguals named pictures in English slower than monolinguals across the board, the gap in performance was significantly larger for low-frequency than high-frequency words. The same holds true between a bilingual's two languages, the dominant and non-dominant; the difference between performance in low- and high-frequency words was larger in the non-dominant language than in the dominant language. Thus, bilinguals in these tasks seem to behave as though their lexicon is composed of lower frequency words, relative to monolinguals (Gollan et al. 2005).

Sharing elements of the above studies, Ivanova and Costa (2008) compared picture naming of repeated stimuli, consisting of both high- and low-frequency words. Participants included Spanish monolinguals and Spanish-dominant Spanish-Catalan bilinguals, and were tested in Spanish; therefore this study included a group of bilinguals being tested in their dominant L1, rather than the L2 (dominant or not) as in most other studies. Bilingual participants named pictures slower than monolinguals. Like Gollan et al. 2008, the performance gap was larger for low-frequency words, though unlike Gollan et al. 2005, the gap shrank but did not close completely by the final repetition. Though not specifically crediting "weaker links," the authors conclude that the most consistent explanation is a "frequency effect in disguise" (p.287).

2.3.2.2.3. L2 learner findings attributed to weaker links

One additional picture naming study (Baus, Costa & Carreiras 2013) found similar frequency effects in L2 learners. Baus et al. (2013) studied the longitudinal performance of German-Spanish L2 students studying abroad. After a semester of L2 immersion, participants named pictures in the L1 slower than upon arrival, but the difference was significant only for low-frequency non-cognate words. The authors interpreted this result as favoring the weaker links account. Specifically, Baus et al. argued the results were problematic according to a competition account, because "one would either expect the same effect for all types of words (global inhibition), or rather, that high-frequency words would suffer a larger detrimental effect because of L2 exposure (local inhibition)" (p. 407), and that the findings therefore aligned more with the weaker links account, as indicating that immersive study abroad experience changes the frequency of use of the L1, therefore weakening specifically the words that were encountered less often before immersion. This conclusion, while aligned with their findings, is problematic in that it presents a different picture of the S-to-P connections than the one proposed originally emphasizing lifetime accumulated frequency.

Baus et al. (2013) also included measures of semantic verbal fluency but found no difference in the total number of words produced nor in the mean frequency of the words produced, though the rate of cognates increased significantly after immersion. While the authors could not account for the discrepancy in results between picture naming and verbal fluency, they suggested that picture naming is perhaps more sensitive than verbal fluency in measuring longitudinal, within-speaker lexical retrieval. The Baus et al. study is limited, however, in that the participant group collapsed students in both intermediate and advanced-proficiency classes, not disaggregating the different proficiency levels for the analysis. Additionally, complicating comparison, the participants in this study were actually multilingual, as they were already fluent in an L2 (English).

Importantly, in Baus et al. (2013) as well as Linck et al. (2009), students were only tested in semantic fluency and were not compared to monolinguals of their L1. To account for those limitations, the participants in the current study were divided into proficiency groups and compared with L1 monolinguals in both semantic and letter/phonemic verbal fluency tasks.

2.3.2.2.4. Findings at odds with weaker links

Though weaker links was proposed in part to account for the HB disadvantage seen in verbal fluency tasks, verbal fluency research based on more than mean word totals has not aligned with predictions made by the weaker links account (Sandoval et al. 2010; Luo et al. 2010). Sandoval et al. (2010) made a series of overlapping predictions based on assumptions of competition, weaker links, and reduced vocabulary accounts. Predictions were based not only on measurements of mean word totals but also mean retrieval latency, which the authors refer to as the “fulcrum point,” and mean word frequency. While their findings regarding retrieval latency were compatible with weaker links, the results of mean word frequency were inconsistent with weaker links' predictions.

Sandoval et al. (2010) calculated the mean CELEX word frequency of the exemplars produced with the program N-Watch (Davis 2005). Mean word frequency was measured to distinguish weaker links from competition. High-frequency words are more present in the input than low-frequency words, so they are earlier to be acquired and have stronger links due to the increased practice. Weaker links predicts that HBs should produce words of a higher average frequency than monolinguals, especially given bilinguals' disadvantage with low-frequency words (Gollan et al. 2005; Gollan et al. 2008). However, high frequency words are also more likely to be concrete, known in both languages and be at a higher state of activation—and therefore competition for activation—during production; thus, a competition model predicts that bilinguals should experience more competition for high-frequency words, and therefore rely more on lower-frequency words (Sandoval et al. 2010). The authors contend that if heritage bilinguals are experiencing cross-linguistic competition, the mean frequency of their exemplars should be lower than that of monolinguals. According to their results, at odds with the weaker links prediction, HBs produced exemplars of lower mean frequency than the monolinguals, at least in semantic categories (Sandoval et al. 2010).

2.3.2.2.5. Weaker links summary

In summary, findings consistent with the weaker links account have come from picture naming, with both early HBs (Gollan et al. 2005; Gollan et al. 2008) and immersed L2 learners (Baus et al. 2013) experiencing frequency effects, especially for low-frequency words. However, Sandoval et al. (2010)'s study, specifically pitting competition and weaker links accounts against one another by comparing both word totals and additional measures, did not find results consistent with weaker links predictions; measurements of mean word frequency supported competition over weaker links in HBs.

For the current study, non-immersed L2 learners were chosen because they were monolingual until the age of exposure to the L2, and therefore grew up developing monolingual-like form-concept "links." Even those who began rudimentary L2 language learning in elementary and middle school still grew up immersed in the L1 and are unlikely to have "weakened" their L1 links through L2 exposure. Baus et al.'s (2013) findings regarding slowed picture naming after a semester of immersion may call that assumption into question. However, Baus et al. did not do a delayed posttest follow-up, so it is possible that, like the L1 verbal fluency performance in Linck et al.'s (2009) study, the L1 impact was temporary and reflective of the unique context of L2 immersion, not of a permanent accumulated frequency effect.

2.3.2.3. Retrieval slowing summary and research question

Much of the bilingual verbal fluency research has cited retrieval slowing to account for reduced word totals in bilinguals relative to monolinguals. Retrieval slowing has been attributed to competition from the target language, the distraction of having to monitor output language in addition to cue restrictions, or the reduced accumulated frequency of words in a bilingual's lexicon. However, the previous studies focused mainly on HB populations, while the work with L2ers has not compared performance to monolinguals.

Based on the findings and limitations of the above described literature, the following research question was proposed:

RQ1: Do L2ers show evidence of retrieval slowing in their L1 production relative to monolinguals?

H1: L2ers would be hypothesized to manifest retrieval slowing through word totals, word frequency effects, and retrieval latencies in the L1 production, which would become evident (or more evident) as proficiency in the L2 increases.

RQ1a: Do L2ers show evidence of competition or dual-tasking in their L1 production?

H1a: This hypothesis predicts that, relative to monolinguals, L2ers will show evidence of competition through: lower word totals, longer retrieval latencies, lower average word frequency, and a time-course with lower intercept and more gradual slope.

RQ1b: Within the competition account, do L2ers show evidence of local cross-linguistic inhibition in their L1 production?

H1b: This hypothesis predicts that, relative to L2ers tested in the dominant L1 first, L2ers tested in the non-dominant L2 first will similarly evidence cross-linguistic inhibition through: lower word totals, longer retrieval latencies, and lower average word frequency.

RQ1c: Do L2ers show evidence of weaker links in their L1 production?

H1c: This hypothesis predicts similar results to competition, but in contrast with competition, L2ers will show evidence of weaker links through: lower word totals and longer retrieval latencies, but higher average word frequency.

2.3.3. Reduced vocabulary

Another factor considered in accounting for performance in lexical retrieval tasks is vocabulary knowledge (Bialystok et al. 2008; Luo et al. 2010; Sandoval et al. 2010). HBs tend to have smaller receptive (Bialystok et al. 2008; Bialystok & Feng 2009; Portocarrero et al. 2007) and productive (Bialystok et al. 2008; Portocarrero et al. 2007) vocabularies in each individual language than monolinguals. Returning to the random-search model as described by Rohrer et al. (1995), smaller search sets more quickly run through the “not-yet-sampled” exemplars and accumulate “previously sampled” exemplars, reaching asymptote. The reduced vocabulary account considers that a reduced lexicon is analogous to a smaller category, and therefore the explanation for disadvantaged HB verbal fluency is as simple as having a smaller pool of exemplars to pull from.

Descriptions of the reduced vocabulary hypothesis (Bialystok et al. 2008; Luo et al. 2010; Sandoval et al. 2010) do not explicitly account for why semantic and letter/phonemic fluency are differently affected. However, it is possible that if overall bilingual vocabulary knowledge in a given language is reduced, the reduced sample size of the relatively smaller semantic cues may be disproportionately impacted relative to the more robust sample size of letter cues.

Multi-measure predictions based on the reduced vocabulary account, like those based on retrieval slowing, also predict bilinguals will produce fewer words than monolinguals. Sandoval et al. (2010) argue that reduced vocabulary would also result in higher average word frequency, like weaker links, because a smaller lexicon is likely to exclude those words that are lowest frequency in use. Based on prior findings that verbal fluency performance is sensitive to category size (Borkowski et al. 1967; Crowe 1998; Rohrer et al. 1995), in which smaller categories result in shorter mean SRLs, reduced target language vocabulary knowledge has been predicted to result in shorter retrieval latencies as the available exemplars more quickly run out[1].

2.3.3.1. Findings attributed to reduced vocabulary

Few studies of HB verbal fluency include an independent vocabulary measure to control for vocabulary knowledge (Bialystok et al. 2008; Luo et al. 2010). Bialystok et al. (2008) conducted two experiments comparing English monolinguals and early HBs of a variety of non-English L1s. Initial results seemed to indicate a bilingual disadvantage for letter/phonemic, but not semantic, verbal fluency, conflicting with other results finding the opposite pattern (Rosselli et al. 2000; Portocarrero et al. 2007) or a disadvantage in both, but especially semantic (Gollan et al. 2002; Sandoval et al. 2010). Importantly, however, receptive vocabulary had been measured based on PPVT scores. When subsequently reanalyzed after median splitting the HBs into high- and low-vocabulary groups (HV/LV, respectively) or using vocabulary scores as a covariate, the pattern of results changed, and the letter/phonemic fluency disadvantage was reduced to only the LVHBs or disappeared completely, respectively. The HVHBs, who were statistically equivalent to the monolinguals in target language vocabulary knowledge, performed equal to the monolinguals in both tasks.

---

[1] Note the alternative SRL prediction proposed in Section 2.2.2 if dual language activation and cross-linguistic competition for selection are assumed. For the purposes of remaining consistent with the predictions of previous literature, that alternative is not proposed in the hypotheses provided in the current chapter. This issue will be taken up again in the Chapter 5: Discussion.

In a second experiment with larger HV and LV bilingual samples, the effect of controlling for vocabulary resulted in several departures from the results of earlier, non-vocabulary matched research. The second experiment replicated the semantic disadvantage found in prior research (Portocarrero et al. 2007; Rosselli et al. 2000), but only for the LVHBs. Rather than finding equivalent (Portocarreero et al. 2007: Rosselli et al. 2000) or disadvantaged (Gollan et al. 2002; Sandoval et al. 2000) letter/phonemic fluency, LVHBs performed equivalently to monolinguals while HVHBs produced *more* words than monolinguals. Based on the results, the authors surmised that decreased semantic fluency productivity in previous studies may reflect vocabulary size.

Given that most verbal fluency studies do not include objective assessment of vocabulary knowledge, it could be argued (as in Bialystok et al. 2008) that previous findings of disadvantaged verbal fluency are actually reflective of vocabulary knowledge in disguise. This could be the case, for example, with Portocarrero et al. (2007), whose bilingual sample produced fewer semantic fluency exemplars but who also performed worse than monolinguals in receptive (PPVT) and expressive (EVT) vocabulary measures; the authors mention this possibility only in passing.

2.3.3.2. Findings at odds with reduced vocabulary

To further investigate the impact of vocabulary size on HB verbal fluency production requires examining beyond word totals. Sandoval et al. (2010) looked at mean word frequency and retrieval latencies to distinguish between possible accounts. Unlike Bialystok et al. (2008), they did not include an objective target language vocabulary measure, but predicted that reduced vocabulary knowledge would result in the HBs producing fewer words combined with shorter mean SRLs and higher mean word frequency relative to monolinguals. Their results found that HBs did produce fewer words than monolinguals, but in contrast with reduced vocabulary predictions, HBs had longer mean SRLs and higher mean word frequency, from which they concluded against reduced vocabulary and in favor of retrieval slowing.

Luo et al. (2010) also examined retrieval latencies, though not word frequency, of HBs median split into LV and HV groups, the latter of which were statistically equivalent with monolinguals in their target language vocabulary knowledge (as in Bialystok et al. 2008). Though their word total results did not replicate the semantic disadvantage of the LVHBs found in Bialystok et al.'s second experiment, they did replicate the letter/phonemic advantage. However, the letter/phonemic mean SRLs for both bilingual groups were equivalent to each other and longer relative to the monolinguals. Luo et al. focus on the HVHBs to conclude that bilingualism affects latencies through enhanced task control (discussed in the presentation of enhanced executive control, below). What Luo et al. (2010) did not address was the interpretation for the LVHBs. As explained above, reduced vocabulary is postulated to result in shorter SRLs, based on research manipulating category size (Borkowski et al. 1967; Rohrer & Wixted 1994; Rohrer et al. 1995), such that lesser vocabulary knowledge is likened to retrieving words from a smaller pool of exemplars. Therefore, LVHBs would be predicted to have shorter mean SRLs than both monolinguals and HVHBs. Instead, the LVHBs, who objectively demonstrated lesser target language vocabulary knowledge, produced mean SRLs longer than monolinguals and equivalent to their HV counterparts.

This result could be interpreted multiple ways. One possibility is that lexicon size as measured by the PPVT highlights differences mainly in low-frequency vocabulary, while knowledge of high-frequency vocabulary of the type typically assessed with verbal fluency measures is similar between the groups; this account is unlikely in that it does not explain why the HB groups differed in words produced. It is also possible that overall lexicon size does not result in the same retrieval latency effects as the size of the particular category. There is currently no retrieval latency research comparing varying levels of general lexical knowledge to assess that possibility.

A third possibility, in line with the authors' conclusion that both lexical knowledge and bilingualism combine to account for performance, is that the task control advantages of bilingualism work to counteract the disadvantage in vocabulary knowledge, making up the difference. If the case, we might expect LVHB mean SRLs to be closer to equivalent to monolinguals (24.1s), or perhaps longer depending on the strength of the bilingual advantage, though still shorter than HVHBs. In fact, the numerical trend did follow this pattern; LVHB mean SRLs (25.9s) fell in between monolinguals (24.1s) and HVHBs (26.7s), though the difference between the HB groups was not significant. This possibility would also account for why Sandoval et al. (2010) found longer mean SRLs in their bilingual population, though they produced fewer words, suggesting that their general HB population was more akin to Luo et al.'s LVHBs. However, this account does not explain Sandoval et al.'s mean word frequency results.

2.3.3.3. Reduced vocabulary summary and research question

In summary, target language vocabulary knowledge has been demonstrated, in the relatively few studies that have measured it objectively, to have an impact on verbal fluency performance. HBs with demonstrably smaller target language vocabularies than monolinguals have produced fewer words in semantic verbal fluency (Bialystok et al. 2008), opening up prior research with similar results to potential reinterpretation. However, the proposed interplay between vocabulary knowledge and executive control complicates the detection and interpretation of vocabulary as a factor in bilingual verbal fluency; retrieval latency predictions based on category size and extrapolated to lexicon size have been found to be inaccurate (Luo et al. 2010) as LVHBs still demonstrate longer mean SRLs than monolinguals instead of the predicted shorter SRLs. However, vocabulary knowledge within HBs is dissociated in initial retrieval, as seen in time-course graphing (Friesen et al. 2015; Luo et al. 2010).

Importantly, the studies that account for vocabulary knowledge have all been performed with HBs, not L2ers. HBs are often only tested in their L2 (or perhaps co-L1, in the case of simultaneous bilingualism); though it may have become the dominant or at least equally proficient language, they typically control a smaller vocabulary than monolinguals. However, L2 learners and proficient bilinguals are assumed to have monolingual-like L1 vocabulary knowledge. While likely a correct assumption, given the years of monolingualism before beginning L2 study and the maintenance of the L1 immersive environment (mainly, apart from stints studying or living abroad), it is important to objectively account for L2er vocabulary knowledge in the L1 to be able to assert that the time spent in L2 study does not result in a reduced L1 lexicon.

Based on the findings and limitations of the above described literature, the following research question was proposed:

RQ2: Do L2ers show evidence of reduced vocabulary in their L1 production?

H2: If time dedicated to formal L2 study results in decreased L1 vocabulary knowledge relative to monolinguals, which would become evident (or more evident) as proficiency in the L2 increases, this hypothesis predicts that L2ers will evidence reduced vocabulary through lower scores on English vocabulary measures. Within the verbal fluency tasks, as with retrieval slowing above, L2ers will evidence lower word totals. Similar to weaker links, this account predicts higher average word frequency. Unlike either retrieval slowing account, this account predicts shorter retrieval latencies.

2.3.4. Executive control advantage

The coexistence of two languages and the need to selectively activate and/or inhibit one language at a time, as presented above in models of bilingual lexical retrieval, is thought to be the locus of retrieval slowing in the competition account. However, dual language management can also constitute mental exercise that arguably results in bilingual advantages in executive control, a nonverbal cognitive ability associated with a variety of nonlinguistic tasks such as conflict resolution, attentional control, and task switching (e.g. Bialystok 2009, 2011; Prior & MacWhinney 2010; though see e.g. Paap & Greenberg 2013; Paap et al. 2014).

Rather than assuming a bilingual lexical retrieval disadvantage, some research with vocabulary-matched bilinguals has found evidence of *greater* word production, specifically in letter fluency, which has been attributed to bilingual advantages in executive control (Bialystok et al. 2008; Luo et al. 2010). In addition to the retrieval slowing (though not distinguishing between cross-linguistic interference and weaker links) and reduced vocabulary predictions made by Sandoval et al. (2010), all of which assume bilinguals will produce lower word totals, Luo et al. propose an additional alternative combining retrieval latency with higher word totals, given Bialystok et al. (2008)'s finding that vocabulary-matched HBs can produce *more* verbal fluency exemplars relative to monolinguals, specifically in letter/phonemic trials. While longer mean SRLs potentially reflect retrieval slowing, Luo et al. point out they could instead indicate a slower decline in production, or an enhanced ability to maintain production over time, due to bilingual executive control advantages.

The executive control advantage hypothesis explicitly considers bilingualism's disproportionate impact on semantic vs. letter/phonemic verbal fluency, suggesting that advantages can accrue in letter/phonemic fluency, as the more challenging task, typically requiring more monitoring of the output due to increased restrictions. While both tasks require lexical knowledge and executive control, semantic fluency is considered to be more reflective of the former, while letter/phonemic fluency is more indicative of the latter (Portocarrero et al. 2007; Rosselli et al. 2000; Shao et al. 2014). Early and heritage bilinguals, having practiced task control all or most of their lives through language switching, have an executive control advantage that manifests as enhanced performance in letter/phonemic fluency after controlling for vocabulary knowledge.

Multi-measure predictions based on the enhanced executive control account, unlike those based on retrieval slowing or reduced vocabulary, predict bilinguals will produce more words than monolinguals due to their being able to more successfully manage the demands of the task over time. Being able to continue producing exemplars later into the time period will also result in longer mean SRLs, but not necessarily longer FRLs, and a more gradual time-course curve, but one with equivalent rather than lower initiation.

2.3.4.1. Heritage bilingual evidence attributed to executive control advantage

Only three HB verbal fluency studies consider the potential impact of executive control (Bialystok et al. 2008; Friesen et al. 2015; Luo et al. 2010), with two studies examining performance beyond word totals. Luo et al. (2010) compared the word total, retrieval latency, and time-course performance of monolingual English speakers and English-speaking HBs of a variety of other languages, median split (as in Bialystok et al. 2008) into HV and LV groups based on receptive vocabulary knowledge measured by the PPVT. They found that vocabulary-matched HVHBs again outperformed monolinguals in word totals, specifically in the letter/phonemic category. In line with their prediction based on enhanced task control, HBs, regardless of vocabulary group, produced longer mean SRLs than monolinguals. Arguing that retrieval slowing would result in lower word totals, the authors reject a retrieval slowing account in favor of enhanced task control, i.e. a more gradual decline in production over time, or the ability to better maintain a rate of production later into the minute-long trial, due to bilingual executive control advantages. This interpretation is supported by the time-course graph and analyses, which showed that HVHBs and monolinguals had equivalent initial recall, but the slope of the HVHBs was more gradual, indicating less of a decline in production over time than the monolinguals.

Friesen et al. (2015) expanded Luo et al. (2010)'s findings to include multiple age groups, by incorporating their HVHBs and monolinguals with additional data on HB and monolingual children and elderly adults. Results with children partially replicated earlier results; 5th grade HBs, who were not vocabulary-matched with their monolingual counterparts, produced fewer semantic exemplars and had a lower initial recall in the time-course analysis, much like the young adult LVHBs. In addition, the performance gap between semantic and letter/phonemic fluency was smaller for the HBs than monolinguals, suggesting compensation through executive control. Elderly bilinguals, like the young adult HBs from Luo et al., also produced more letter/phonemic exemplars and longer mean SRLs than the monolinguals.

2.3.4.2. L2 learner evidence attributed to executive control advantage

Though substantial research has been published regarding executive control advantages in HB populations, especially children (e.g. Bialystok 2011), the context of late, formal L2 learning is quite different. Multiple studies (Carlson & Meltzoff 2008; Kaushanskaya et al. 2014; Poarch & van Hell 2012) compared executive control in children, overall finding that HBs, but not L2 learners, outperformed monolinguals. Specific to Poarch and van Hell (2012), the child L2 learners' performance fell in between monolinguals and HBs such that there was a numeric but not statistically significant advantage over monolinguals, and the authors optimistically conclude that "the L2 learners' enhanced attentional control was emerging" (p.548). These results indicate that short periods of L2 learning, even in immersion classrooms, are insufficient for children to develop demonstrably advantaged executive control, though continued study and exposure could result in ultimate advantages.

One study considers the potential impact of executive control on the bilingual verbal fluency of L2ers (though the primary focus of the study was episodic memory). Ljungberg et al. (2013) drew data from the Betula Prospective Cohort Study, comparing Swedish monolinguals with Swedish-speaking L2ers (L2 typically English). Participants were tested in the L1, completing only one cue per category. Oddly, the semantic fluency category was actually a dual-restriction category, consisting of a semantic restriction and additional letter/phonemic restriction (occupations starting with B). The analyses only examine word totals, with the results finding that bilinguals produced equivalent words relative to monolinguals in semantic fluency but more words in letter/phonemic fluency, an effect that was maintained over multiple longitudinal testing periods.

Though the authors correctly predicted no semantic fluency advantage, based on the findings of Luo et al. (2010), it is in fact curious that such an effect was not found, given the letter/phonemic component in the supposed semantic task, which has been shown previously (Azuma et al. 1997; Gollan et al. 2002) to increase task difficulty and therefore should have required more executive control to manage. However, lack of a semantic disadvantage, which Bialystok et al. (2008) and others have associated with reduced vocabulary, suggests that the groups were similar in target language lexical knowledge and therefore formal L2 study is unlikely to result in an L1 vocabulary deficit as is frequently found with heritage populations. Though the study is limited given its methodology, e.g. the singular cues and conclusions based only on word totals, it serves as one example demonstrating that the language experiences characteristic of HBs (early, natural, immersed exposure) are not necessary conditions to observe bilingual advantages.

2.3.4.3. Findings at odds with executive control advantage

Friesen et al. (2015) expanded on Luo et al. (2010) by comparing HBs and monolinguals of various age groups. In the youngest group, 2nd graders who were matched in target language vocabulary, HBs produced longer mean SRLs but equivalent words, consistent with accounts of retrieval slowing. Though multiple factors including age and literacy may account for the differences, this result does suggest that vocabulary and executive control do not retroactively account for all previous HB verbal fluency findings, and retrieval slowing may not be ruled out entirely as a factor.

2.3.4.4. Executive control advantage summary and research question

Across the relatively brief history of bilingual verbal fluency research, studies considering the impact of executive control advantages are recent and encouraging. They have mainly found that, lexical resources being equal, bilingualism likely confers an advantage in task control, allowing bilinguals to produce more words specifically in the task most associated with executive control processes. This result in HB populations has been successfully replicated, and preliminary evidence exists for L2er populations as well.

Based on the findings and limitations of the above described literature, the following research question was proposed:

RQ3: Do L2ers show evidence of enhanced executive function in their L1 production relative to monolinguals?

H3: L2ers would be hypothesized to manifest enhanced executive function through increased retrieval control, which would become evident (or more evident) as proficiency in the L2 increases. Unlike the previous accounts, this hypothesis predicts that, relative to monolinguals, L2ers will show evidence of executive control through: higher word totals and a time-course with equivalent intercept but more gradual slope, combined with (similar to competition and weaker links accounts) longer retrieval latencies.

2.3.5. Verbal fluency summary and conclusions

To summarize, verbal fluency tasks measure the ability to quickly retrieve exemplars from the lexicon given a particular category restriction. Both of the two common category types, semantic and letter/phonemic, recruit both lexical knowledge and executive control, though semantic fluency is more associated with the former, and letter/phonemic fluency the latter. Semantic fluency is classically considered the easier of the two category types, because it is conceptually mediated and can take advantage of semantic network organization of lexical storage, which aligns with the meaning-centric direction of processing during production, meaning→form. Letter/phonemic fluency, on the other hand, involves production without regard to meaning, potentially bypassing conceptual mediation and able to take advantage only of phonemic word connections, which run counter to productive retrieval processes.

Previous verbal fluency research comparing monolinguals and bilinguals, both heritage and L2, has found mixed results. Word totals are the most common measure reported, with earlier studies reporting a bilingual disadvantage and more recent work countering with evidence of a bilingual advantage. Regarding category type, the findings are also not entirely consistent. Some studies have found disadvantages in both (Gollan et al. 2002; Sandoval et al. 2010) or just semantic (Linck et al. 2009; Portocarrero et al. 2007; Rosselli et al. 2000), while others found equivalent semantic and advantaged letter/phonemic (Bialystok et al. 2008; Friesen et al. 2015; Ljungberg et al. 2013; Luo et al. 2010) after controlling for target language vocabulary knowledge. Beyond word totals, verbal fluency production data can be analyzed for mean word frequency and retrieval latency. HBs have been found to produce longer mean SRLs, which when combined with fewer, lower-frequency words and longer FRLs was interpreted to support retrieval slowing through competition (Sandoval et al. 2010), while combined with more words and a more gradual decline over time was interpreted to indicate a task control advantage (Friesen et al. 2015; Luo et al. 2010).

Of the modest amount of bilingual verbal fluency research to the current time, most has focused on heritage and early bilinguals, though even within this subgroup there has been much variation in population. Across studies, participants have differed dramatically in age and language background, studying heterogeneous groups with mixed reports of first vs. second language, dominant vs. non-dominant language, age of acquisition of the L2 (some before and some after puberty), and the similarity between the L1 and the language of testing (typically English). However, language background and participant classification has relied almost exclusively on self-ratings and only rarely (Bialystok et al. 2008; Friesen et al. 2015; Linck et al. 2009; Luo et al 2010) on objectively measured proficiency through vocabulary knowledge or other linguistic performance.

Only a handful of studies have examined the production of L2 learners/bilinguals. Of that subset, only one compared bilinguals to monolinguals and studied production in both semantic (though see above description for reservations regarding that label) and letter/phonemic categories (Ljungberg et al. 2013). No studies have looked at L2ers at multiple proficiency levels, and no published work has examined L2 verbal fluency production beyond word totals, though an unpublished analysis of the data from Linck et al. (2008) suggested L1 retrieval slowing during L2 immersion.

Given the limitations of the previous research, the current project compares the L1 verbal fluency performance of monolingual English speakers to English-Spanish L2ers at a variety of proficiency levels (low, mid, high). Though L2 production is not the focus of the dissertation and therefore will not be reported, L2ers were tested in both languages in order to allow for examination of cross-linguistic order effects. Participants completed trials in both semantic and letter/phonemic fluency, since the two have been shown to be differentially impacted by bilingualism. To facilitate distinguishing between the various theories set forth to account for bilingual production differences, word total analyses were complemented by comparisons of mean word frequency, retrieval latencies, and time-course graphing. The following chapter presents the methodology of the study in detail, including description of participant characteristics, materials, coding and scoring of performance.

Chapter 3: Methodology

3.1. Introduction

The overarching questions posed in this dissertation are how L1 lexical retrieval of L2ers compares to lexical retrieval in monolinguals, whether L2 proficiency impacts L1 lexical retrieval, and whether L2ers demonstrate evidence of retrieval slowing, reduced L1 vocabulary knowledge, or enhanced executive control. One measure of lexical retrieval which can be analyzed to address those questions is performance on tasks of verbal fluency, which measure the speed and ease with which someone can access words in their lexicon according to a given category. Much of the verbal fluency work has been done with HBs, for whom the performance results are mixed. Early studies found poorer performance on these tasks relative to monolinguals (Gollan et al. 2002), a finding that has been attributed variously to cross-linguistic interference (Rosselli et al. 2000; Sandoval et al. 2010), frequency effects (Gollan et al. 2002), and vocabulary deficits (Bialystok et al. 2008). On the other hand, studies comparing monolinguals with vocabulary-matched HBs (Bialystok et al. 2008; Friesen et al. 2015; Luo et al. 2010) have found that HBs can produce *more* words than monolinguals, which the authors attribute to a bilingual executive control advantage.

Regarding L2ers, relatively less is known. L2 immersion has been found to temporarily result in decreased L1 semantic verbal fluency relative to classroom-instructed peers (Linck et al. 2009; Gerfen et al. in preparation), while the one study comparing L2ers to monolinguals (Ljungberg et al. 2013) replicated the HB finding of greater word totals. The current dissertation measures the L1 verbal fluency performance of L2ers at multiple proficiency levels, to investigate how it compares to monolinguals. In this chapter, I describe the four participant groups who participated in the study, as well as the data collection procedures and materials.

This study examined English-Spanish L2ers at three different Spanish ability levels (low, mid, high), as well as English monolinguals. The L2er population is well suited for distinguishing between theories that have been proposed for HB verbal fluency performance, as well as investigating the impact of age and manner of acquisition on bilingual lexical retrieval. Language chronology and dominance are often mixed in HBs, making it difficult to disentangle the effects of proficiency, age of acquisition, and amount of exposure. In contrast, L2ers spent their formative years monolingual, and even after beginning to study the L2 were predominantly immersed in the L1, resulting in a more stable and homogenous pattern of language chronology and dominance and reducing the likelihood of two of the proposed theories: weaker links or vocabulary deficits.

In this dissertation, the research question inquires as to the impact of a relatively late-acquired, non-immersive L2 on L1 lexical retrieval, specifically how developing L2 proficiency is reflected in L1 performance relative to monolinguals in L1 verbal fluency tasks. Whether the L2ers' dominant L1 is affected by the presence and proficiency of the L2 can be observed by comparing the L1 verbal fluency performance of monolinguals with L2ers of varying proficiency levels. Multiple hypotheses are presented according to various accounts that have been proposed: retrieval slowing through competition or weaker links; reduced vocabulary; and enhanced executive control. If cross-linguistic competition or reduced L1 vocabulary due to L2 exposure is responsible for bilingual disadvantages in lexical retrieval, L2ers should demonstrate decreased L1word totals relative to their increasing proficiency in the L2. If instead, bilingualism creates an executive control advantage, the higher proficiency L2 bilinguals may evidence increased L1 production relative to monolinguals.

To examine the hypotheses presented, this dissertation presents performance data for monolingual and bilingual participants in the following: 1) English vocabulary measures (picture vocabulary identification, lexical judgment), 2) English verbal fluency tasks. English vocabulary measures provide an objective way to determine whether monolingual and bilingual L1 vocabulary knowledge is in fact comparable, while the English verbal fluency tasks examine the specific linguistic phenomenon under investigation: L2 learner/bilingual lexical retrieval in the L1. To assess the L2 proficiency of the L2ers, they also completed: 3) Spanish vocabulary measures (picture vocabulary identification, lexical judgment) and a 4) Spanish grammatical proficiency test. Spanish vocabulary knowledge, as opposed to simple self-rating, was used to classify participants into three ability groups, which were also compared on grammatical knowledge to further confirm group proficiency differences. Finally, while not the focus of the present dissertation, L2ers also completed 5) Spanish verbal fluency tasks, to allow for the testing of language order effects and cross-linguistic inhibition by comparing participants who completed the verbal fluency tasks in English first versus Spanish first.

This chapter is outlined as follows: In section 3.2., I describe the participant groups, including their demographic and linguistic background information. In section 3.3, I summarize the data collection, the settings and procedures used as well as the materials. In section 3.4, I detail the data analysis and the various measures coded from the verbal fluency tasks.

3.2. Participants

Participants in this study consisted of 122 monolingual and bilingual individuals, 32 English monolingual controls and 90 English-Spanish L2 learners/bilinguals who were classified at three levels of Spanish ability according to their scores on measures of Spanish vocabulary knowledge: 28 low (LVL2), 32 mid (MVL2), 30 high (HVL2).

3.2.1. Monolinguals

Thirty-two monolinguals (15 male), ranging in age from 19 to 37 (mean age = 25.06, s.d.= 4.76) volunteered to participate in the study. Monolingual participants were recruited from undergraduate classes in linguistics and education, as well as referred by bilingual participants. Those in undergraduate classes participated for class credit, while the rest volunteered to participate. Given that most universities have second language study requirements for admissions, some form of high school level language study was expected and unavoidable; therefore, only students who reported no functional knowledge of a language other than English were considered. Seventeen of the 32 monolingual participants indicated some previous language study on their own or in school, including ASL, Chinese, French, German, Italian, Latin, Spanish and Tagalog, but rated themselves an average of 1.06 out of 10 in communication ability (see Table 3.2.).

3.2.2. L2 learners/bilinguals

The L2ers were learners of Spanish at various stages of study and ability, currently studying or regularly using Spanish. Many participants in the MVL2 and HVL2 groups had experience studying and/or living abroad in a Spanish-speaking country, but had primarily learned in formal second language classes in the United States. Linck et al. (2009) found temporarily diminished L1 access in immersed L2 learners relative to their non-immersed classmates, an effect which disappeared when tested again six months after returning home from study abroad. Given that participants were not posttested at any other interval, the study does not indicate precisely how long any potential effect of L2 immersion lasts, just that performance matches non-immersed levels by at least six months post-immersion. Therefore, only participants who had not been studying or living abroad in a Spanish-speaking country in the last six months were eligible to participate.

One of the primary research questions addressed by the current study is whether a late-learned L2 *can* compete with the L1 for production; therefore, participants with experience studying an L2 in addition to Spanish were considered eligible for participation. Of the 90 L2 subjects, 23 reported having studied an additional foreign language on their own or in school, including Arabic, ASL, Chinese, French, German, Italian, Hebrew, Japanese, and Portuguese, and self-rated their communication ability at 3.65 out of 10 (see Table 3.2.).

3.2.2.1. Low vocabulary L2

L2 participants scoring 55% or below on the combined vocabulary measure were classified into the low vocabulary (LV) group. The LVL2 group consisted of 28 students

(4 males), ranging in age from 18 to 26 (mean age = 19.39, s.d.= 1.66). Subjects were undergraduate students taking intermediate and advanced level courses in the Spanish department and participated for class credit. One subject reported having studied in a Spanish-speaking country, for five weeks, but not within the six months preceding participation.

3.2.2.2. Mid vocabulary L2

L2 participants scoring between 55-75% on the combined vocabulary measure were classified into the mid vocabulary (MV) group. The MVL2 group consisted of 32 subjects (8 males) ranging in age from 18 to 29 (mean age=20.31, s.d. = 2.26). Subjects included undergraduate students in advanced level courses and graduates of Spanish bachelor's degree programs who continued to make use of their Spanish. The undergraduate students were recruited from upper level literature, linguistics, and translation courses, and were typically in their final year or two of university Spanish study; they participated for class credit. The previous graduates included teachers of Spanish as well as others who were referred to the study by other Spanish students; they volunteered to participate. Ten of the 32 subjects reported having studied abroad in a Spanish-speaking country, seven of whom had studied for five weeks or less, while the other three had spent a semester or more abroad, but as stated earlier, not within the six months preceding participation.

3.2.2.3. High vocabulary L2

L2 participants scoring higher than 75% on the combined vocabulary measure were classified into the high vocabulary (HV) group. The HVL2 group consisted of 30 subjects (12 males) ranging in age from 18 to 40 (mean = 29.63, s.d. = 6.03). Subjects primarily consisted of current and former graduate students as well as undergraduate students in advanced courses. The undergraduate students participated for class credit, while the graduates were colleagues of the researcher and referrals from other participants, all who volunteered to participate. Twenty-three of the 30 participants reported having studied abroad in a Spanish-speaking country. Thirteen reported having spent less than a year, and the other ten reported spending a year or more abroad. As with the other L2 groups, none of the time spent studying or living abroad had been in the six months prior to participation.

3.2.3. Demographics and education

Overall, participants consisted of 39 males and 83 females between the ages of 18 and 40 (mean = 23.61, s.d. = 5.74). More women than men tend to enroll in university language courses, making gender difficult to control for the purposes of this study; however, normed monolingual data found no significant contribution of gender to production variation (Tombaugh et al. 1999), so there was little reason to suspect it would be problematic.

Age and education are significant contributors to verbal fluency performance, with age accounting for more variance in semantic categories, and education accounting for more variance in letter/phonemic categories. Regarding age, the differences are concentrated in the transition from childhood to adulthood and again after age 60, while the bulk of the adult years 18-60 vary little (Tombaugh et al. 1999). Given that L2 proficiency grows with experience, it is unsurprising that the HVL2 group is older than their LVL2 and MVL2 counterparts, and since all participants were well within the adult age range, the age difference in groups was not considered problematic. In education, meanwhile, monolingual production means increase from high school education to college education—in both semantic and letter categories—but further education beyond university level shows little increase. To attempt to control for education, all participants had at least some university education. Reported highest level of education is summarized in Table 3.1.

Table 3.1. Highest level of education completed

	Monolingual	LVL2	MVL2	HVL2
Some undergraduate	11	26	20	4
2-year undergrad degree	6	1	2	1
4-year undergrad degree	8	1	8	2
Some graduate	3	0	0	2
Masters Degree	1	0	1	15
Ph.D or Prof. Degree	3	0	1	6
Total	32	28	32	30

As can be seen from Table 3.1., the HVL2 group had the highest concentration of advanced degree seekers and recipients. As was the case with age, since L2 proficiency grows with experience and the HVL2 group consisted in large part of graduate students and credentialed professionals who had continued formal study of Spanish beyond the undergraduate major, it is unsurprising that the groups differ in educational attainment. Though this group difference was largely unavoidable, its impact is weighed in the analysis of the results and subsequent discussion of limitations.

3.2.4. Language background

All subjects reported growing up speaking only English in the home, excluding non-native English speakers and heritage speakers of other languages. Subjects completed a language background questionnaire, reporting at what age they began learning Spanish (birth assumed for English), as well as their self-rated communication ability in English, Spanish, and any other language they had studied. All participants rated themselves 10 out of 10 in English. Table 3.2. presents the self-reported second language characteristics.

As can be seen in Table 3.2., the average age of exposure to Spanish was approximately middle school. A series of one-way ANOVAs were conducted to compare language background characteristics. A significant group difference was found in the age of beginning Spanish learning, $F(3,103) = 4.736$, $p = .004$. Post hoc comparisons determined that the MVL2 group started learning Spanish at a younger age than the HVL2 group, while no other group differences were significant. However, age of exposure did not correlate with any of the objective proficiency measures (described below), making it unlikely that group differences accounted for performance. Group differences in Spanish self-rating were significant, $F(3,118) = 176.48$, $p < .001$, in

Table 3.2. Second language background characteristics

	Monolingual		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Age learning Spanish	12.53[b]	3.67[b]	11.63	2.60	10.32	3.12	13.07	2.68
Self-rating[a] Spanish	1.34[b]	1.07[b]	5.04	0.92	6.16	1.37	8.06	1.26
Self-rating[a] other language	1.06[c]	1.18[c]	-	-	3.25[c]	2.70[c]	3.87[c]	2.33[c]

[a] Proficiency level based on self-rating using a scale of 0-10, "0 being no ability to communicate, 10 being able to communicate like a native speaker."
[b] Calculated based on the 19 of the 32 monolinguals who reported having studied some Spanish
[c] Calculated based on the 16 monolingual, eight mid-proficiency, and 15 high-proficiency participants who reported having studied a non-Spanish language

that each group self-rated significantly differently than every other (all p values < .005). There was also a group difference in other L2 self-rating, F (2,36) = 8.037, p = .001, in that both the HVL2 (p = .001) and MVL2 (p = .05) rated themselves more proficient in a non-Spanish L2 than the monolinguals (further emphasizing the monolinguals' lack of proficiency in any L2), with no difference between the two bilingual groups, though the score is still rather low (around 3/10).

3.2.5. Second language proficiency

Previous studies have relied primarily on self-rating to determine language proficiency. Notable exceptions include Bialystok et al. (2008) and Luo et al. (2010), who used target language (English) vocabulary knowledge, as measured by the PPVT, to classify HBs into high and low vocabulary groups. Linck et al. (2009) used accuracy and latency on an L2 translation-recognition task as a covariate to account for L2 proficiency in their L2ers. While self-rating of dominance and/or proficiency is common, it is not without problems. For instance, in both Bialystok et al. (2008) and Luo et al. (2010), LV and HV groups differed in their objective English vocabulary knowledge, but did not differ in their self-rated English proficiency, calling into question the accuracy of self-rating. Especially given the dissociable effects of bilingualism and linguistic resources (vocabulary) found by Luo et al. (2010), it was deemed important to objectively measure the L2 proficiency of the L2ers, rather than simply trusting self-rating.

Participants completed two vocabulary measures: a Spanish multilingual picture vocabulary test (MPVT-Esp) adapted from the Multilingual Naming Test (MiNT, Gollan et al. 2012) and a Spanish lexical judgment task, the Lextale-Esp (Izura et al. 2014). Spanish proficiency for the purposes of participant classification was calculated by averaging the percentage scores of the two vocabulary measures. The ensuing score was used to categorize L2er participants into the three groups described above.

Percentage cut-offs were arbitrary, but a one-way ANOVA found significant group differences in combined vocabulary, $F(2, 87) = 361.127$, $p < .001$. A mixed ANOVA with group (LVL2, MVL2, HVL2) as the between-subjects variable and vocabulary test (MPVT-Esp, Lextale-Esp) as the within-subjects variable found a significant interaction between group and test $F(2, 87) = 18.588$, $p < .001$, partial $\eta^2 = .299$. Univariate analyses of the simple main effect of group found significant differences in both measures, (MPVT-Esp $F(2,87) = 244.503$; Lextale-Esp $F(2,87) = 143.328$), such that each learner/bilingual group performed significantly differently than the other two (ps $\leq .002$).

Since the participants were formal L2 learners (as opposed to HBs), a grammatical proficiency measure, a portion of the DELE, was added to complement the information provided by measures of receptive vocabulary knowledge. The DELE score highly correlated with both individual and combined measures (ps $< .001$). Additionally, a one-way ANOVA found significant group differences, $F(2, 87) = 126.688$, $p < .001$, and post hoc tests confirmed that each group performed significantly differently than the other two (ps $< .001$). Table 3.3. presents the mean scores and standard deviations for both individual vocabulary tests, as well as the DELE grammatical proficiency test. Though objective vocabulary and grammar knowledge measures were included in addition to self-ratings, it can be noted that Spanish self-rating did in fact highly correlate (ps $< .001$) with all Spanish proficiency measures.

Table 3.3. Spanish vocabulary and grammar scores (%)

	LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*
MPVT-Esp	48.61	8.47	68.42	8.33	91.69	4.97
Lextale-Esp	52.68	4.96	59.04	5.78	81.67	9.21
DELE	32.14	10.27	48.18	11.94	78.61	11.63

In summary, participants included monolingual English speakers and English-Spanish L2 learners/bilinguals at three different ability levels—low, mid, and high. The monolingual English group served as a control for comparison with the L1 performance of L2ers. Dividing L2ers in three different levels enabled investigation of the impact of varying degrees of L2 proficiency on L1 lexical retrieval. Specifically, the study sought to determine whether non-immersed (currently or recently) L2ers demonstrated L1 lexical retrieval effects relative to monolinguals, and at what level of proficiency effects manifested, if at all.

3.3. Materials and procedures

Participants began by completing a short demographic and language background questionnaire to determine their eligibility to participate. All subjects completed a series of vocabulary tasks in English to measure vocabulary knowledge. HBs typically have a smaller vocabulary in each of their languages relative to a monolingual (Bialystok et al. 2008; Hoff et al. 2012), and vocabulary size has been proposed to account for verbal fluency performance. Therefore, L1 vocabulary was measured to assess whether L2 learners/bilinguals have monolingual-like vocabulary knowledge, or whether time spent formally learning an L2 reduces L1 exposure sufficiently to result in a reduced L1 vocabulary pool. In addition to measures of English vocabulary, L2ers completed corresponding measures of Spanish vocabulary, which were used to classify them into groups, as well as a brief multiple choice sentence completion task to measure Spanish grammatical proficiency. The questionnaire and all vocabulary and grammar measures were completed through an online survey.

The experimental tasks for all participants included verbal fluency trials in English, to assess whether non-immersed formally acquired L2 bilingualism impacts L1 lexical retrieval for production, perhaps through increased competition (Gollan et al. 2003) or alternatively, enhanced executive control (Luo et al. 2010), additional theories hypothesized to account for mixed results in studies with HBs. Of the limited L2 learner verbal fluency literature, some tested only one category type (Baus et al. 2013; Linck et al. 2009; Van Assche et al. 2013), while others (Ljungberg et al. 2013; Snodgrass & Tsivkin 1995) as well as various HB studies (Bialystok et al. 2008; Friesen et al. 2015; Gollan et al. 2003; Luo et al. 2010) have suggested that semantic and letter/phonemic fluency categories experience differential effects of bilingualism. Therefore, both semantic and letter/phonemic categories were included. Bilinguals also completed the same verbal fluency tasks in Spanish, to allow for subsequent investigation of cross-linguistic inhibition through comparison of results by language testing order. Verbal fluency tasks were digitally recorded separately from the online survey.

3.3.1. Biolinguistic questionnaire

The biolinguistic questionnaire included demographic questions on age, sex, and highest level of education completed (see Section 3.2.3) as well as language background questions to determine bilingual status, heritage speaker status, and self-rated proficiency. Participants were asked to state their home language to identify heritage speakers of Spanish and other languages. In addition, they were asked whether they had studied abroad in a Spanish-speaking country, and participants who reported study abroad in the last six months were excluded. Participants also reported whether they had studied any other language and finally, self-rated their ability to communicate in (see Table 3.1.) English, Spanish, and any other language they had studied.

3.3.2. Vocabulary measures

After completing the biolinguistic questionnaire, as part of the online survey, participants completed vocabulary measures. Participants indicated their monolingual or bilingual status in the biolinguistic questionnaire, and were directed to the rest of the tasks accordingly. Monolinguals were provided with the English versions of two complementary vocabulary measures. Bilinguals completed the same vocabulary measures and then went on to complete corresponding Spanish versions of the same tasks. None of the tasks was timed.

English vocabulary measures were included to determine whether L2 bilinguals have monolingual-like vocabulary size, regardless of L2 study. Previous research has found that vocabulary size can account for bilingual verbal fluency performance (Bialystok et al. 2008; Friesen et al. 2015; Luo et al. 2010), in that bilinguals with lower vocabulary knowledge produce fewer words than monolinguals, while disadvantages in verbal fluency production disappear when the bilinguals are matched with monolinguals on vocabulary knowledge.

Spanish vocabulary measures were included in order to objectively classify the L2ers into their respective ability groups, rather than relying on self-rating. Scores for both measures were converted to percentages, which were then averaged to create an overall proficiency score. Participants scoring above 75% were classified as high, those scoring between 55% and 75% were classified as mid, and those scoring 55% or below were classified as low. The proficiency ranges were selected by the researcher, in consultation with participant scores, to be ample but evenly distributed (approximately 20% increments). Scoring ranges were chosen in lieu of the commonly used median split

(Bialystok et al. 2008; Luo et al. 2010), the specific cut-off point of which can shift depending on the participant pool.

The two receptive vocabulary measures included a picture vocabulary test and a lexical judgment task, selected to measure vocabulary knowledge of words at a range of frequencies and because both had corresponding versions in Spanish and English. The two receptive vocabulary measures were chosen to complement each other, by including both images and words, as well as together representing a broader range of vocabulary than each did individually. The picture vocabulary tests, adapted from the MiNT naming test (Gollan et al. 2012), focus on visual identification of high- and medium-frequency words. The lexical judgment tests (Izura, Cuetos, & Brysbaert 2014; Lemhöfer and Broersma 2012) were developed to distinguish vocabulary knowledge at more advanced levels. Because all three participant groups are highly fluent in English, the additional vocabulary measure was chosen in the hopes of preventing or reducing the ceiling effect.

3.3.2.1. Picture vocabulary test

The first vocabulary measure was a picture vocabulary test, a receptive adaptation to the Multilingual Naming Test (MiNT; Gollan et al. 2012). The MiNT test consists of 67 line drawings presented in order of increasing difficulty, with participants instructed to name them as quickly as possible. The MiNT was chosen as the basis for the picture vocabulary adaptation because it was designed with parallel versions in English and Spanish (as well as Mandarin and Hebrew). It was designed to exclude cognates in any of the languages, control for difficulty and variability across languages, and include a higher proportion of words of intermediate difficulty, excluding words of very low frequency commonly found in other naming tests. The authors posit that "sensitivity to bilingual naming skills might be better with a slightly easier test… especially important for assessing naming ability in a non-dominant language" (Gollan et al. 2012, p.598), which was considered appropriate for the current study, because L2er participants completed the task in Spanish as well as English.

With permission from the author, the English and Spanish versions of the test were adapted to measure receptive rather than productive vocabulary. In the receptive adapted version, now the Multilingual Picture Vocabulary Test (MPVT), 67 line drawings are accompanied by multiple word options (Appendix A): the a.) target word, b.) semantic distractor, c.) phonemic distractor and d.) neutral control distractor, e.g. a drawing of a hand is accompanied by the options (English version/*Spanish version*) a.) hand/*mano*, b.) ear/*oreja*, c.) house/*manta* (blanket), d.) cup/*taza*.

Distractors were chosen taking into account the relative frequency of the target word and maintaining high levels of concreteness and imageability, in consultation with the MRC Psycholinguistic Database, a computer usable resource for words in English (see Coltheart 1981), and EsPal, an internet-accessible repository of Spanish word properties (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013). The online test presented each picture in the original MiNT order, with one picture appearing on the screen at a time, along with its accompanying multiple choice options. Before beginning the task, participants were instructed to select the word corresponding to the picture and to skip a question rather than guessing.

3.3.2.2. Lexical judgment task

The second vocabulary measure was a lexical judgment task, in which the participant is presented with a word-like string of letters and must indicate whether the string of letters is a word or not. It was selected because similar versions exist in both English and Spanish, and it provides a broader range of vocabulary than the picture vocabulary test, with more emphasis on low-frequency words than the MiNT. Because the participants in the current study were native speakers of English, a more difficult task was included in addition to the picture vocabulary test to reduce the probability of all participants testing at ceiling. The online survey presented the instructions as given in the original form, followed by each word in the original order, with one letter string and the yes/no option appearing on the screen at a time.

The English version of the lexical judgment task is called LexTALE (Lemhöfer and Broersma 2012; see www.lextale.com). It was developed to study vocabulary knowledge at the medium to highly advanced proficiency level. LexTALE was designed to be presented either by paper or digitally. Each trial consists of a string of letters following English lexical rules, and for each string, the participant is to indicate whether it is an existing word in English or not. The test consists of 60 total items with a 2:1 ratio of words to nonwords: 40 words, 20 nonwords. Before beginning the task, participants were presented with instructions per Lemhöfer and Broersma (2012).

The Spanish counterpart to LexTALE is the Lextale-Esp (Izura, Cuetos, & Brysbaert 2014). It was developed patterning off the LexTALE model, and was also designed to be delivered by paper or digitally. Differences include an increase to the total number of items from 60 to 90, to increase the reliability and range of word-frequency, though the 2:1 proportion of words to nonwords was maintained: 60 words, 30 nonwords. Also, to discourage a yes-bias, the instructions include an express warning that errors are penalized. This warning does not appear in the original English version of the task (Lemhöfer & Broersma 2012), because Izura et al. elected to modify the scoring procedure (described below). Before beginning the task, participants were presented with instructions partially adapted from Izura et al. (2014), to align with the digital format of the test. The original instructions were specific to a paper and pencil presentation of the test, in which items are presented alongside a column to checkmark the words (leaving nonwords blank), so phrases like "tick the box" and "adding tallies" were removed.

3.3.3. Spanish grammar measure

After completing the English and Spanish vocabulary tasks, L2ers additionally completed a short measure of Spanish grammatical knowledge. Verbal fluency studies with HBs have either relied on self-ratings of proficiency (e.g. Gollan et al. 2003, Sandoval et al. 2010) or used vocabulary size as an indicator of proficiency (Bialystok et al. 2008; Luo et al. 2010). Of the few studies on L2 learners, self-rating is also a common measure of proficiency (Snodgrass & Tsivkin 1995; Baus et al. 2013), while Linck et al. (2009) additionally compared accuracy and latencies in a translation-recognition task, one of their experimental tasks, as a measure of proficiency. Given that the current research study did not compare groups of similar proficiency, but rather divided participants into leveled groups, a Spanish grammar measure was included to complement and confirm the vocabulary knowledge score that was the basis of the group classification.

The grammatical proficiency test was excerpted from the Diploma de Español como Lengua Extranjera (DELE) (also used in Sagarra & Ellis 2013; Seibert Hanson & Carlson 2014). Participants completed a portion of the DELE focused on grammar, which consisted of three blocks (basic, intermediate, and advanced) of multiple choice sentence completion questions, 12 questions per block. The online test presented the sentence and its multiple choice options in the original order. Questions appeared on the screen one block at a time. Before beginning the proficiency portion of the test, participants were instructed to complete the questions using their knowledge or intuition, but without consulting outside sources or guessing wildly.

3.3.4. Experimental tasks: verbal fluency

The experimental task in this dissertation is the verbal fluency task. Verbal fluency tasks direct participants to produce as many words as possible for a given cue in a determined amount of time—most often one minute (however, see Baus et al. 2013; Linck et al. 2009; Roberts & Le Dorze 1997 for examples of other time limits). To be counted correct, words must conform to a specific category. The most commonly used category types are semantic[2] and letter[3], the latter of which this dissertation will refer to as letter/phonemic, as explained below. For semantic fluency trials, participants produce words belonging to a specific category of concrete nouns, like Animals or Fruits. For

[2] Sometimes called category (Luo, Luk & Bialystok 2010) or categorical (Snodgrass & Tsivkin 1995)

[3] Sometimes called formal (Schmid & Köpke 2009), phonemic (Rosselli et al. 2002), phonologic (Butman, Allegri, Harris & Drake 2000), phonetic (Portocarrero, Burright & Donovick 2007), alphabetic (Snodgrass & Tsivkin 1995) or lexical (Piatt, Fields, Paolo, Koller, & Tröster 1999).

letter/phonemic fluency trials, participants are directed to produce words beginning with a given letter/phoneme, most commonly the set of F, A, and S.

Verbal fluency tasks measure lexical retrieval, specifically the speed with which an individual can access their stored lexicon, retrieve and produce exemplars in a short period of time based on content or sound association. Like other measures of lexical retrieval, such as picture naming and translation, greater speed and higher total production are interpreted as measures of the ease with which the mind retrieves lexical information for production. It differs from other measures of lexical retrieval in that it is relatively unconstrained. Participants are not bound to produce the specific exemplar presented to them, but rather have a greater degree of freedom to produce the words that come to mind.

In the current study, the verbal fluency tasks consisted of three semantic categories and three letter/phonemic categories. The same tasks were assigned in both English and Spanish. Category order was randomized so that half the participants completed the semantic categories first, and the other half the letter/phonemic categories, though the order of presentation within the category itself was constant. For the L2ers, the language order was also randomized, with half the participants completing the set of six English tasks first and the other half the set of Spanish tasks first. Thus, monolingual participants completed a total of six trials in one of two task orders, and L2 participants completed a total of twelve trials in one of four task orders.

Before beginning the verbal fluency task, subjects were told they would be doing a series of 1-minute recordings in which they should say as many words as possible according to each cue they were given. They were told to expect two different types of category, one based on objects (e.g. vegetables) and the other based on word properties, in which they would be given a letter and should produce words starting with that letter or sound (e.g. "words starting with B or /b/"). To minimize anxiety and prevent participants, especially the LVL2ers, from giving up partway through each minute trial, subjects were informed that it is common (though not guaranteed) to run out of words before the end of the minute, but that the minutes would not be cut short, and to keep thinking the entire time in case another word came to mind towards the end of the trial. After subjects had any questions clarified and indicated they understood, they were digitally audio recorded, and their responses were later transcribed.

3.3.4.1. Semantic verbal fluency

The semantic fluency cues used in this dissertation were Animals, Fruits, and Clothes. Animals is the most common cue used in the literature, often as the only semantic trial (Bialystok et al. 2008; Butman et al. 2000; González et al. 2005; Rosselli et al. 2002; Tombaugh et al. 1999). Two additional cues were chosen to allow for an average word total that was less dependent on the performance in any one single category and to equal the FAS set of letter/phonemic trials. The specific semantic cues were chosen to be as culturally neutral as possible, to avoid soliciting a category of vocabulary more common to one culture than another and therefore biasing production in favor of the language in which those ties were created (Roberts and Le Dorze 1997). Cues were also selected that were broad enough to represent both superordinate and subordinate categories. For instance, the superordinate Animals can be organized into subordinate groups of pets, farm animals, forest animals, zoo animals, etc. Fruits can be subcategorized into berries, stone fruits, tropical fruits, etc. And clothes can be grouped by season (e.g. winter, summer), occasion (party, beach), etc. Though grouping strategies and schematic organization do not form part of the research questions in this dissertation and are therefore not discussed in the results, categories with this type of organizational potential were chosen to allow for potential future schematic analysis of the current data.

3.3.4.2. Letter/phonemic verbal fluency

The letter/phoneme prompts for this study were F, A, and S. Many studies have used a variety of letters/phonemes, but FAS is the most common set (Tombaugh et al. 1999), and therefore aligns with the largest percentage of the previous literature. Though languages may have different rank frequencies for different initial letters/phonemes, previous work across languages (Rosselli et al. 2000, 2002) did not find differences between English and Spanish monolinguals in the standard FAS prompts. The standard instructions for the letter fluency task, "name as many words as you can that start with the letter…," were modified to minimize emphasis on orthography as much as possible. Subjects were explicitly directed before beginning the verbal fluency tasks that they could name words beginning with the given letter or the sound represented by the letter.

3.3.4.2.1. A note on terminology

By far the commonest designator of the letter/phonemic fluency is "letter," which is representative of the standard instructions to produce words starting with a given letter. This title and the corresponding instructions may be linguistically inappropriate, as they confound phonologic and orthographic information. This tension can be seen in the wide array of other terms sometimes substituted for “letter” fluency, including alphabetic (Snodgrass & Tsivkin 1995), formal (Schmid & Köpke 2009), phonemic (Rosselli et al. 2002), phonologic (Butman et al. 2000), phonetic (Portocarrero et al. 2007), or lexical (Piatt et al. 1999a,b), with some terms emphasizing letter features and other phonemic features.

Use of the term “letter” is problematic for multiple reasons. While the title “semantic” in semantic fluency points to the semantic networks activated by the prompt, the use of the word “letter” in letter fluency assumes that the letter, an orthographic rather than linguistic feature, is the organizing principle upon which participants rely to retrieve exemplars. Referencing letters also assumes a certain amount of literacy and metalinguistic knowledge on the part of the participant. While not an issue for the literate, university educated, participants in the current study, it is the case that letter fluency is more sensitive to education levels than semantic fluency (Tombaugh et al. 1999), which further indicates that literacy may be a confounding variable in this task as it is commonly administered.

Determining error in this category is also difficult. Guidelines for verbal fluency error classification include “intrusions (a word belonging to another category is offered (e.g., when naming fruits, the subject names vegetables)” (Ardila et al. 1994, p.49). However, this description is problematic, even if the example is amended to reflect a letter/phonemic category instead of semantic. Given the cue S, if the subject names a word starting with B, it is clearly an error. However, it is more challenging to determine how to interpret exemplars that conform phonetically but not orthographically to the given cue, e.g. cigarette. Alternatively, words can conform orthographically, but not phonetically, e.g. shirt. The title “letter” fluency suggests that ‘cigarette’ would be considered an error, but ‘shirt’ would be counted as correct, while the use of other titles like “phonologic” or “phonetic” suggests a definition of “correct” as those words beginning with the given initial phoneme, resulting in the opposite of the above, i.e. counting ‘cigarette’ but excluding ‘shirt.’ Studies typically do not formalize or explicitly explain whether these types of crossovers are considered errors, and in the rare case that it is mentioned, the name given to the category can conflict with the formalized rules, as in referring to the category as “phonetic fluency” but specifying that ‘phone’ would be considered a phonetic intrusion error for an F trial (Portocarrero et al. 2007).

Given the above concerns, the current study chose to represent the category under the title “letter/phonemic” verbal fluency. The addition, rather than substitution, of phonemic to letter serves two purposes. It maintains clarity in the review of previous literature, which most commonly uses “letter.” It also represents the amended instructions given the participants in this study, who were directed to produce words “beginning with that letter or that sound.” The addition of “sound” to the instructions clarifies that examples like the above “phone” for F are considered correct, maximizing flexibility within the category constraint. A focus on sound alone was deemed problematic in part due to the variability in pronunciation of one of the trials, A, in English. Word initial ‘a’ can be pronounced /a/ in “at,” /ae/ in “ate,” and /au/ in “art,” in which case a directive based on initial phoneme would be more restrictive than letter.

3.4. Setting

The data collection setting varied according to the convenience of the participants. The online survey was completed at the discretion of the participant, either at home or at a computer lab. Participants were directed to complete the survey on their own and not to consult outside sources, including dictionaries, books, the internet or other people. The verbal fluency recordings were either collected in a university language lab using Sony Soloist® Digital PC Comparative Recorder and Virtuoso™ Instructional Control softwares, or in individual meetings with participants, using a Sony digital recorder.

3.5. Data coding and scoring

This section describes the data coding procedures used in the current study. First, I detail the scoring procedures for the various non-experimental tasks participants completed, including measures of English and Spanish vocabulary knowledge and Spanish grammatical proficiency. Then I describe the transcription and coding of the verbal fluency trials. The verbal fluency section is further divided into the scoring of mean word totals, mean word frequency, mean retrieval latency, and time-course of retrieval.

3.5.1. Vocabulary measures

Four tasks made up the vocabulary measures in this study: picture vocabulary tests in English and Spanish, and lexical judgment tasks in English and Spanish. The picture vocabulary tests were versions of the same task and contained the same number of items. The Spanish version of lexical judgment task, though based on the English version, contained more items. Scores were converted to percentages to allow for comparison.

3.5.1.1. Picture vocabulary test

The English and Spanish versions of the Multilingual Picture Vocabulary Test consisted of 67 line drawings accompanied by four multiple choice options. Participants were assigned one point for each correct answer and zero points for incorrect answers or for not selecting an answer. Therefore, the maximum possible total score was 67 and minimum possible score was 0. The raw total was then divided by 67 to convert it to a percentage.

3.5.1.2. Lexical judgment task

Lemhöfer and Broersma (2014) and Izura et al. (2014) detail different scoring procedures for the LexTALE and Lextale-Esp, respectively. Given that both the LexTALE and the Lextale-Esp were given to the participants of the current study, it would not be possible to compare scores across languages if different scoring methods were used. The choice was made, therefore, to select one of the methods and apply it to both tests for reporting purposes.

Lemhöfer and Broersma (2012) compared several methods to score the LexTALE, and selected the method that correlated the highest with a separate translation task, which they called *%correct*$_{av}$. The scoring instructions (see www.lextale.com/scoring.php), involve assigning one point to correct answers and zero points to incorrect and skipped items. The items are then grouped into words and nonwords, and the percent correct of words and nonwords (individually) is averaged, as follows:

((number of words correct/40*100) + (number of nonwords correct/20*100) / 2

This method reports scores as a percentage of 100. It corrects for the fact that two thirds of the items are words and one third are nonwords. It does not additionally penalize guessing and/or a yes-bias, the tendency to incorrectly judge nonwords as words.

Izura et al. (2014) based the creation of the Lextale-Esp on the LexTALE but designed a different scoring method, one not included in Lemhöfer and Broersma (2012)'s comparisons. Izura et al. assign one point for every correct response to a word or nonword, zero points to an incorrect response to a word, and two points for every incorrect response to a nonword. The score for incorrect responses to nonwords is then subtracted from the score for correct responses to words, as follows:

$N_{\text{yes to words}} - 2{*}N_{\text{yes to nonwords}}$

This method reports a raw score, with a maximum possible of 60 (representing one point each for correctly identifying all words and not losing any points for incorrectly identifying nonwords as words). It separates the calculation of words from nonwords, as does the $\%correct_{av}$, but it doubly penalizes guessing and/or a yes-bias. Random guessing would result in a score around 0. With this method, it is also possible to score a negative number.

To allow for comparison between the LexTALE and the Lextale-Esp among the current study's participants, Izura et al. (2014)'s scoring method was not used; instead, the original $\%\ correct_{av}$ from Lemhöfer and Broersma (2012) was adapted to account for the increased number of tokens:

((number of words correct/60*100) + (number of nonwords correct/30*100) / 2

By using the *%correct*$_{av}$ for both tests, both scores are reported as a proportion of 100 and neither penalizes one type of incorrect response more than another. Therefore, participant scores across languages can be compared fairly.

3.5.2. Spanish grammar measure

The selected portion of the DELE consisted of 36 multiple choice sentence completion questions. Participants were assigned one point for each correct answer and zero points for incorrect answers or for not selecting an answer. Therefore, the maximum possible total score was 36 and minimum possible score was 0. The raw total was then divided by 36 to convert it to a percentage, in line with the vocabulary measures.

3.5.3. Verbal fluency

The following section details the coding and scoring of the experimental verbal fluency tasks, including descriptions of how errors were determined. The verbal fluency tasks included three semantic and three letter/phonemic trials, completed in English and Spanish (the latter for L2ers only, results not presented). Participants' production was digitally recorded and later transcribed and coded. The recordings were first processed to transcribe all responses, during which incorrect responses were identified.

3.5.3.1. Word totals and errors

Word totals were calculated for each individual trial (Animals, Fruits, Clothes, F, A, S) by assigning one point for each correct exemplar and counting the number of correct responses produced within each one-minute trial. Overall semantic and letter/phonemic fluency was calculated as well by averaging individual trials across semantic (Animals, Fruits, Clothes) and letter/phonemic (FAS) categories.

3.5.3.1.1. Semantic verbal fluency

In semantic categories, participant responses were transcribed in their neutral singular form. One point was given for each correct exemplar. Both superordinate (e.g. pants) and subordinate (e.g. jeans, slacks, etc.) words were counted, even if both were produced in the same trial. This was done because superordinate labels may carry distinct conceptual information relative to subordinate examples. For instance, though “jeans” and “slacks” are types of “pants,” “pants” perhaps conjure up a different image than either of the others. Participants were therefore credited for all responses belonging to a category.

A word was considered an error if it fit into one of the following categories: (1) repetitions (producing the same word multiple times during the same trial); (2) intrusions (naming a word not belonging to the given category, e.g. sequoia for the Fruits trial); (3) nonwords; (4) cross-language insertions (producing an exemplar not in the target language of the trial, for L2ers only). Responses that were not considered errors, as mentioned above, included producing both a superordinate and subordinate response during a single trial.

3.5.3.1.2. Letter/phonemic verbal fluency

In letter/phonemic categories, participant responses were transcribed in their neutral singular form. One point was given for each correct exemplar, which included any word beginning with the correct letter or phoneme. So called "derivatives," words that are morphologically related or with the same lexical root (see below) were counted as correct if the words came from different word classes or otherwise clearly represented distinct concepts.

Homophones present a particular challenge to the transcription of words in this category. Context was used as much as possible to aid in determining which of several possible words had been produced, e.g. in the series "forever, /fɔr/, forego," in which the middle word is ambiguous, context suggests that the intended word is "for" and not "four." In the case of remaining ambiguity, the higher frequency word was chosen for transcription.

Error determination in letter/phonemic verbal fluency is somewhat more complex than in semantic fluency. For the purposes of the current project, a word was considered an error if it fit into one of the following categories: (1) repetitions (producing the same word multiple times during the same trial, including derivatives of tense and number, e.g. sneeze, sneezed or sock, socks; see below for details); (2) intrusions (naming a word not beginning with the given letter or phoneme); (3) nonwords; (4) cross-language insertions (producing an exemplar not in the target language of the trial, for L2ers only).

Repetitions may be difficult to distinguish from homophones and homonyms. For example, a subject may repeat the word /fɔr/ at two different times during an F trial, leaving open the possibility that they meant "for" the first time and "four," or even "fore" the second time, or simply that they forgot they had already said the word "for." Given that participants were not debriefed following the recording, and therefore clarification questions were not possible, such instances were counted as repetitions unless the subject clarified during the recording itself, e.g. “so… the other /so/, s-e-w.”

Regarding derivations, the justification for exclusion is easy to understand, lest a participant artificially increase his word count by focusing on one word. The issue of clearly delineating “derivatives” is problematic, however, at least based on the type of example provided in guidelines to error classification: “the subject produces a word and then begins to say other words with the same lexical root—e.g. sun, sunny” (Ardila et al. 1994, p.49). If the intent of general verbal fluency tasks is to count the number of concepts retrieved and produced according to the criteria, “sun, sunny” is, at the least, activating distinct syntactic features and therefore distinct lemmas (Levelt et al. 1999), and may also be activating distinct concepts (the sun itself versus a sunny day or even a sunny personality). A clearer example of a derivative would be inflections of number or tense, e.g. “swim, swims, swimming,” since they are generated from a singular lemma (Levelt et al. 1999); derivations of this type were counted as repetition errors.

Responses that were not considered errors included: (1) words that conformed phonetically to the given category, even if not orthographically, and vice versa (e.g. ‘cigarette’ or ‘shirt’ in an S trial) (2) compounds and words with bound morphemes that changed the word’s syntactic category (“sun, sunny” or “swim, swimmer, swimming pool”); (3) proper names.

3.5.3.2. Word frequency

As described in the literature review, Sandoval et al. (2010) measured mean word frequency in addition to mean word totals, in order to distinguish between two theories of retrieval slowing: competition and weaker links. If lifetime reduced exposure to linguistic input results in weaker links, low-frequency words should be disproportionately impacted, making them less likely to be produced, resulting in an overall higher mean frequency for bilinguals than monolinguals. Conversely, the authors argue that high-frequency words, being used more often and recently, would experience a higher level of activation and thus would provide stronger competition for their cross-language translation equivalents. Therefore, lower-frequency words would be more likely to be produced, resulting in an overall lower mean frequency for bilinguals than monolinguals.

Following Sandoval et al. (2010), the CELEX (Baayen et al. 1995) word frequency of correct words was obtained using N-watch, aka NeighborWatch (Davis, 2005), a Windows program for deriving neighborhood size and other psycholinguistic properties of both words and nonwords. As Davis explains, the CELEX English linguistic database "is derived from the COBUILD corpus of 17.9 million words, 16.6 million of which were sampled from written sources… the remaining 1.3 million being sampled from spoken English" (p.67). Frequency counts can be obtained from only the written, only the spoken, or from the combined sources, the latter option being the case in the current study. The output describes the occurrence of a given word per million words; the larger the number, the higher frequency the word. After the occurrence per million was determined for each word produced, it was averaged across cues and categories to create the mean frequency scores.

3.5.3 3. Acoustic measures

Beyond word information like total words produced and mean word frequency, analysis of acoustic measures allows for further calculations including mean retrieval latencies and time-course of retrieval. Acoustic data, like word frequency, can be used to differentiate between results in support of various theories for bilingual verbal fluency performance. The following section will describe how the acoustic time-stamp data was coded, followed by a description of the measures it provides: retrieval latency and graphic time-course of retrieval.

3.5.3.3.1. Time-stamping procedure

After initial transcription, the associated time-stamp for the beginning of each trial and each correct word was identified using Audacity® on Windows. Luo et al. (2010) identified word onset by observing the sound wave form, using the mouse to click the onset point on the screen, then recording the timestamp displayed (Luo, personal correspondence, December 3, 2015). In an attempt to increase the precision, accuracy, and efficiency of the process for the current study, time-stamps were initially placed using the 'sound finder' feature, which places a 'region label' (a tag that encompasses the sound) around defined periods of sound. Minimum silence duration was set to a tenth of a second, and labels started and ended one hundredth of a second before and after the identified sound.

Following the 'sound finder' procedure, labels of false positives (extraneous vocalizations and noises identified as sound, e.g. 'ummm,' finger tapping, throat clearing) were deleted, while labels of words were edited to match the word given. The Audacity®

program is somewhat limited regarding the detection of initial fricatives [f] and sibilants [s], so each label boundary was double-checked to ensure it contained the onset of production and was adjusted as necessary. The time-stamp of each label was recorded as the time (in minutes, seconds, and milliseconds) elapsed since the beginning of the recording to the starting point of the region label.

The time-stamp was used in calculating the retrieval latencies as well as to group responses into 5-s bins, the twelve 5-s periods over the course of the 1-m trial. Bin numbers were used to graph the time-course of retrieval. Overall, the following codes were assigned for each correct response: (1) serial number, indicating the order or responses in each trial; (2) total latency; (3) first-response latency (FRL) ; (4) subsequent-response latency (SRL); (5) bin number, indicating the 5-s period during which the response was given, according to the total latency.

3.5.3.3.2. Mean retrieval latency

Typically, verbal fluency production follows an exponentially declining curve (Wixted & Rohrer 1994), wherein subjects produce exemplars at the greatest rate at the beginning of the trial and slowly taper off as time passes. Retrieval latency (Rohrer et al. 1995) measures the time until articulation of each response, beginning from the onset of recall. Total response latency (the average time between the beginning of the trial and each response) can be broken into two important, but distinct, measures of retrieval latency: FRLs and mean SRLs (Rohrer et al., 1995).

FRLs measure the time between the beginning of the trial and the first response. In this study, FRLs were calculated by subtracting the time-stamp of the beginning of the trial from the time-stamp of the first correct response. The beginning of each trial was marked as the endpoint of the initiation instructions, e.g. the end of the 'n' in "You may now begin." FRLs are separated from SRLs because they represent "the processing of the retrieval cue and the initiation of the search process," while the mean SRL "provides a more accurate depiction of the retrieval process" (Rohrer et al. 1995, pg.1131).

SRLs measure the time between the first response and each remaining response. For example, in an abridged trial of 5 responses occurring at 1, 3, 5, 8, and 16 s, the FRL equals 1 s. The SRLs are 2, 4, 7, and 15 s respectively (subtracting 1 s from each subsequent response), which result in a mean SRL of 7 s. In this study, mean SRLs were calculated by subtracting the time-stamp of the first correct response from the time-stamp of each subsequent response, and then averaging. Mean SRLs are an indicator of the spread of production over time. Shorter mean SRLs indicate that responses were clustered at the beginning of the time period with fewer at the end (e.g. 5, 6, 7, 8, 19; mean SRL: 5 s). Longer mean SRLs indicate that responses were more evenly spread throughout the time period (e.g. 5, 10, 15, 20, 25; mean SRL: 12.5 s), which can arguably be due to retrieval slowing or a slower rate of decline through the trial (Luo et al. 2010; Sandoval et al. 2010).

3.5.3.3.3. Time-course of retrieval

The time-course of word retrieval, similar to mean retrieval latency, is a measure of the declining rate of production. It plots the average number of responses over time on a graph, with number of responses on the *y*-axis and time (60 seconds, divided into 5-s increments) on the *x*-axis. Plotting the time-course allows for a visual examination of production over time, making evident the exponential decline curve.

In the current study, time-stamps were used to group correct responses (both first and subsequent) into 5-s bins over every 1-m trial, according to the calculated total retrieval latency. The decision to include FRLs was made so that the 60-s graph represented the 60-s trial for everyone. For example, if participant A starts producing exemplars right at the beginning of the trial, she will have nearly all 60 seconds to produce exemplars. If person B pauses and thinks for 5-10 seconds, by the time he gets started, he actually only has the remaining 50-55 seconds to produce exemplars. Grouping only SRLs into bins (Rohrer et al. 1995) ignores the initiation process of retrieval, information which is excluded from the graph. Additionally, the tail end of the 60 seconds may be artificially low, because participant B's data has been left-shifted.

Two main elements of the time-course can be used to assess the results. The line of best fit provides the intercept (which Luo et al. 2010 refer to as the initiation, given that no production is occurring at time (t) = 0, and therefore the line cannot ever truly reach the *y*-axis) and the slope of production. Luo et al. (2010) and Friesen et al. (2015) argue that the intercept (initiation) reflects differences in target language knowledge, while the slope reflects task control.

3.6. Summary

In summary, 122 English monolinguals (n=32) and English-Spanish L2 learners/bilinguals (n=90; 28 low, 32 mid, 30 high) participated in the current study. They provided biographic information and were tested on their linguistic knowledge in the L1 (and additionally in the L2 for the L2ers). They subsequently completed six English verbal fluency trials (as well as six in Spanish for the L2ers, language order counterbalanced, results not reported here) to examine their L1 lexical retrieval process. Scores on Spanish lexical and grammatical knowledge were tallied and used to identify three levels of ability within the L2 participants. English verbal fluency performance was coded for word totals and word frequency. In addition, acoustic analyses identified the time-stamp of each word produced, resulting in measures of retrieval latency and time-course of retrieval. The next chapter will compare the results of the measures described above across participant groups to address the research questions posted in this study.

Chapter 4: Results

4.1. Introduction

This chapter will present the results and statistical analysis of the English vocabulary and verbal fluency tasks. The information is organized as follows: First I will present the results of the English vocabulary measures, to examine the hypothesis that L2 learners/bilinguals' L2 vocabulary knowledge does not result in decreased L1 vocabulary knowledge, unlike HBs, who tend to have smaller vocabularies in each of their languages than a respective monolingual (Bialystok et al. 2008; Bialystok & Feng 2009; Portocarrero et al. 2007). The rest of the chapter will describe the English verbal fluency performance in detail, with a subsection each of word totals, word frequency and acoustic measures of retrieval latency and time-course of retrieval. Word totals are the primary basis for interpretation of verbal fluency, with more words produced being a sign of speedier and easier lexical access as well as greater executive control (Shao et al. 2014). Further measures add nuance to the interpretation of word total performance, allowing for distinctions between varying hypotheses, including retrieval slowing due to competition or weaker links, vocabulary deficits, and executive control advantages (Friesen et al. 2015; Luo et al. 2010; Sandoval et al. 2010).

Each subsection contains descriptive data followed by statistical analyses comparing performance across groups. Within each subsection, the verbal fluency performance is analyzed in terms of 1) aggregated category types: semantic and letter/phonemic; 2) disaggregated trials: Animals, Fruits, Clothes, F, A, S; and 3) comparing performance divided by language testing order: English first versus Spanish first. Recall that Spanish verbal fluency data was collected to look for evidence of cross-linguistic inhibition, specifically whether activation of the L2 by testing Spanish first had an impact on L1 performance. However, only the verbal fluency data produced in the English trials is presented in this dissertation, in alignment with the research questions presented earlier which aim to investigate the impact of a formally acquired L2 on L1 lexical retrieval.

4.2. English vocabulary measures

The following section will present the results of the English vocabulary measures, including the English picture vocabulary test (MPVT-Eng) and LexTALE lexical judgment task (Lemhöfer & Broersma 2012), beginning with descriptive means and followed with statistical analyses.

4.2.1. English vocabulary descriptive data

Table 4.1. presents the percentage means and standard deviations of both the English picture vocabulary test (MPVT-Eng) and LexTALE lexical judgment task. Recall that reduced vocabulary has been presented as a hypothesis to account for verbal fluency disadvantages seen in HBs (Bialystok et al. 2008; Luo et al. 2010; Sandoval et al. 2010), given that they often do not know as many words in each of their languages as a corresponding monolingual (e.g. Bialystok 2006; Bialystok et al. 2008), though they often report dominance in the L2. Receptive vocabulary knowledge was tested to examine whether L2ers demonstrate reduced vocabulary knowledge even in the dominant L1. Scores equivalent to monolinguals would suggest that verbal fluency performance,

whether higher or lower than monolinguals, could not be attributed to differences in vocabulary knowledge.

Table 4.1. English vocabulary scores (%)

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
MPVT-Eng	98.97	1.28	98.40	2.07	99.30	0.85	98.86	2.37
LexTALE	91.37	7.73	92.50	7.885	92.23	9.91	94.46	8.72

Comparing vocabulary measures, we can see that participants in all four groups scored higher on the MPVT-Eng, near ceiling. This is unsurprising given that the MiNT stimuli (Gollan et al. 2012) from which it was developed target high- and medium-frequency words, which are unlikely to be challenging for native English speakers. Scores on the LexTALE, which includes lower-frequency vocabulary, were slightly lower. Across participant groups, there is little observable difference in score on the MPVT-Eng between the monolinguals and the various L2er groups. Interestingly, rather than a trend of decreasing score with increasing L2 proficiency, the HVL2ers score slightly higher on the LexTALE than the other groups. Possible accounts for this will be briefly considered in the discussion chapter, in terms of education differences and cross-linguistic influence.

4.2.2. English vocabulary statistical analyses

To determine whether any significant differences existed across groups in the two English vocabulary measures, statistical analyses were conducted. The mean correct responses were analyzed in a mixed ANOVA with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and task (MPVT-Eng, LexTALE) as the within-subjects factor. The main effect of task was significant, $F(1, 118) = 63.677$, $p < .001$, partial $\eta^2 = .350$, in that participants scored higher on the MPVT-Eng than the LexTALE. As expected, there was no significant main effect of group, $F(3,118) = .632$, $p = .956$, partial $\eta^2 = .016$, or interaction between group and task, $F\ (3, 118) = .836$, $p = .477$, partial $\eta^2 = .021$. Thus, there was no significant difference in score for either vocabulary measure based on bilingual status or proficiency.

4.2.3. Effects of other variables on English vocabulary knowledge

A multiple regression was run to investigate whether some other variable besides group impacted performance on the LexTALE, as the more sensitive test with less evidence of ceiling effects. Variables included age, sex, and highest level of education, none of which significantly predict LexTALE performance, $F(3, 118) = 1.472$, $p = .226$, $R^2 = .036$.

4.2.4. English vocabulary summary

In summary, native English-speaking L2 learners/bilinguals of Spanish do not appear to differ from English monolinguals, according to the given measures, in English vocabulary knowledge. It is therefore unlikely that reduced L1 vocabulary knowledge disadvantages L2ers as it is theorized to do in HB populations (Gollan et al. 2008; Luo et al. 2010; for counterevidence, see Sandoval et al. 2010). Given that all four groups performed near ceiling, it is possible that the two measures—whose target items were originally selected to measure heritage and/or second language vocabulary knowledge—were too easy for native English speakers and therefore failed to discern more subtle differences in their knowledge of obscure, very low-frequency words. However, given the (nonsignificant) numerical trend of *increasing* LexTALE score for the highly proficient bilingual group, it is unlikely that a more sensitive English vocabulary measure would find decreased vocabulary knowledge corresponding to increasing L2 proficiency.

The next section presents the data from the experimental tasks, measures of verbal fluency. It begins with word totals and progresses to more fine-grained analysis, including word frequency and retrieval latency, with subsections for first-response (FRL) and mean subsequent-response latency (SRL), and the time course of retrieval.

4.3. Verbal fluency measures

Verbal fluency tasks are used as neuropsychological measures of cognitive functioning (Bialystok et al. 2008) and are typically divided into two category types: semantic and letter/phonemic. By requiring word production from memory over a span of time, they gauge both lexical access and executive control (Shao et al. 2014), with the relative contribution hypothesized to be stronger for the former in the case of semantic and the latter in the case of letter/phonemic. The current section presents the results from the verbal fluency trials, three each in semantic (Animals, Fruits, Clothes) and letter/phonemic (F, A, S). The three subsections of analysis include the following: word totals, including error analyses (4.2.1.), word frequency (4.2.2.) and retrieval latency (4.2.3.), which is further subdivided into sections for first-response latency (FRL), mean subsequent-response latency (SRL), and the time course of retrieval.

4.3.1. Word totals

4.3.1.1. Word total descriptive data

The 122 subjects who participated in the current study produced a total of 13,356 correct words in English. Table 4.2. reports the means and standard deviations for word totals in semantic and letter/phonemic verbal fluency categories as a whole as well as disaggregated into their respective individual tasks.

Table 4.2. Verbal fluency word totals

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	19.39	3.62	19.51	3.43	20.84	3.50	21.96	4.17
Animals	22.19	4.582	23.63	5.109	24.57	5.118	24.72	4.376
Fruits	15.97	4.721	14.87	4.531	16.89	4.383	17.89	4.549
Clothing	20.03	3.972	21.10	3.827	23.51	4.788	22.78	6.431
Letter/phonemic	14.93	3.57	15.30	4.46	16.56	3.73	17.47	4.26
F	14.88	4.29	14.83	4.08	17.41	4.19	19.28	4.52
A	13.69	5.10	12.87	5.57	14.35	5.20	17.83	4.97
S	16.66	4.36	17.87	5.67	19.24	4.02	19.28	4.84

Looking first across categories, monolinguals produced more words in semantic than letter/phonemic categories, as did each of the L2er groups. Among the semantic categories, the most productive task for each group was Animals followed by Clothing, and finally Fruits. Among the letter/phonemic categories, the most productive task for all but the HVL2ers was the letter/phoneme S, followed by F, and finally A. The HVL2ers produced equally well in S and F, while A was less productive. Comparing individual semantic and individual letter/phonemic categories, it can be seen that at least one of the letter categories, S, frequently resulted in greater production than the least productive semantic Fruits category (Azuma et al. 1997), while F was occasionally slightly more productive than Fruits.

Turning to an inspection across groups, the monolinguals and LVL2ers appear to perform quite similarly in the mean semantic and letter/phonemic categories. Among the categories, LVL2ers produced fewer words on average in Fruits and A. The HVL2ers produced more words than the other three groups, with the MVL2ers performing in between, except in Clothes, in which the MVL2 group produced slightly more words. LVL2ers produced similar amounts of words as the HVL2 group in the Animals, Clothing, and S categories but fewer words in Fruits, F, and A.

4.3.1.2. Word total statistical analyses

4.3.1.2.1. Aggregated word totals

To determine whether the between-group differences were significant, statistical analysis were conducted. The mean correct responses, grouped by category type, were analyzed by a mixed ANOVA with participant group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects variable. The main effect of category type showed a statistically significant difference between semantic and letter/phonemic verbal fluency, $F(1, 118) = 166.368$, $p < .001$, partial $\eta^2 = .585$, in which participants produced more words in semantic trials than letter/phonemic trials. The main effect of group showed a statistically significant difference in verbal fluency performance based on bilingual status, $F(3, 113) = 3.779$, $p = .012$, partial $\eta^2 = .088$. Post hoc pairwise comparisons revealed that HVL2 generated more words than the monolinguals ($p = .02$). No other group differences were significant, though the difference between HVL2 and LVL2 groups approached significance ($p = 0.61$). Finally, there was no statistically significant interaction between category type and bilingual status on verbal fluency performance, $F(3, 118) = .040$, $p = .989$, partial $\eta^2 = .001$.

4.3.1.2.2. Disaggregated word totals

Following Azuma et al.'s (1997) recommendation to analyze disaggregated verbal fluency tasks in addition to overall means by category, an additional mixed ANOVA were conducted with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and trial (Animals, Fruits, Clothes, F, A, S) as the within-subjects factor. The main effect of trial was significant, $F(5, 560) = 132.917$, $p < .001$, partial $\eta^2 = .530$. Post hoc pairwise comparisons ($p < .01$ in all cases, unless specified) found significant differences among each pair of trials except Fruits and F, such that the order of productivity (highest to lowest) was as follows: Animals, Clothes, S, F/Fruits, A. In summary, comparing across category types, while semantic trials were on average more productive than letter trials, there was a bit of crossover in that the Fruits semantic trial was less productive than the letter/phonemic trial S, and equal to the letter/phonemic trial F. The interaction between trial and bilingual status was not significant, $F(15, 590) = 1.473$, $p = .110$ partial $\eta^2 = .036$.

4.3.1.3. Word total language order effects

4.3.1.3.1. Word total language order descriptive data

Language testing order was counterbalanced across the three L2 groups to test the hypothesis that activating the L2 first results in inhibition of the L1, interfering with L1 production and resulting in reduced total words produced. Table 4.3. presents the means and standard deviations for word totals divided by language testing order, i.e. those tested in English first or Spanish first. Results are provided for semantic and letter/phonemic verbal fluency as a whole, as well as disaggregated into their respective individual tasks.

Analyzing across trials, comparing language testing order, it appears that LVL2ers produced on average approximately 2.25 more words when tested in Spanish first. The MVL2ers show the opposite visual trend, producing approximately 2.4 more words when tested in English first. Finally, the HVL2ers produced very similarly regardless of which language was tested first, less than a quarter of a word more when Spanish was presented first.

Table 4.3. Verbal fluency word totals by language testing order

Category	Order[a]	LVL2		MVL2		HVL2	
		M	*SD*	*M*	*SD*	*M*	*SD*
Semantic	English	18.44	3.31	21.33	3.07	22.16	5.01
	Spanish	20.44	3.36	20.21	4.01	21.78	3.29
Animals	English	21.31	5.14	23.71	3.25	24.73	5.12
	Spanish	25.47	5.21	23.50	5.61	24.07	4.03
Fruits	English	14.38	3.86	16.94	4.36	19.07	5.39
	Spanish	14.00	4.34	15.36	3.82	17.73	2.82
Clothing	English	19.62	3.69	23.33	3.79	22.67	6.13
	Spanish	21.87	2.56	21.79	5.73	23.53	4.84
Letter/phonemic	English	14.59	4.84	16.91	3.88	18.04	2.57
	Spanish	15.91	4.18	16.12	3.62	16.89	5.50
F	English	14.61	4.65	17.56	4.29	18.33	2.72
	Spanish	15.93	3.86	15.86	2.57	17.07	6.32
A	English	11.46	5.03	14.00	5.39	16.73	3.62
	Spanish	13.67	6.08	13.79	5.40	15.93	5.55
S	English	17.69	6.01	19.17	4.18	19.07	5.74
	Spanish	18.13	4.82	18.71	5.00	17.67	4.79

[a] The language tested first

4.2.1.3.2. Word total language order statistical analyses

To determine whether the numerical differences according to language testing order resulted in any significant differences, a three-way mixed ANOVA was conducted using proficiency group (LVL2, MVL2, HVL2) and language order (English first vs. Spanish first) as between-subjects factors and category type (semantic, letter/phonemic) as the within-subjects factor. The three-way interaction between category type, proficiency group and language testing order was not significant, $F(2, 84) = .208$, $p = .813$, partial $\eta^2 = .005$, nor were the two-way interactions between category type and language testing order, F (1, 84) = .228, $p = .634$, partial $\eta^2 = .003$ or proficiency group and language testing order, $F(2, 84) = 1.264$, $p = .288$, partial $\eta^2 = .029$, or the main effect of language testing order, $F(2, 84) = .001$, $p = .980$. The same pattern of results was found when analyzing disaggregated trials and will therefore not be reported.

4.3.1.4. Effects of other variables on word totals

A multiple regression was run to assess the impact of various other variables on each of semantic and letter/phonemic verbal fluency. Variables included age, sex, and highest level of education, which did not significantly predict letter/phonemic fluency, $F(3, 118) = 1.274$, $p = .287$, $R^2 = .031$. The set of variables did significantly predict semantic fluency, $F(3, 118) = 3.544$, $p = .017$, $R^2 = .083$; however, the only variable that significantly added to the prediction was sex, $t(3, 118) = 2.287$, $p = .024$ A follow-up three-way mixed ANOVA with group (monolingual, LVL2, MVL2, HVL2) and sex (male, female) as between-subjects factors and semantic trial (animals, fruits, clothing) as the within-subjects factor found no significant three-way interaction, $F(6, 228) = 1.388$, $p = .221$ and no two-way interaction between group and sex, $F(3, 114) = .784$, $p = .784$, partial $\eta^2 = 0.009$. The two-way interaction between trial and sex was significant, $F(2, 228) = 5.111$, $p = .007$, partial $\eta^2 = .043$. There was a statistically significant simple main effect of sex for the Fruits trial, $F(1, 114) = 13.621$, $p < .001$, partial $\eta^2 = .107$, in which females produced more fruit exemplars than males.

4.3.1.5. Error analyses

4.3.1.5.1. Error descriptive data

In addition to the 13,356 correct words, participants produced a total of 429 words discounted as errors, or 3.2% on average. 166 (38.7%) of the errors were produced during semantic trials, while the remaining 263 errors were produced during the course of the letter/phonemic trials. Table 4.4. reports the means and standard deviations for errors in semantic and letter/phonemic verbal fluency categories as a whole as well as disaggregated into their respective individual trials. Semantic and letter/phonemic errors are totals across the three trials, as opposed to averages.

Table 4.4. Verbal fluency errors

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	1.25	1.30	1.18	1.19	1.47	1.59	1.47	2.11
Animals	0.41	0.56	0.29	0.54	0.56	0.80	0.20	0.61
Fruits	0.41	0.67	0.57	0.69	0.53	0.72	1.07	1.62
Clothing	0.44	0.72	0.32	0.55	0.38	0.83	0.20	0.41
Letter/phonemic	2.19	3.92	2.43	2.19	2.09	2.74	1.93	2.43
F	0.59	0.84	0.64	1.62	0.69	1.15	0.70	1.18
A	0.59	1.37	0.82	1.12	0.37	0.49	0.77	1.59
S	1.00	2.70	0.96	1.14	1.03	1.68	0.47	0.68

Analyzing within categories, Fruits appears to result in greater errors than Animals or Clothes, while S results in greater errors than F or ‘A.’ Across participant groups, there is no clear pattern of one group producing more or fewer errors than the other groups.

As described in the Methodology chapter, total errors were classified into four types: 1) repetitions and derivations (the latter relevant for letter/phonemic trials only), 2) intrusions (words from other categories), 3) non-words, and 4) cross-language insertions. Among the semantic categories, the vast majority of errors were Type 1 (146/164). Ten were Type 2, five were Type 3, and only three were Type 4 (2 instances from participants in the HVL2 group, and one from the MVL2). Numbers were similar among the letter/phonemic categories: 247 Type 1, eight Type 2, two Type 3, and six Type 4 (all of which were produced by one participant).

Type 1 responses consisted of true repetitions, i.e. forgetting that a word had been produced previously and repeating it, as well as rehearsals, in which participants reiterated previously produced words or strings of words aloud, indicating they were aware of having said the word(s) already. Repetitions and rehearsals can be difficult to tell apart, a change in tone (e.g. speaking under one's breath) often being the only signal of a rehearsal. Getting stuck on 'repeat,' so to speak, and vocalizing as such, resulted in inflated error counts for some participants. Using derivations to increase word totals also inflated the error counts for a few participants.

4.3.1.5.2. Error statistical analyses

4.3.1.5.2.1. Aggregated mean error

The mean error count was analyzed by a mixed ANOVA with participant group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects variable. The main effect of category type was significant, $F(1, 118) = 8.736$, $p = .004$, partial $\eta^2 = .069$, in that more errors were produced in letter/phonemic categories than semantic categories. The main effect of group was not significant, $F(3, 118) = .019$, $p = .996$, partial $\eta^2 = .000$, nor was the interaction of category and group $F(3, 118) = .377$, $p = .770$, partial $\eta^2 = .009$.

Of the error types, only Type 1 had a high enough count to analyze. The main effect of category remained significant, $F(1, 118) = 42.225$, $p = .002$, partial $\eta^2 = .076$, but neither the main effect of group nor the interaction between group and category were significant ($p > .05$).

4.3.1.5.2.2. Disaggregated mean error

To examine the disaggregated error counts, an additional mixed ANOVA was conducted with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and trial (animals, fruits, clothing, F, A, S) as the within-subjects factor. With a Greenhouse-Geisser correction, the main effect of trial was the only significant finding, $F(5, 590) = 4.659$, $p = .002$, partial $\eta^2 = .038$. Post hoc pairwise comparisons found that fewer errors were produced in Animals than Fruits ($p = .035$) and S trials ($p = .042$), and fewer in Clothes than S ($p = .023$), while no other differences were significant.

4.3.1.6. Word total summary

In summary, participants produced more words in semantic than letter/phonemic verbal fluency, like in previous work with monolinguals and HBs. Comparison between groups found that HVL2ers produced significantly more words than monolinguals across both category types.. No significant impact was found for age, education, or language testing order, though sex did appear to impact performance, in the Fruits trial alone. Analyses of errors found no significant difference between groups, though participant across groups produced more errors in letter/phonemic than semantic trials. Most errors fell into the category of repetitions, rehearsals, and derivations, while very few errors in category or language (either non-words or non-English words) were produced.

To attain a more in-depth understanding of the differences between the monolinguals and L2ers, the following sections analyze word frequency, retrieval latency, and time-course of retrieval. These measures help gain a deeper picture of the verbal fluency retrieval process to better ascertain the cause of the difference in word total performance.

4.3.2. Word frequency

4.3.2.1. Word frequency descriptive data

Table 4.5. reports the means and standard deviations for word frequency of the words produced in semantic and letter/phonemic verbal fluency categories as a whole as well as disaggregated into their respective individual trials. Across categories, both monolinguals and L2ers produced words of lower average frequency in semantic than letter/phonemic categories. Among the semantic categories, the Animals trial prompted the highest frequency words, followed by Clothes, and finally Fruits. Among the letter/phonemic categories, the highest frequency words were produced in the A trial, followed by F and then S, except for the high-proficiency L2 bilinguals, who produced words of very similar frequency in both the F and S trials.

Table 4.5. Verbal fluency word frequency (occurrence per million)

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	13.47	2.61	15.06	3.01	14.44	2.62	13.75	2.08
Animals	19.75	5.58	23.20	5.98	22.08	5.29	21.61	4.75
Fruits	5.44	1.45	5.34	1.30	5.03	1.21	4.67	1.32
Clothes	15.23	4.27	16.63	5.37	16.15	4.52	14.94	3.55
Letter/phon	299.35	406.23	538.43	503.63	369.39	370.69	323.02	259.30
F	186.61	193.56	274.70	302.61	250.56	237.82	110.38	126.30
A	612.39	1129.59	1232.01	1433.92	747.11	1049.89	751.54	744.38
S	100.36	104.85	108.57	83.05	110.49	88.70	114.02	88.66

Turning to an inspection across groups, the various participant groups appear to produce words of quite similar frequency in the semantic trials, overall and individually. In the letter/phonemic category, LVL2 subjects produce words of higher mean frequency than the other groups, due primarily to their performance in the A trial.

4.3.2.2. Word frequency statistical analyses

4.3.2.2.1. Aggregated word frequency

To determine whether any significant differences exist, statistical analyses were conducted. The mean word frequency count per million, was analyzed by a mixed ANOVA with participant group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects variable. The main effect for category type was highly significant, $F(1, 118) = 107.039$, p

< .001, partial η^2 = .476, in that participants produced significantly lower frequency words in semantic than letter/phonemic categories. The main effect for group, $F(1, 118)$ = 2.219, p = .089, partial η^2 = .053, as well as the interaction between category and group were not significant, $F(3, 118)$ = 2.160, p = .096, partial η^2 = .052.

4.3.2.2.2. Disaggregated word frequency

To examine the disaggregated word frequency counts, an additional mixed ANOVA was conducted with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and trial (animals, fruits, clothing, F, A, S) as the within-subjects factor. With a Greenhouse-Geisser correction, the main effect of trial was significant, $F(1.090, 126.582)$ = 59.331 , p < .001, partial η^2 = .335, but there was no significant interaction between group and trial, $F(3.271, 128.653)$ = 1.703, p = .165, partial η^2 = .042. The main effect of group was not significant, $F(3, 118)$ = 2.212, p = .090, partial η^2 = .053. Pairwise comparisons for trial (p < .001) found that each trial resulted in words of significantly different frequency than all other trials, indicating that mean word frequency is highly dependent on the individual trials chosen for testing. In summary, different trials within each category resulted in different mean frequency counts, but frequency across category types was consistently lower in semantic than letter/phonemic trials.

4.3.2.3. Word frequency language order effects

4.3.2.3.1. Word frequency language order descriptive data

To examine whether activating the L2 first would cause inhibition of the L1 and result in lower mean word frequency, Table 4.6. presents the means and standard deviations for word frequency divided by language order, i.e. those tested in English first or Spanish first. Results are provided for semantic and letter/phonemic verbal fluency as a whole, as well as disaggregated into their respective individual tasks.

Table 4.6. Verbal fluency word frequency by language testing order

		LVL2		MVL2		HVL2	
Category	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	English	14.76	3.82	13.65	2.39	13.67	2.21
	Spanish	15.31	2.21	15.41	2.59	13.83	2.01
Animals	English	22.24	6.66	21.53	4.74	22.66	5.11
	Spanish	24.03	5.41	22.80	6.03	20.77	4.21
Fruits	English	5.33	1.26	4.97	0.98	4.77	1.26
	Spanish	5.35	1.37	5.12	1.48	4.59	1.37
Clothing	English	16.72	6.95	14.47	4.86	13.58	3.51
	Spanish	16.55	3.76	18.32	2.97	16.13	3.14
Letter/phonemic	English	568.92	562.19	371.68	402.55	338.65	265.97
	Spanish	512.00	465.45	366.43	340.22	307.39	260.80
F	English	218.45	234.41	218.07	201.37	127.95	170.88
	Spanish	323.44	352.17	292.33	280.19	87.43	47.66
A	English	1411.57	1699.91	809.77	1133.83	786.41	775.06
	Spanish	1076.40	1232.59	666.54	966.92	716.68	737.88
S	English	76.73	58.99	87.20	68.87	101.59	110.32
	Spanish	136.17	92.54	140.44	104.12	120.59	62.36

Analyzing between categories, comparing language testing order, L2ers' word frequency count was often similar regardless of which language was tested first. The semantic trials are especially consistent regardless of order, while the letter/phonemic trials include more variation. Word frequency was often a bit higher in the F and S trials, when tested in Spanish first, but in the A trial somewhat lower. The standard deviation for the letter/phonemic trials was also quite high.

4.3.2.3.2. Word frequency language order statistical analyses

To determine whether any significant differences exist, a mixed ANOVA was conducted using proficiency group (LVL2, MVL2, HVL2) and language order (English first vs. Spanish first) as between-subjects factors and category type (semantic, letter/phonemic) as the within-subjects factor. No two-way or three-way interactions involving language testing order were significant ($p > .05$). The same results were found when analyzing disaggregated trials. L2ers' mean word frequency counts were statistically equivalent regardless of which language was tested first.

4.3.2.4. Effects of other variables on word frequency

Multiple regressions were run to assess the impact of various other variables on the word frequency of both semantic and letter/phonemic verbal fluency. Variables included age, sex, and highest level of education, which did not significantly predict semantic fluency, $F(3, 118) = 1.071$, $p = .364$, $R^2 = .027$. The set of variables did significantly predict letter/phonemic fluency, $F(3, 118) = 2.976$, $p = .034$, $R^2 = .070$, but none of the individual variable coefficients were significant.

4.3.2.5. Word frequency summary

In summary, participants produced higher frequency words in letter/phonemic categories than in semantic categories, the opposite pattern of Sandoval et al. (2010). Also diverging with Sandoval et al., there were no significant group differences, indicating that word frequency was not impacted by learner/bilingual status or L2 ability. It was also not impacted by language testing order or demographic variables.

4.3.3. Retrieval latencies

4.3.3.1. First-response latency

4.3.3.1.1. First-response latency descriptive data

Table 4.7. reports the means and standard deviations for FRLs in semantic and letter/phonemic verbal fluency categories as a whole as well as disaggregated into their respective individual tasks.

Table 4.7. Verbal fluency first-response latency (seconds)

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	1.78	0.60	1.86	0.78	1.88	0.92	1.95	0.59
Animals	1.74	0.90	1.86	0.90	1.95	1.17	2.07	0.79
Fruits	1.57	0.69	1.90	0.99	1.98	1.16	1.85	0.76
Clothes	2.04	0.91	1.83	1.00	1.71	0.91	1.93	0.83
Letter/phon	1.89	0.94	1.95	1.17	1.92	0.87	1.98	0.65
F	2.17	1.38	2.30	1.88	2.16	1.41	2.10	0.60
A	1.57	0.97	1.75	1.00	2.05	1.47	1.99	1.79
S	1.92	1.32	1.81	1.19	1.55	0.84	2.18	1.36

Across categories, FRLs appear very similar in both semantic and letter/phonemic categories overall as well as individually. FRLs across groups also appear quite similar, though the latency is slightly longer for the L2ers than the monolinguals in several of the trials.

4.3.3.1.2. First-response latency statistical analyses

4.3.3.1.2.1. Aggregated first-response latency

To determine whether any significant differences exist, statistical analyses were conducted. The mean FRLs were analyzed by a mixed ANOVA with participant group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects variable. Neither the main effect for category type, $F(1, 118) = .735$, $p = .393$, partial $\eta^2 = .006$, nor group, $F(1, 118) = .179$, $p = .910$, partial $\eta^2 = .005$ was significant, nor was the interaction between category and group, $F(1, 118) = .050$, $p = .985$, partial $\eta^2 = .001$. Participants in all groups produced equivalent FRLs across category types and regardless of bilingual status.

4.3.3.1.2.2. Disaggregated first-response latency

To examine the disaggregated FRLs, an additional mixed ANOVA was conducted with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and trial (animals, fruits, clothing, F, A, S) as the within-subjects factor. With a Greenhouse-Geisser correction, there was no significant interaction between group and trial, $F(12.051, 464.005) = 1.090$, $p = .367$, partial $\eta^2 = .027$. The main effect of trial only approached significance, $F(4.017, 464.005) = 2.350$, $p = .053$, partial $\eta^2 = .020$.

4.3.3.1.3. First-response latency language order effects

4.3.3.1.3.1. First-response latency language order descriptive data

To examine whether activating the L2 first results in inhibition of the L1 and extended FRLs, Table 4.8. presents the means and standard deviations for FRLs divided by language order, i.e. those tested in English first or Spanish first. Results are provided for semantic and letter/phonemic verbal fluency as a whole, as well as disaggregated into their respective individual tasks.

Table 4.8. Verbal fluency first-response latency by language testing order

		LVL2		MVL2		HVL2	
Category	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	English	2.19	0.75	2.06	0.99	1.83	0.53
	Spanish	1.58	0.72	1.65	0.79	2.07	0.65
Animals	English	2.13	0.77	2.17	1.38	1.87	0.81
	Spanish	1.61	0.96	1.66	0.80	2.26	0.75
Fruits	English	2.17	0.99	2.29	1.23	1.78	0.71
	Spanish	1.67	0.97	1.59	0.96	1.92	0.82
Clothing	English	2.26	1.29	1.71	0.90	1.84	0.60
	Spanish	1.46	0.46	1.69	0.97	2.02	1.02
Letter/phonemic	English	2.48	1.34	2.32	0.93	1.87	0.49
	Spanish	1.50	0.80	1.41	0.39	2.10	0.78
F	English	3.03	2.16	2.74	1.60	2.05	0.58
	Spanish	1.67	1.38	1.41	0.58	2.16	0.64
A	English	2.23	1.10	2.51	1.79	1.56	0.38
	Spanish	1.32	0.66	1.46	0.55	2.43	2.46
S	English	2.17	1.53	1.72	1.02	1.99	1.10
	Spanish	1.50	0.69	1.34	0.51	2.37	1.60

Comparing language testing order across trials, LVL2ers and MVL2ers appear to have longer FRLs when producing in English first, from about one to two-and-a-quarter seconds (mean = 1.2 s). The HVL2ers' FRLs were more similar regardless of which language was tested first, with around half a second difference (mean = .78 s) though the numerical trend is the opposite, for some longer FRLs when producing in Spanish first.

4.3.3.1.3.2. First-response latency language order statistical analyses

To determine whether any significant differences exist, a three-way mixed ANOVA was conducted using group (LVL2, MVL2, HVL2) and language order (English first vs. Spanish first) as between-subjects factors and category type (semantic, letter/phonemic) as the within-subjects factor. The three-way interaction of category, group, and language order was not significant, $F(2, 84) = .705$, $p = .497$, partial $\eta^2 = .017$, and neither was the two-way interaction of category and language order, $F(1, 84) = 2.609$, $p = .110$, partial $\eta^2 = .030$. The two-way interaction of group and language order was significant, $F(2, 84) = 5.068$, $p = .008$, partial $\eta^2 = .108$. There was a statistically significant simple main effect of language testing order for the LVL2 group, $F(1,84) = 9.578$, $p = .003$, as well as the MVL2 group, $F(1, 84) = 7.573$, $p = .007$, but not for the HVL2 group, $F(1, 84) = .898$, $p = .346$. Pairwise comparisons found longer FRLs when the language testing order was English first, for both the LVL2ers and MVL2ers, while the language testing order resulted in no significant differences for the HVL2ers. The same results were found when analyzing disaggregated trials. In summary, LVL2 and MVL2ers' FRLs were longer when tested in English first, while HVL2ers' FRLs were statistically equivalent regardless of which language was tested first.

4.3.3.1.4. Effects of other variables on first-response latency

Multiple regressions were run to assess the impact of various other variables on the FRLs of both semantic and letter/phonemic verbal fluency. Variables included age, sex, and highest level of education, which did not significantly predict semantic fluency, $F(3, 118) = 2.311$, $p = .080$, $R^2 = .055$ or letter/phonemic fluency, $F(3, 118) = .723$, $p = .540$, $R^2 = .018$.

4.3.3.2. Subsequent-response latency

4.3.3.2.1. Subsequent-response latency descriptive data

Table 4.9. reports the means and standard deviations for SRLs in semantic and letter/phonemic verbal fluency categories as a whole as well as disaggregated into their respective individual tasks.

Table 4.9. Verbal fluency subsequent-response latency (seconds)

	Monolinguals		LVL2		MVL2		HVL2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Semantic	21.62	2.75	20.37	2.84	20.20	3.26	21.83	2.25
Animals	24.28	3.01	23.43	2.81	21.80	3.31	23.52	3.53
Fruits	19.24	4.65	17.04	5.77	17.85	5.95	19.76	2.82
Clothes	21.36	3.84	20.64	3.25	20.94	3.43	22.20	3.04
Letter/phon	23.65	2.42	23.95	2.89	23.98	2.21	25.04	2.29
F	22.71	4.16	23.72	3.80	23.33	3.12	23.55	4.52
A	23.29	3.94	23.81	5.21	23.98	4.24	25.61	3.53
S	24.96	3.62	24.32	3.19	24.62	2.62	26.24	3.05

Looking first across categories, monolinguals and L2ers produced shorter mean SRLs in semantic than letter/phonemic categories. Among the semantic categories, Fruits had the shortest mean latency and Animals the longest. Among the letter/phonemic categories, F had the shortest mean latency and S the longest. Turning to an inspection across groups, the monolinguals and L2ers appear to perform quite similarly in both the mean and individual semantic and letter/phonemic categories, while the LVL2 and MVL2ers' mean SRLs appear slightly shorter than the latencies of the HVL2ers.

4.3.3.2.2. Subsequent-response latency statistical analyses

4.3.3.2.2.1. Aggregated subsequent-response latency

To determine whether any differences were significant, statistical analyses were conducted. The mean SRLs, grouped by category type, were analyzed by a mixed ANOVA with participant group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects variable. Unsurprisingly, the main effect of category type showed a statistically significant difference between semantic and letter/phonemic verbal fluency, $F(1, 118) = 106.316$, $p < .001$, partial $\eta^2 = .474$, in which participants' SRLs were shorter in semantic categories than in letter/phonemic categories. The main effect of group was significant, $F(3, 118) = 2.809$, $p = .043$, partial $\eta^2 = .067$. However, pairwise comparisons found no significant group differences (only that the difference between HVL2 and MVL2 approached significance, $p = .060$). There was no statistically significant interaction between category type and bilingual status on SRL, $F(3, 118) = 1.742$, $p = .162$, partial $\eta^2 = .042$.

4.3.3.2.2.2. Disaggregated subsequent-response latency

To explore disaggregated SRLs, an additional mixed ANOVA was conducted with group (monolingual, LVL2, MVL2, HVL2) as the between-subjects factor and trial (animals, fruits, clothing, F, A, S) as the within-subjects factor. The main effect of trial was significant, $F(4.368, 515.436) = 52.354$, $p < .001$, partial $\eta^2 = .307$. Posthoc pairwise comparisons ($p < .005$ in all cases) found significant differences among the pairs of trials except S/A, and A/Animals/ F, such that the length of the SRLs (longest to shortest) was as follows: S/A, A/F/Animals, Clothes, Fruits. The main effect of group was again significant, $F(3, 118) = 3.050$, $p = .031$, partial $\eta^2 = .072$, and with the disaggregated trials, pairwise comparisons found significantly longer mean SRLs for the HVL2ers than the MVL2ers ($p = .045$), but no significant differences between the monolinguals and any of the L2 groups. With a Greenhouse-Geisser correction, the interaction between trial and bilingual status was not significant, $F(13.104, 515.436) = 1.192$, $p = .281$ partial $\eta^2 = .029$.

4.3.2.3. Subsequent-response latency language order effects

4.3.2.3.1. Subsequent-response latency language order descriptive data

To examine whether activating the L2 first would cause inhibition of the L1, resulting in longer SRLs, Table 4.10. presents the means and standard deviations for SRLs divided by language order, i.e. those tested in English first or Spanish first. Results are provided for semantic and letter/phonemic verbal fluency as a whole, as well as disaggregated into their respective individual tasks.

Table 4.10. Verbal fluency subsequent-response latency by language order

Category	Order[a]	LVL2		MVL2		HVL2	
		M	*SD*	*M*	*SD*	*M*	*SD*
Semantic	English	20.92	2.40	21.20	3.41	21.69	2.17
	Spanish	19.90	2.18	18.90	2.63	21.97	2.39
Animals	English	23.51	3.44	21.73	3.14	22.66	3.65
	Spanish	23.37	2.25	21.89	3.62	24.38	3.30
Fruits	English	17.85	5.39	19.95	6.00	19.82	2.90
	Spanish	16.34	6.18	15.16	4.84	19.71	2.83
Clothing	English	21.39	3.12	21.92	2.86	22.58	2.48
	Spanish	19.99	3.33	19.66	3.77	21.81	3.55
Letter/phonemic	English	23.86	3.33	23.76	2.28	25.18	1.84
	Spanish	24.03	2.58	24.25	2.16	24.90	2.72
F	English	23.02	3.73	22.99	3.22	23.99	2.58
	Spanish	24.33	3.87	23.76	3.04	23.12	5.94
A	English	24.32	5.83	23.55	4.28	25.56	2.88
	Spanish	23.37	4.77	24.53	4.29	25.66	4.18
S	English	24.26	4.19	24.74	2.42	25.98	2.93
	Spanish	24.37	2.13	24.48	2.95	26.50	3.05

[a] The language tested first

Comparing language testing order across trials, the SRLs of all three L2 groups were often similar regardless of which language was tested first, and no clear pattern of differences appears. For the LVL2 and MVL2, Fruits and Clothes trials resulted in somewhat longer mean subsequent-response latencies when testing in English first, but the F trial resulted in somewhat shorter means.

4.3.3.2.3. Subsequent-response latency language order statistical analyses

To determine whether any significant differences exist, a three-way mixed ANOVA was conducted using group (LVL2, MVL2, HVL2) and language order (English first vs. Spanish first) as between-subjects factors and category type (semantic, letter/phonemic) as the within-subjects factor. The main effect of language testing order was not significant, F (1, 84) = .923, p = .339, partial η^2 = .011, nor were the two-way interactions of group and language order, F(2, 84) = .334, p = .717, partial η^2 = .008, or category type and language order, F(1, 84) = 3.209, p = .077, partial η^2 = .037, or the three-way interaction between category type, group, and order, F(2, 84) = 2.387, p = .098, partial η^2 = .054.

Similar results were found when analyzing disaggregated trials. With a Greenhouse-Geisser correction, the two-way interaction of trial and language testing order approached significance, F(5, 420) = 2.290, p = .054. Exploratory univariate tests found significant simple main effects for the Fruits (p = .043) and Clothing (p = .032) trials only. In summary, L2ers across proficiency levels produced statistically equivalent mean SRLs regardless of which language was tested first.

4.3.3.2.4. Effects of other variables on subsequent-response latency

Multiple regressions were run to assess the impact of various other variables on the mean subsequent-response latency of both semantic and letter/phonemic verbal fluency. Variables included age, sex, and highest level of education, which did not significantly predict semantic fluency, F(3, 118) = 1.640, p = .184, R^2 = .040, but did significantly predict letter/phonemic fluency, F (3, 118) = 4.023, p = .009, R^2 = .093.

Both sex (p = .047) and highest education (p = .023) significantly added to the prediction. Follow-up one-way ANOVAs found no significant differences between men and women's mean subsequent-response latency, $F(1, 120) = 3.112$, $p = .080$, nor differences based on education level, $F(5, 116) = 1.653$, $p = .152$.

4.3.3.3. Retrieval latency summary

In summary, FRLs were equivalent between categories and across monolingual and L2er groups. Among the L2ers, language testing order did significantly impact the FRLs of the LVL2 and MVL2 groups, in being longer when participants were tested in English first, but FRLs were statistically equivalent for the HVL2 group regardless of which language was tested first. Mean SRLs were longer for letter/phonemic categories than semantic categories. When examining disaggregated trials, the mean SRLs of HVL2ers were longer than the MVL2ers, but there were no significant differences between monolingual and L2 groups. In general, latencies were statistically equivalent regardless of which language was tested first, though the disaggregated trials Fruits and Clothes appear to result in longer mean SRLs when tested first in English, irrespective of bilingual proficiency.

4.3.4. Time-course of retrieval

While mean SRLs represent the spread of exemplars over time, their primary limitation is in reducing all response times to one average. The more robust complement to verbal fluency word totals is the time-course of retrieval, as it takes into consideration the concentration of responses in each 5-s bin over the course of the minute. Both participants who produce few and many words can end up with equivalent SRLs if the responses are similarly spread throughout the minute, whereas graphing and analyzing the time-course allows for the visualization of the density of responses and decline of recall over time.

4.3.4.1. Time-course descriptive data

Though exponential functions have commonly been used to describe free recall performance (Rohrer et al. 1995), Luo et al. (2010) fit scatterplots with both exponential and logarithmic functions and found that logarithmic functions accounted for a larger proportion of the variance in their data. The data in the current study were fitted with both exponential and logarithmic functions, and consistent with previous findings, the logarithmic functions account for more of the variance. Therefore, like both Luo et al. and Friesen et al. (2015), time-course analyses were based on logarithmic functions. Table 4.11. presents the estimated functions from the multilevel models.

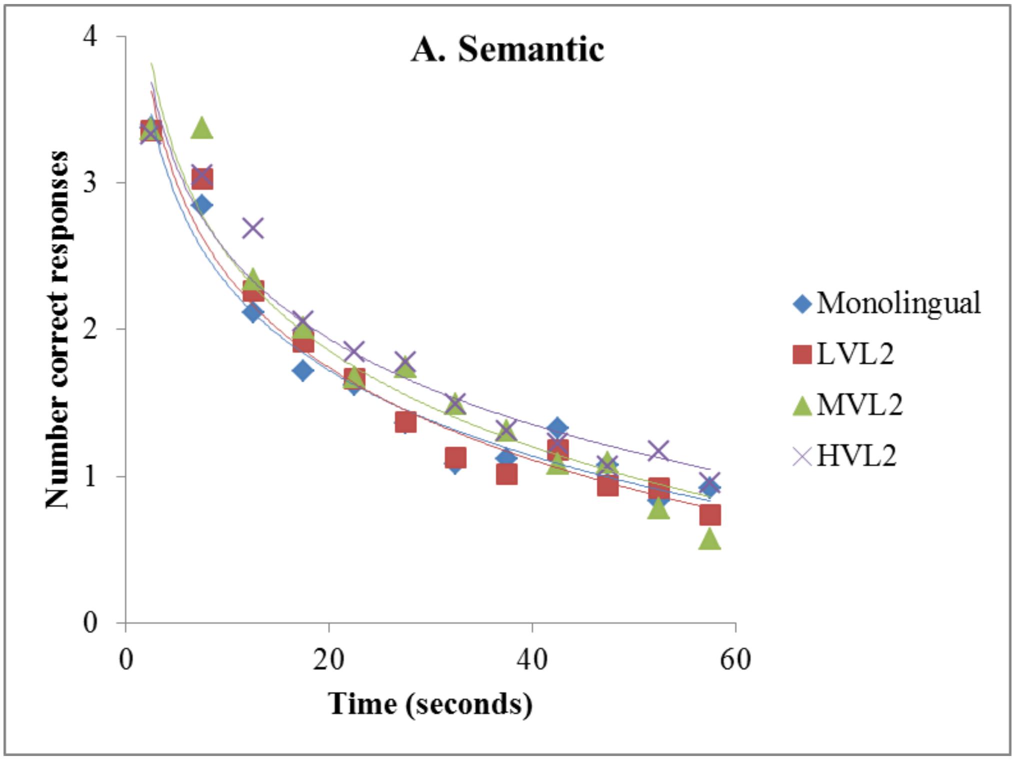
Table 4.11. Best fitting multilevel model functions for time-course of retrieval

	Semantic	Letter/phonemic
Monolingual	$y = 4.99 – 1.03\ln(t)$	$y = 3.32 – 0.63\ln(t)$
LVL2	$y = 5.29 – 1.12\ln(t)$	$y = 3.18 – 0.58\ln(t)$
MVL2	$y = 5.58 – 1.17\ln(t)$	$y = 3.41 – 0.62\ln(t)$
HVL2	$y = 5.26 – 1.05\ln(t)$	$y = 3.21 – 0.53\ln(t)$

Figures 4.1 and 4.2 graph the responses of semantic and letter/phonemic verbal fluency, respectively, with time in seconds on the *x*-axis and mean correct responses on

the *y*-axis. Responses were grouped into 5-s bins based on total latency[4] and group means for response are plotted in the mid-point of each 5-s bin (Rohrer et al. 1995; Luo et al. 2010).

Figure 4.1. Semantic exemplars produced as a function of time

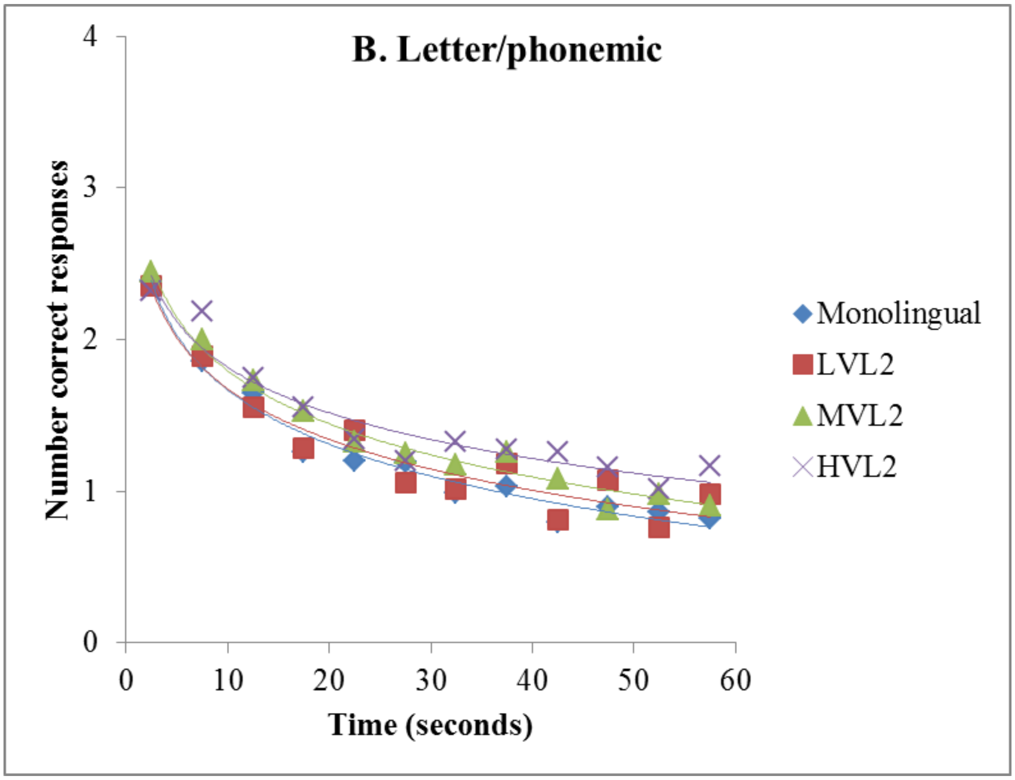


Beginning with (A) semantic verbal fluency, a visual inspection of the best fit lines finds that the MVL2ers' initiation is slightly higher than that of the HVL2 and LVL2ers, which are in turn greater than that of the monolinguals. Examining the slope,

[4] Rohrer et al. (1995) grouped responses into 5-s bins based on subsequent-response latency rather than total latency. They do not provide a rationale for doing so, and the decision to include first-response latencies in the current project is explained in Section 3.5.3.3.3. Luo et al. (2010) cite Rohrer et al. for their procedure in calculating first- and subsequent-response latencies, but state that responses were grouped into bins according to time-stamp. It is unclear from their description whether their time-course represents all responses as in the current study or subsequent responses, as in Rohrer et al.

by the end of the trial, the fit line of the HVL2ers is higher than that of the other three groups.

Figure 4.2. Letter/phonemic exemplars produced as a function of time



Moving to the (B) letter/phonemic verbal fluency, the initiation according to the best fit line is visually equivalent across groups. Looking at the slope, the best fit line by the end of the trial is highest for the HVL2ers, followed by MVL2ers, LVL2ers, and finally monolinguals.

4.3.4.2. Time-course tatistical analyses

To determine whether any significant differences exist, statistical analyses were conducted. Following Luo et al. (2010), group differences were assessed with multilevel modeling with the lme4 package in R. Twelve observations were included from each individual, one per each of the twelve 5-s bins, resulting in 1464 total observations in the whole model. The multilevel model was fitted by maximum likelihood methods and based on logarithmic transformations. The main effect of group represented the intercept, or “initiation parameters” (Luo et al. 2010), while the interaction of group and time represented the slope of the curve.

Analyses were conducted separately for semantic and letter/phonemic fluency. In the semantic fluency time-course analysis, there was a significant main effect of group, $F(3, 118) = 3.381$, $p = .021$. Exploratory group contrasts were conducted to further distinguish between participant groups. The only significant difference in intercept, $t(118) = 2.609$, $p = .01$ was between the MVL2ers and monolinguals. Though the interaction between group and time was not significant, $F(3, 1338) = 2.026$, $p = .108$, group contrasts found a significant difference in slope, again between the MVL2ers and monolinguals, $t(1338) = 2.189$, $p = .029$, and the difference in slope between the MVL2ers and HVL2ers approached significance, $t(1338) = 1.904$, $p = .057$. Time-course analysis of letter/phonemic fluency also found a significant effect of group, $F(3, 118) = 2.803$, $p = .043$, but no group contrasts were significant (ps > .05). Though there was no significant interaction between group and time, $F(3, 1338) = 1.471$, $p = .221$, group contrasts found that the difference in slope between the HVL2ers and monolinguals approached significance, $t(1338) = 1.915$, $p = .056$.

4.3.4.3. Time-course summary

In summary, in semantic fluency, MVL2ers had a higher intercept than the monolinguals as well as a steeper slope. In letter/phonemic fluency, no groups differed in intercept or slope, though the slope of the HVL2ers appeared more gradual and the difference with the steeper monolingual slope approached significance.

4.4. Summary

Overall, analysis of the performance in the various tasks employed in the current study found the following results: English monolinguals and English-Spanish L2ers at multiple levels of L2 ability did not differ significantly in their English receptive vocabulary knowledge, according to two measures. In verbal fluency, HVL2ers produced more English words across both semantic and letter/phonemic fluency than monolinguals, while the other groups did not differ. Across participant groups, mean word frequency, FRL, and SRL were equivalent in both category types, though LVL2 and MVL2 participants who were tested in English first were slightly slower to start producing words than those tested in Spanish first. Finally, in the time-course of semantic retrieval, MVL2ers had higher intercepts and steeper slopes than monolinguals, while in the time-course of letter/phonemic retrieval, the difference in slope between HVL2ers and monolinguals approached significance. The following chapter will discuss the results of the data analysis in light of the previous literature and how it supports or conflicts with the various accounts presented in the research questions and hypotheses.

Chapter 5: Discussion

5.1. Introduction

The current study aimed to examine the impact of formal L2 learning on L1 lexical retrieval by comparing English-Spanish L2ers with English monolinguals on their performance in English verbal fluency, a relatively unconstrained psycholinguistic task measuring lexical access and retrieval for production. Measures included total correct words produced, errors (see 3.5.3.1 for error classification), mean word frequency (3.5.3.2), retrieval latency (3.5.3.3.2) including first-response latency (FRL) and mean subsequent-response latency (SRL), and intercept and slope of the time-course of retrieval (3.5.3.3.3).

In short, group differences were found in the word totals and time-course of retrieval, while group differences in the rest of the measures were not significant. Among the four participant groups, monolinguals and L2ers at three levels of L2 ability (LVL2, MVL2, and HVL2), word totals differed between the HVL2 bilinguals and the monolinguals. Contrary to theories of retrieval slowing (Hypothesis 1, 1a,c) or reduced vocabulary (Hypothesis 2), and in line with theories of bilingual executive control advantages (Hypothesis 3), HVL2ers produced more English words than English monolinguals. In contrast with previous findings that HBs produce words of lower average frequency in semantic fluency than monolinguals (Sandoval et al. 2010), monolinguals and all L2 groups produced words of equivalent frequency. Also in contrast with findings that HBs produce longer SRLs in letter fluency (Friesen et al. 2015; Luo et al. 2010; Sandoval et al. 2010), all groups produced equivalent response latencies.

Finally, in the time-course of retrieval, MVL2ers, but not HVL2ers, had a higher intercept and steeper slope than monolinguals in semantic fluency. Partially in line with findings that vocabulary-matched HBs have gentler letter/phonemic slopes than monolinguals (Friesen et al. 2015; Luo et al. 2010), the difference in slope between the HVL2ers and monolinguals neared significance.

These findings imply, counter to Hypothesis 1a, that cross-linguistic competition is not a hindrance to L1 lexical retrieval in L2ers, at least not in the relatively unconstrained environment of a verbal fluency task. Rohrer and Wixted (1994) claim that latency of retrieval is a reflection of the size of the search set, while Rohrer et al. (1995) add the speed of processing, such that "the average time needed to retrieve the items within the search set increases when either the size of the search set increases or the duration of each random sampling increases" (p.1130). If cross-linguistic competition or inhibition (Hypothesis 1a,b) were a factor during verbal fluency for L2ers, these learners/bilinguals could be susceptible to either or both factors. If activated words from both languages compete for selection, the search set is larger than for a monolingual, and if bilinguals must inhibit the non-target language, or if the conceptual links are weaker due to decreased exposure (Hypothesis 1c), retrieval may be slowed, resulting in fewer words produced and longer retrieval latencies. The results of the current study do not present evidence that either of the above processes were at play for the L2ers at any L2 ability level.

In contrast with the predictions of several accounts that predicted lower word totals (Hypotheses 1 and 2), all learner/bilingual groups performed at least as well in word totals and had statistically equivalent mean SRLs relative to monolinguals. Looking only at the search set factor, these results could imply that L2ers are enacting global inhibition of the non-target language, reducing their effective vocabulary to the same as a monolingual, but inhibition would likely entail slowed retrieval. Alternatively, in line with the language-specific selection mechanism proposed by Costa et al. (1999), L2ers may be selecting only among competitors in the target language without specific inhibition of the non-target language. However, Costa and colleagues propose differential language control mechanisms for low- and high-proficiency bilinguals (irrespective of age of acquisition, Costa et al. 2006), in which the former rely on inhibitory mechanisms while the latter have developed a language-specific selection mechanism, which does not explain the lack of disadvantage seen in the LVL2 (and potentially MVL2) participants, unless the required inhibition of the weak L2 was so minimal as to not create a measurable difference in verbal fluency performance.

More interestingly, the HVL2ers produced more words than the monolinguals. This result, though potentially surprising given the predictions of the retrieval slowing or reduced vocabulary accounts, partially replicates findings with vocabulary-matched HBs (Bialystok et al. 2008; Luo et al. 2010; Friesen et al. 2015), which the authors attributed to bilingual executive control advantages (Hypothesis 3). The findings of the current project differ in two key ways, however. First, the current project found no group by category interaction, whereas the previous studies found production advantages for the letter/phonemic category but not semantic, which was attributed to the differential level of difficulty of the two types of category, with the letter/phonemic category relying more on executive control. Second, the current project found no significant difference in mean

SRL for any of the L2 groups relative to monolinguals. Luo et al. (2010) and Friesen et al. (2015) reported that bilinguals (regardless of vocabulary knowledge) evidenced longer mean SRLs in the letter/phonemic category, from which they interpreted that “the group has superior control… and could continue generating responses longer” (Friesen et al. 2015: 242). These studies do not provide an anticipated interpretation if more words are produced but with equal mean SRLs. Mean SRL is a measure of the spread of words over time; therefore, equivalent mean latency combined with higher word totals, rather than evidence of performance endurance, indicates a higher concentration of words in the same time period, suggesting increased retrieval efficiency. In essence, both HBs and highly proficient L2 bilinguals show signs of an executive control advantage, but with slightly different manifestations within task control.

The remainder of this chapter is divided into the following sections: Section 5.2. will describe how the results of the current study align with each of the theories presented in the Research Questions from Chapter 2. Section 5.3. will compare semantic and letter/phonemic fluency in light of previous studies, and Section 5.4. will discuss the current project’s limitations and the directions for future research. Finally, Section 5.5. will offer concluding remarks.

5.2. Review of Research Questions and Hypotheses

5.2.1. Retrieval slowing

The early research that found disadvantaged verbal fluency production for HBs (Gollan et al. 2002; Portocarrero et al. 2007; Rosselli et al. 2000) attributed findings to retrieval slowing due to competition between the two languages, variously from the effort of suppressing the target language in general or individual translation equivalents. An alternative account, in line with frequency effects in picture naming (e.g. Gollan et al. 2008), was presented to account for retrieval slowing not from competition but from weaker connections between each of a bilingual's given languages and the conceptual representation. Given the previous research attributing HB verbal fluency performance to retrieval slowing, the first research question of the current study asked whether L2ers show evidence of retrieval slowing while producing in their L1.

5.2.1.1. Competition

Retrieval slowing through competition was hypothesized to manifest (Friesen et al. 2015; Luo et al. 2010; Sandoval et al. 2010) through lower word totals, longer retrieval latencies and a time-course with a lower intercept and more gradual slope than monolinguals. Additionally, average word frequency was predicted to be lower than monolinguals. The results of the study did not confirm any of the predictions of the competition model. Rather than producing fewer words, L2 groups produced equivalent or more (in the case of HVL2) words in English than English monolinguals, across semantic and letter/phonemic fluency. The word frequency, FRLs, and mean SRLs were statistically equivalent across all groups and both category types. The intercept of the time-course was equivalent or higher (for MVL2ers in semantic fluency) than monolinguals, and the slope was generally equivalent as well. The exception in the last case is that the MVL2ers had a steeper slope than monolinguals in semantic fluency. The results of the current study conflict with Linck et al. (2009), who found temporarily disadvantaged L1 retrieval during an L2 study abroad immersion experience. The finding of equivalent or advantaged L1 word totals in the current study, along with the fact that the L1 disadvantage found by Linck et al. was temporary and had disappeared within six months of returning to the L1 majority environment, underscores that the effect was due to the specific L2 immersive environment rather than increased proficiency from the study abroad experience.

The contrast between the results of the current study and those attributing results to competition-based retrieval slowing might suggest that HBs experience cross-linguistic competition in their verbal fluency production that L2ers do not, perhaps due to the specifics of the language acquisition experience or the relation of language chronology and dominance. The HBs recruited and examined in these studies are likely to have or at least self-report more balanced proficiency in each of the languages than an L2er; in addition, the target language is typically the L2 (albeit often a dominant L2) for HBs while being the dominant L1 for the L2ers. This combination of proficiency and chronology may mean the non-target language provides stronger competition. In an attempt to minimize the proficiency imbalance, the current study included a group of highly proficient L2 bilinguals, for whom the strength of the L2 would provide the strongest potential competition. Rather than reduced production based on the strength of the L2, the opposite result was found, in which the group with the highest proficiency in the L2 produced the most words. While the differences between the other groups were not significant, the numerical trend favored incremental increases in verbal fluency performance as proficiency increased. That direct relationship between proficiency and performance, rather than the expected inverse relationship, is strong evidence against the competition account in relation to L2 learners/bilinguals, insofar as wheter it negatively impacts L1 verbal fluency performance.

5.2.1.1.1. Language order effects

Testing L2er participants in both languages offers an additional angle from which to look for evidence of competition, in the form of asymmetric cross-linguistic inhibition. Cross-linguistic inhibition was hypothesized to impact verbal fluency measures through disadvantages for participants tested in the non-dominant L2 (Spanish) first, as that is the direction in which the dominant L1 would be suppressed before production. Converse to those predictions, the order in which languages were tested, English first or Spanish first, did not significantly impact the results for the majority of L1 verbal fluency measures in the current study for any group. This result fails to replicate the word total findings of Van Assche et al. (2013), in which Dutch-English L2ers tested in letter/phonemic categories produced fewer L1 words if tested in the L2 first. Their finding was restricted to repeated trials, i.e. English FAS followed by Dutch FAS, and was interpreted as evidence for item-specific (though not restricted to translation equivalent) inhibition. Multiple characteristics of the Van Assche et al. study are shared by the current project: participants were L1 dominant, learning the L2 mostly formally in adolescence; testing occurred in an immersive L1 environment (Belgium), and trials were completed in language blocks (all Dutch then all English, or vice versa). Participants self-rated quite fluent in the L2 (English; 7.3/10), which falls in between the self-ratings of the MVL2 and HVL2 groups. Proficiency is not an explanatory factor for the different results, given that multiple proficiency levels were tested in the current study, and none of them evidenced a language testing order effect (no interaction between group and language order, and no language order effects in follow-up planned comparisons). The main difference between studies was that the current project also included three semantic fluency trials, which might be argued to break up the block and give time for the inhibition to abate, except that Van Assche et al. also included three non-repeated letter trials, i.e. letters that were tested in one language but not the other, which would serve the same purpose.

The one exception to the null language testing order findings in the current study was the FRLs of the LVL2 and MVL2 groups. An inhibition account would predict asymmetric switch costs such that prior activation of the L2 would trigger greater suppression of the L1 than vice versa, resulting in slower initiation of production when tested in the L2 first. Surprisingly, rather than longer FRLs after producing in the L2, FRLs were longer for participants producing in the L1 *English* first, the opposite of the predicted pattern. The original analyses do not uncover whether the difference is evidence of a boost in speed for the LVL2 and MVL2 participants who produced in Spanish first or a delay in production for those who produced in English first. To examine those possibilities, LVL2 and MVL2 participants were divided according to language testing order (English-first or Spanish-first) and compared to the monolinguals. The descriptive data is presented in Table 5.1.

Table 5.1. Verbal fluency first-response latency by language testing order, monolinguals and low/mid-vocabulary L2 learners

	Monolinguals		English-first L2		Spanish-first L2	
	M	*SD*	*M*	*SD*	*M*	*SD*
Semantic	1.78	0.60	2.11	0.89	1.61	0.74
Letter/phonemic	1.89	0.94	2.39	1.10	1.45	0.63

Considering the monolinguals as neutral controls, visually it appears the FRLs of the Spanish-first L2ers are slightly shorter than the monolinguals (average .31 s), suggestive of an advantage. Simultaneously, the FRLs of the English-first L2ers are somewhat longer than the monolinguals (average .41 s), indicating a disadvantage. To examine whether either difference was significant, a two-way mixed ANOVA was conducted using group (monolingual, Spanish-first L2, English-first L2) as the between-subjects factor and category type (semantic, letter/phonemic) as the within-subjects factor. Neither the main effect of category, $F(1, 89) = .574$, $p = .451$, nor the interaction between category and group, $F(2, 89) = 1.689$, $p = .191$, were significant. Unsurprisingly, the main effect of group was significant, $F(2, 89) = 7.952$, $p = .001$. Multiple comparisons maintained the significant difference between the English-first and Spanish-first L2 groups, $p < .001$. The difference between the monolinguals and Spanish-first L2ers was not significant, $p = .220$, while the difference between monolinguals and English-first L2ers approached significance, $p = .053$. The statistics thus cannot confirm either account definitively, though it appears more likely that the order effect difference for the LVL2 and MVL2 groups was due to a delay in their initial English production when tested in English first.

These results are surprising but may be suggestive of a processing strategy favoring the L2 when L2 production is anticipated. Costa and Santesteban (2004) found that participants in a language switching picture naming task were slower in the L1, even in non-switch trials. To account for this, they hypothesized that the bilinguals had “bias[ed] the lexicalization process towards their weak language” (p.502), creating an artificially high threshold for activation of the L1 in anticipation of producing in the L2, though they were unable to demonstrate or reduce the hypothesized bias through further manipulation. Kroll et al. (2002) found that less proficient L2ers named L1 words slower than more proficient L2ers, suggesting a cost to the L1 from the process of L2 learning. Applying that hypothesis to the current study, LVL2 and MVL2 participants assigned to the English-first condition may have been distracted by the knowledge that they would be later expected to produce in Spanish as well, and the expectation of future L2 production was sufficient to prompt momentary inhibition, enough to result in initial hesitation but temporary enough not to affect overall latencies through longer SRLs.

5.2.1.2. Weaker links

Of the possible accounts for retrieval slowing, weaker links was considered to be less likely because of the results of prior studies and because L2ers, having grown up monolingual and maintained a mostly immersive L1 environment (minus periods studying or living abroad), were less likely to experience the frequency effects hypothesized to impact HBs. The predictions of retrieval slowing through weaker links (Sandoval et al. 2010) differed from competition only in average word frequency, in which L2ers would produce words of higher mean frequency than monolinguals. The results of the study do not support the weaker links alternative retrieval slowing account, as word frequency was equivalent across all groups. Like with competition, the relationship of L2 proficiency to L1 performance, in which HVL2ers produced more words, is evidence in direct contrast with the predictions of the weaker links account and retrieval slowing in general.

5.2.2. Reduced vocabulary

More recent heritage bilingual verbal fluency studies found word total results that conflicted with the earlier studies, in which HBs matched with monolinguals on vocabulary knowledge produced equivalent word totals in semantic categories and *more* words in letter/phonemic categories. Only those with lower vocabulary knowledge produced fewer words, and only in semantic fluency, prompting the consideration of whether "many of the bilingual disadvantages reported in previous studies would be more accurately interpreted as reflecting a smaller vocabulary than weaker control of lexical processing" (p.535). Though the reduced vocabulary account was not considered a likely contributor to L2er verbal fluency performance, for the same reasons as weaker links (above), the second research question asked whether L2ers show evidence of reduced L1 vocabulary as a sacrifice for the time spent learning and interacting in the L2. This question was addressed directly through English vocabulary testing as well as indirectly through verbal fluency performance.

Like competition, reduced vocabulary was predicted to manifest as reduced verbal fluency word totals. However, rather than longer response latencies and lower average word frequency, reduced vocabulary was hypothesized to result in shorter mean SRLs and higher average word frequency. Again, the results of the verbal fluency analyses did not confirm any of the above predictions. Additionally, the scores on the English vocabulary tests did not decrease as Spanish vocabulary scores increased and were equivalent across groups, evidence against the possible claim that time spent in L2 study decreases exposure to the L1sufficient to impact vocabulary knowledge.

5.2.2.1. Problematizing the reduced vocabulary retrieval latency prediction

Importantly, the retrieval latency predictions set forth by Sandoval et al. (2010) and Luo et al. (2010) regarding vocabulary size were adapted from category size experiments by Rohrer et al. (1995) based on the random-search model. As discussed in the literature review, while this model is useful for describing the decline of retrieval over time, it is simplistic. Among its limitations is that it is restricted to activation of exemplars and since "[e]ach item has the same probability of being sampled," (p.1129), it does not incorporate competition. As a result, the reduced vocabulary account implicitly excludes competition (both within and across languages) from the retrieval process. However, the results comparing the SRLs of Luo et al.'s (2010) LV and HVHBs call into question the validity of the reduced vocabulary prediction regarding retrieval latency. It is therefore instructive to reconsider the assumptions behind the reduced vocabulary retrieval latency prediction.

The reduced vocabulary account as presented (Bialystok et al. 2008; Luo et al. 2010; Sandoval et al. 2010), and upon which predictions were made in previous research as well as the current study, considers whether production differences between monolinguals and HBs are due to vocabulary pool *rather than* from cross-linguistic competition from translation equivalents (semantic fluency) or non-target language word-initial neighbors (letter/phonemic fluency). The implicit exclusion of competition in the retrieval process is made apparent in that only the words in the target language are assumed to count as part of the exemplar pool, which is how it could be argued that HBs have fewer lexical resources than monolinguals. However, nonselective access theories, as presented in the literature review, argue that exemplars from both languages can and do compete during lexical retrieval. While bilinguals tend to have smaller individual vocabularies relative to monolinguals in each of their languages, the combined vocabulary from both languages is larger than a monolingual. If nonselective access were assumed, a very different mean SRL prediction would emerge, resulting in an additional possible interpretation to account for Luo et al.'s LVHB performance. The larger combined (i.e. dual language) exemplar pools for both the LV and HVHBs, relative to monolinguals, could account for the longer SRLs (in both Luo et al. 2010 and Sandoval et al. 2010), while the size of the relevant (i.e. target language) vocabulary would explain the difference in word totals between the HBs and monolinguals.

Though the predictions put forth by Sandoval et al. (2010) and Luo et al. (2010) implicitly measure and take into account the vocabulary knowledge of only the target language, it is unclear whether the reduced vocabulary account rejects or ignores dual language activation entirely or if it instead aligns itself with the language-specific selection mechanism account (Costa et al. 1999; Costa & Santesteban 2004; Costa et al. 2006) that maintains dual-language activation but excludes the non-target language from the selection process. As Costa and colleagues argue that the selection mechanism develops through bilingual proficiency and is not present in low proficiency bilinguals, it could be appropriately assigned to the high proficiency HBs studied previously (Bialystok et al. 2008; Friesen et al. 2013; Luo et al. 2010; Sandoval et al. 2010) and would likely apply to the HVL2 participants in the current study.

It would not, however, appropriately predict the selection mechanism of the LVL2ers (and potentially MVL2ers). Assuming dual-language activation, if Costa and colleagues (Costa et al. 1999; Costa & Santesteban 2004; Costa et al. 2006) are correct that non-target language lemmas compete for selection in lower proficiency learners while proficient bilinguals, even those with L2 learning backgrounds, develop language-specific selection mechanisms allowing them to exclude the non-target language from competition for selection, then multiple predictions emerge for the participants in the various ability levels. For participants at the LVL2 level, either non-target lemmas are still competing for selection, which would result in a larger search set size, or non-target lemmas that were activated by the prompt were inhibited, which requires effort and would likely increase the duration of sampling, either of which would result in longer SRLs relative to a monolingual (Rohrer et al. 1995). Meanwhile, participants at the HVL2 level, if sufficiently proficient in the L2, would be selecting only from the target language exemplars, resulting in monolingual-like search set size, which arguably would result in equivalent SRLs. The matter is complicated further by accounts that combine vocabulary knowledge and executive control. HBs and L2ers of varying proficiency levels could be predicted to have longer SRLs for a variety of reasons—non-target language competition for low-proficiency learners (though the general weakness of the L2 would require relatively little suppression and might not result in measurable differences; there is evidence that at early stages of L2 learning, the impact of L2 on L1 processing can be detected even when behavioral differences are not significant, e.g. Bice & Kroll 2015) and executive task control advantages for high-proficiency bilinguals.

Having considered alternative predictions of how bilingual vocabulary knowledge impacts retrieval latencies according to the random-search model, it is important to examine the results of the current study for any light they might shed. The HVL2ers did produce mean SRLs equivalent to monolinguals, which could be interpreted as indicating that only target language words competed for selection, in support of the language-specific selection mechanism. However, the LVL2ers did not produce longer SRLs, implying that the non-target language was not competing for selection at that level of proficiency either, possibly at odds with dual language activation and inhibition accounts. Given the breadth of evidence for dual-language activation, it is more likely the case that the weak L2 did not provide sufficient competition or require substantial enough inhibition to impact retrieval latencies.

The case of the MVL2ers remains somewhat muddled, as there is no definitive cut-off for what defines high and low proficiency learners/bilinguals. As an in-between group, their L2 may have been weak enough not to impact SRLs or strong enough to have developed language-specific selection. Schwieter and Sunderman (2008) found a threshold beyond which asymmetric switch costs disappeared, based on the robustness of L2 lexical knowledge, which interestingly enough was measured with semantic verbal fluency. The threshold their statistical analyses uncovered corresponded roughly to a score of 110 from 10 semantic trials, averaging out to 11 responses per trial. The MVL2ers in the current study produced an average of 9 responses per L2 semantic trial, suggesting they were not proficient enough to have developed language-specific selection, though that conclusion is purely speculative given the differences in cues.

In summary, the retrieval latency effects of category size have been extrapolated to make predictions about bilingual verbal fluency. These predictions, however, do not clarify how the mathematical random-sample model blends with psycholinguistic models assuming competition for lexical selection. This consideration is even more crucial when predicting lexical selection for people who speak two languages, and requires explicit justification for which language's exemplars form part of the search set, given substantial evidence for dual-language activation, competition, and inhibition. Combining theoretical accounts with bilingual proficiency levels results in multiple possible overlapping predictions. The results of the current study do not definitively offer evidence in favor of any particular interpretation. Further consideration of the implications of lexical knowledge and bilingual proficiency suggests that mean SRLs do not reliably distinguish between accounts involving vocabulary.

5.2.2.2. Inconsistencies in time-course predictions by category vs. lexicon size

Regarding intercept and slope of the time-course, there are mixed graphic representations of reduced vocabulary. Rohrer et al.'s (1995) hypothesized recall predicts a lower initiation and similar initial slope that reaches floor earlier, while the results of their experiment comparing large and small categories resulted in similar initiation, but a much steeper slope. Sandoval et al.'s (2010) vision of a reduced vocabulary timeline reflects what Rohrer et al. found, in which bilingual retrieval mimics monolingual retrieval rates at the beginning of the trial but runs out of exemplars earlier. However, their results found lower initiation and more gradual curve, from which they conclude against reduced vocabulary entirely. However, Luo et al.'s (2010) hypothesized recall for reduced vocabulary was just that—lower initiation and more gradual curve. Luo et al. attributed the intercept to differences in vocabulary resources but the difference in slope to bilingualism. Taken together, the results of experiments controlling for category size vs. overall vocabulary size suggest that the intercept and slope patterns are not equivalent or directly comparable; in other words, knowing fewer words may not result in the same time-course pattern as categories with fewer possible exemplars. Without normed monolingual retrieval latency data based on vocabulary knowledge (or at least education), it is difficult to conclude how vocabulary size alone impacts the initiation and slope of the time-course.

The time-course results of the current study partially diverge from Luo et al.'s (2010) assertion that initiation reflects lexical resources. In line with the equivalent vocabulary scores across groups, there were no differences in letter/phonemic fluency intercept. However, the MVL2ers had a higher semantic fluency intercept than the monolinguals, which should indicate greater vocabulary knowledge on the part of the MVL2ers. Though not significant, the numeric trend across groups was for English vocabulary scores to increase from the monolinguals to the L2 groups and along with L2 ability. This trend is partially reflected in the best fit line intercepts, though a vocabulary-based intercept would predict the highest intercept for HVL2ers.

In summary, L1 vocabulary knowledge was not considered a likely factor for verbal fluency performance in L2ers, given their L2 acquisition experience. Results of vocabulary measures and verbal fluency tasks show no sign of a reduced L1 lexicon as a consequence of L2 study, even for the HVL2 participants who had dedicated substantial time to L2 learning.

5.2.3. Executive control advantage

To account for findings that vocabulary-matched HBs produce more letter/phonemic exemplars than monolinguals, Bialystok et al. (2008) propose that, when bilingual performance is not hindered by reduced vocabulary knowledge, a bilingual executive control advantage allows for enhanced task management in the letter/phonemic fluency task. Findings have confirmed that HBs have longer SRLs regardless of vocabulary, which combined with the greater letter/phonemic word totals of HVHBs, have been interpreted as further evidence of enhanced executive control.

Given the more recent verbal fluency studies found advantages after accounting for vocabulary knowledge, the third research question of the current study asked whether L2ers show signs of enhanced executive control in their L1 verbal fluency production. Enhanced control was predicted (Friesen et al. 2015; Luo et al. 2010) to manifest as higher word totals, longer mean SRLs and a time-course with similar initiation but a flatter slope. Results partially aligned with these predictions, in that the HVL2 group produced more words than monolinguals. This effect spanned both category types (there was no interaction between group and category), in contrast with other studies (Friesen et al. 2015; Ljungberg et al. 2013; Luo et al. 2010), who found bilingual advantages only for letter/phonemic fluency, not semantic. It is difficult to account for these results, especially since most previous verbal fluency studies with L2ers (Baus et al. 2013; Linck et al. 2009) did not compare with L1 monolinguals and only tested semantic categories and therefore cannot be used for comparison. Ljungberg et al. (2013) found a bilingual advantage for Swedish-native L2ers in letter/phonemic fluency but not semantic fluency, though comparison is difficult because each category tested only one trial, and the so-called semantic task combined a semantic with a phonemic restriction (occupations starting with B).

The results of the current study could be interpreted to imply that highly proficient L2ers gain an executive control advantage that early heritage bilingualism does not impart, one that is not restricted to the task endurance required by the more difficult letter/phonemic task. On the other hand, distinctions in population (beyond the L2/heritage divide) and/or methodology may have played a part. The bilingual populations in Bialystok et al. (2008), Luo et al. (2010), and Friesen et al. (2015) represented a variety of non-English language backgrounds. Little research has systematically compared verbal fluency performance between speakers of various languages. Kempler et al. (1998) compared the semantic fluency of Chinese, Hispanic, and Vietnamese immigrants in their respective native languages, finding that more words were produced in Vietnamese than Spanish, which was attributed primarily to the differences in average syllable length. However, that study did not test participants in the shared L2 (English) to determine whether the native language's (dis)similarity to English impacted word totals. It is possible that the mixed L1 of the bilingual studies obscures language-related effects specific to semantic verbal fluency such as cognate status. Additionally, the current study did not vocalize during the participant instructions a number of additional restrictions to letter/phonemic fluency (no numbers, no names of people or places) that had been included in previous studies for the purpose of increasing the executive processing demands relative to the semantic task (Bialystok et al. 2008; Friesen et al. 2015; Luoet al. 2010). Therefore, it is possible that the gap in executive load required by the two tasks was smaller than in the previous studies, which could account for the lack of interaction between group and category. Importantly, it does not explain why a word total advantage was still found for *both* category types, as no additional restrictions were placed on the semantic instructions that would have necessitated increased task control.

Another important distinction between the current results and the predictions of the executive control advantage account resides in the retrieval latencies. FRLs and mean SRLs were equivalent across groups, whereas both Luo et al. (2010) and Friesen et al. (2015) found longer mean SRLs for HBs. Though they considered how to interpret longer mean SRLs with either lesser or greater word total, a combination not explored by Friesen et al. is one in which one group produces more correct responses with an *equivalent* mean SRL. Rather than producing words at a similar rate but stretching longer into the time-course, indicative of task endurance, such a result involves fitting more responses into the same amount of time, which could be understood as task efficiency, thus implying that high proficiency L2 bilingualism impacts L2ers by increasing the efficiency of their lexical retrieval process irrespective of category type.

One potential caveat to the above deserves brief mention. Due to Friesen et al.'s (2015) and Luo et al.'s (2010) finding of increased SRLs specifically for letter/phonemic fluency, an exploratory one-way ANOVA of the letter/phonemic SRLs comparing only the monolinguals and HVL2ers (i.e. excluding LVL2 and MVL2 data) was significant, $F(1, 60) = 5.418$, $p = .023$, in that the SRL of the HVL2ers was longer. Of course, this result should be taken with a grain of salt as evidence of at best a very weak group effect, given that significance was lost in the more robust mixed ANOVA considering all participant groups and both category types. Interestingly, it suggests further alignment with the prior HB studies' conclusion of enhanced task control, referred to here as task endurance (i.e. more words over longer period of time). As mentioned above regarding word totals, a stronger letter/phonemic effect may have been lost with the reduction in complexity of the restrictions specified by the instructions. Importantly, a corresponding one-way ANOVA of semantic SRLs was not significant, and thus the distinction between task endurance and task efficiency (i.e. more words in equivalent time) remains relevant.

The final measure provided by the verbal fluency data was the time-course, and the results of the graphing and analysis also vary in multiple ways from previous studies. As mentioned above in the discussion of reduced vocabulary predictions, Luo et al. (2010) found no intercept or slope group differences in semantic fluency, regardless of vocabulary knowledge, while the current results examining time-course of retrieval in L2ers found intercept and slope differences in semantic fluency between the monolinguals and MVL2ers. Specifically relevant to the impact of bilingualism, the MVL2 group's slope was steeper, a surprising result given that it was not found in either the LVL2 or HVL2 groups. It might suggest an emerging bilingual task efficiency advantage that disproportionately impacts the beginning of the trial but cannot yet be sustained throughout the trial. Moving on to letter fluency, the predicted outcome based on prior research was that HVL2ers should match monolinguals in intercept but have a more gradual slope, in line with vocabulary-matched HBs. There was a trend toward such a difference in slope between monolinguals and HVL2ers, though it did not reach significance.

As the only hypothesis predicting higher bilingual word totals than monolinguals, the results of the current study most closely align with the bilingual executive control advantage account.

5.3. Semantic vs. letter/phonemic fluency

The current study found no significant interaction between category and group, but there were consistent main effects of category, indicating differences between semantic and letter/phonemic fluency in general. The following section will break down the differences by measure and how they compare to previous literature.

5.3.1. Word totals

Consistent with previous literature (though see Gollan et al. 2002 who found no category differences), semantic categories overall were more productive than letter/phonemic, though similar to Azuma et al. (1997), there was some overlap between the individual trials such that not all semantic categories were more productive than all letter categories. The particular overlap found here did not entirely match what Azuma et al. reported, though there were similarities. Between the two studies, Animals and S scored near the top while Fruits and the letters F and A scored near the bottom, while the other categories were not tested in both. Within each category type, the order corresponds roughly to the size of the respective exemplar pool for the various categories (Rohrer et al. 1995), such that there are more potential Animal exemplars than Fruits, and S is a more common starting letter for English words than A or F; in contrast, F was significantly more productive in the current study, though A is more common of the two. How the current study's third semantic category, Clothes, fits into the rank order is unclear, though the results suggest a larger pool of Clothing exemplars than Fruits.

5.3.1.1. Impact of trial duration

An important methodological distinction between the current study, which found a bilingual advantage in L1 verbal fluency, and Linck et al. (2009), who found an immersion disadvantage in L1 verbal fluency, was the length of the verbal fluency trial, the typical one minute versus 30 seconds, respectively. To explore the impact of the difference in trial duration on the results of the current study, word totals were adjusted to account for only the first 30 seconds and were then reanalyzed in a two-way mixed ANOVA. Given the equivalent retrieval latencies across groups, which did not provide evidence of retrieval slowing, it was not expected that the word total trend would reverse to evidence disadvantaged L1 retrieval. Results followed the same trend as the minute-long data. There was a significant main effect of category, $F(3, 118) = 296.335$, $p < .001$, in that semantic fluency was more productive than letter/phonemic fluency. There was also a significant main effect for group, $F(3, 118) = 2.903$, $p = .038$, partial $\eta^2 = .069$. However, none of the pairwise comparisons were significant. Though the interaction of category and group was not significant, $F(3, 118) = .761$, $p = .518$, exploratory univariate tests found a significant group effect only for semantic fluency, $F(3, 118) = 3.348$, $p = .021$, with significant differences between monolinguals and HVL2ers ($p = .039$), aligned with the results from the minute-long word totals, while no significant group effect was found for letter/phonemic fluency, $F(3, 118) = 1.103$, $p = .351$.

Thus, it appears that the latter half of the minute-long trial does somewhat strengthen the performance effect, especially for the letter/phonemic trials. Luo et al. (2010) and Friesen et al.’s (2015) finding of greater word totals combined with longer SRLs was interpreted to indicate that HB production is maintained more efficiently later into the minute. However, the current study failed to find group differences in retrieval latency, so the portion of the minute studied should not impact the results. Though the trend is the same in the 30-s as in the 1-m trial, the loss of some statistical differences implies that the latter half of the minute is crucial, even though the differences in retrieval latency did not reach significance.

To further explore that possibility, the word totals from the final 30 seconds were analyzed in a separate two-way mixed ANOVA. Results found no significant main effect of category or interaction of group and category, but the main effect of group was again significant, $F(3, 118) = 3.593$, $p = .016$, partial $\eta^2 = .084$. This time, pairwise comparisons found significant differences between the HVL2ers and both the monolinguals ($p = .028$) and the LVL2ers ($p = .036$). Though the interaction was not significant, exploratory univariate tests found the opposite pattern of the data from the first 30 seconds, namely a significant group effect for letter/phonemic fluency, $F(3, 118) = 3.9520$, $p = .010$, though only between the HVL2ers and monolinguals, but no significant effect for semantic fluency, $F(3, 118) = 2.087$, $p = .106$, lending support to idea that the latter half of the minute impacts performance totals especially for letter/phonemic fluency, even though retrieval latencies were statistically equivalent. This distinction is highlighted by the time-course graphs, where visually (though not statistically) the HVL2 group's best fit line is above the monolinguals' throughout the minute in the semantic graph, while the differences become more pronounced with time in the letter/phonemic graph.

5.3.1.2. Error rates

The categorized error rates in the current study contrast with some previous work. Semantic fluency overall resulted in significantly fewer errors than letter/phonemic fluency, in line with the findings of Sandoval et al. (2010), but distinct from other studies finding either more errors in semantic fluency (Gollan et al. 2002) or no difference (monolinguals in Sandoval et al. 2010). Surprisingly, Gollan et al. (2002) also found more intrusions than repetitions, though according to their error classification, "morphologically related" words were counted with intrusions rather than repetitions, which likely accounts for much of the difference. As well, it is unclear whether a broad definition of morphologically related, such as the one described by Ardila et al. (1994) or a more narrow definition such as the one used in the current study (i.e. restricted to suffixes that do not change the word class, e.g. say, says, stick, sticks) was used. However, these differences would increase the likelihood of inflating the letter/phonemic errors. Therefore, the contrasting pattern of results is more likely attributable to differences not in error classification and counting, but in the categories selected for testing, which included the three studied here but also more intricate categories, such as 'countries in Europe,' 'musical instruments,' and 'things with wheels.' The categories presented in Sandoval et al. (2010) were more complex than the norm both in semantic fluency (e.g. ‘things that cost < $1,’ ‘spices,’ ‘airplane trip medications’) and letter/phonemic fluency (double-letter combinations like ‘ex,’ ‘lu,’ ‘sn’), which may account for why no differences were found for the monolinguals, while perhaps the HBs found the double-letter categories difficult. No comparison of error type was provided.

5.3.2. Word frequency

In the current study, the average word frequency of semantic fluency was significantly lower than letter/phonemic, in contrast with Sandoval et al. (2010), who found the opposite pattern. Differences in category likely account for that switch. Double-letter categories such as those used in Sandoval et al. may have prevented participants from producing very high-frequency two- and three-letter words, while in contrast, two of their semantic categories, which are typically restricted to concrete nouns, were Adjectives and Function Words, the latter of which is particularly likely to result in abnormally high frequency. Sandoval et al. did not disaggregate the mean word frequency of the individual trials. In the current study, unlike for the word totals, there was no overlap between semantic and letter/phonemic categories; all letters resulted in words of higher mean frequency than the semantic trials. The letter with the highest mean frequency per million was A, likely due to words like “and” (28,767.93), “as” (6,933.46) or “at” (5596.03), which as function words, may have contributed to the higher semantic frequency found by Sandoval et al.

5.3.3. Response latencies

FRLs, according to Rohrer et al. (1995), measure the “processing of the retrieval cue and the initiation of the search process” (p.1131). Interestingly, no statistical differences between category types were found for FRLs, indicating that the initiation of retrieval was equivalent for both semantic and letter/phonemic categories. This finding replicates the results from Sandoval et al. (2010) who also found no statistical difference between categories and presumably Luo et al. (2010), who did not report a main effect for category.

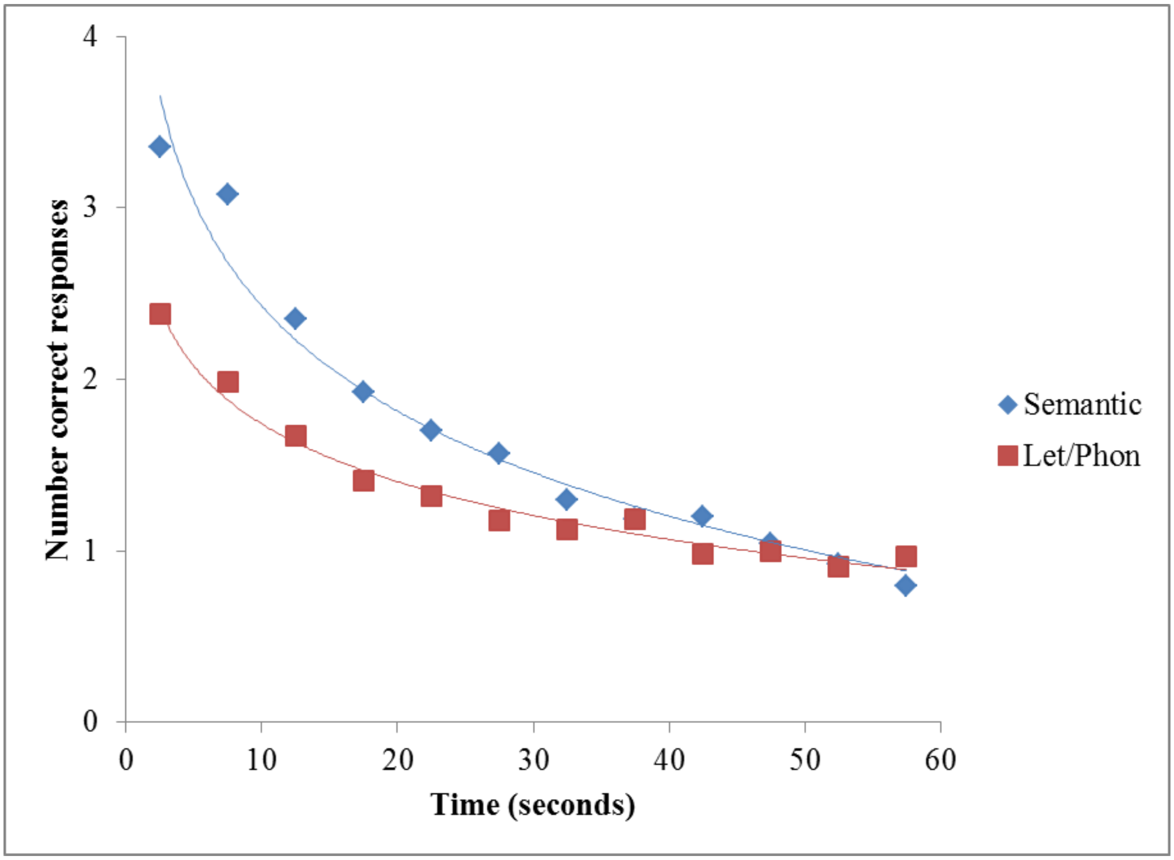
While FRLs measures the initiation of the search, mean SRLs measures the search process itself, and represent the spread of exemplars over time. Shorter mean SRLs indicate production that was concentrated toward the beginning of the trial, while longer mean SRLs represent production that spread later into the trial. In the current study, mean SRLs did differ significantly by category type, in that the mean latency was shorter for semantic fluency than letter/phonemic fluency. According to Rohrer et al. (1995) retrieval latency should increase “when either the size of the search set increases or the duration of each random sampling increases” (p.1130). Letter categories have larger exemplar pools than semantic categories, so search set size could reasonably account for those differences.

However, larger search set sizes are also correlated with greater word totals (Borkowski et al. 1967; Crowe 1998; Rohrer et al. 1995), while in the current study, as is typical, more words were produced during aggregated semantic trials than letter/phonemic trials. According to Friesen et al. (2015):

> Longer mean subsequent response latency indicates that performance extends later into the time course, but the interpretation of this variable depends on the total number of correct responses. If one group produces more correct responses than another group and has a longer mean subsequent-response latency, then the interpretation is that the group has superior control (and equivalent or better vocabulary) and could continue generating responses longer. If one group produces fewer or equivalent correct responses but has longer mean subsequent-response latency, then the interpretation is that the control is more effortful because it took longer to generate the same or a fewer number of items. (p.242)

Expanding Friesen et al.'s interpretation across groups to category types, longer SRLs combined with fewer total words is an indicator of retrieval difficulty; therefore, it is more likely that the processing of semantic fluency is faster and less effortful than letter/phonemic fluency. To further investigate the combination of word totals and SRLs, mean production in both categories was graphed to compare the time-course (Fig. 5.1.). Visually, the intercept is higher for semantic fluency. Since letter/phonemic categories are larger than semantic categories, the intercept cannot be attributed to differences in size, as Luo et al. (2010) and Friesen et al. (2015) argue for LVHBs. Therefore, the shorter mean SRL for semantic fluency is due to a higher concentration of semantic words at the beginning of the trial rather than greater control during the letter/phonemic trial, reflecting the increased difficulty of the letter/phonemic fluency task. This interpretation aligns with results reported by Friesen et al. (2015), who reported the same main effect of category type, attributing it to "the more demanding nature" (p.244) of letter/phonemic fluency.

Figure 5.1 Time-course of semantic vs. letter/phonemic retrieval across groups



5.3.4. Summary

The combination of results in both the previous studies and the current investigation highlight the importance of considering semantic and letter/phonemic fluency separately. The overall impression comparing semantic and letter/phonemic fluency in the current study is that the letter/phonemic fluency is a more difficult task, resulting in fewer words produced, more errors, longer mean SRLs, and a lower time-course intercept.

5.4. Limitations and directions for future study

It is important to note the limitations of the current study and how they can be addressed with future research. One limitation is differences in and difficulty controlling for education level. While efforts were made to control for education by requiring all participants to have completed at least some postsecondary schooling, the nature of finding highly proficient L2 bilinguals resulted in that group containing a disproportionate number of participants pursuing or having attained advanced degrees. Monolingual norms that take education into account (Tombaugh et al. 1999) do not distinguish between education levels beyond secondary, which could be interpreted as evidence that differences beyond that point (e.g. between Bachelors and Masters degrees) are minimal and nonsignificant. However, it is suboptimal to make conclusions based on the absence of analysis. To account for this, reported education levels were submitted to regression analyses, and in none of the instances were they found to be significant predictors of verbal fluency performance. It is possible, notwithstanding, that the nominal measure, which was chosen to simplify the question from the perspective of the participant, was not sensitive enough, or that the unequal distribution of education across groups disguised an effect. Therefore, future research should control educational attainment even more strictly or include it as a variable in the study design.

There is reason to believe, however, that education differences are unlikely to be the cause of the differences reported in the current dissertation. Though the HVL2ers have the highest concentration of advanced degrees, the monolinguals had the most diverse educational attainment, including several with doctoral or professional degrees, while it was the LVL2ers whose educational attainment was the lowest on average, as they were primarily undergraduates in intermediate Spanish classes. If education were driving performance, differences would be less likely to occur between monolinguals and HVL2ers and more likely to be found between the HVL2ers and LVL2ers.

Furthermore, while monolingual norms exist for word totals, there is no established norm for retrieval latency. Longer mean SRLs combined with increased word totals are interpreted as evidence of "superior control (and equivalent or better vocabulary)" (Friesen et al. 2015, p. 242). As there are no norms for how education impacts retrieval latency, the current project's finding of increased word total spread equivalently over time may be indicative of differences in education—those with higher education have a more efficient lexical retrieval system (though direction of potential causation would be difficult to determine). According to the random-search model as described by Rohrer et al. (1995), the most likely impact of education on retrieval latency is that more highly educated monolinguals have longer mean SRLs than less educated monolinguals, due to having a larger vocabulary pool (though see the above reservations regarding extrapolating the effect of category size to lexicon size in latency predictions).

Additional limitations of the current study are the confounds of years of L2 study and L2 proficiency, and L2 attainment and innate individual differences. Learning a second language formally, with study focused after the critical period, is a long process. The nature of any learned skill is that ability comes with time and practice, so it is unsurprising that the HVL2 group had spent more time learning Spanish than the LVL2 and MVL2 groups. Therefore, it cannot be determined from the current results whether the advantage in word production evidenced by the HVL2 group is a function of their L2 proficiency, their length of study, or a combination of the two factors. Similarly, given that the current study is cross-sectional and not longitudinal, a causative relationship between L2 proficiency and lexical retrieval ability cannot be concluded with certainty; in other words, whether the word total advantage found for the HVL2ers is a result of their advanced L2 proficiency and accompanying executive control and task efficiency advantage or instead reflects an innate retrieval advantage that perhaps contributed to their successful acquisition of the L2 at the advanced level. While the former confound is by and large unavoidable, longitudinal research following verbal fluency performance as L2 proficiency develops would help establish whether any gains are caused by L2 attainment or whether the L2 attainment itself is indicative of prior nonlinguistic ability.

Additional opportunities for future research are suggested by unexpected results that did not align with the predictions, to confirm the trends evidenced here. For instance, further investigation with highly proficient L2ers in both semantic and letter/phonemic fluency is needed to confirm whether the word total advantage, if replicated, extends across verbal fluency categories rather than being restricted to letter/phonemic fluency as has been found for HBs (Luo et al. 2010; Friesen et al. 2015). Additionally, the finding that MVL2ers had a higher intercept and steeper slope than monolinguals necessitates further research with this L2 ability group to better determine whether this result is an artifact of the specific participants in the group or whether the trend is reliable. In contrast with Van Assche et al. (2013), the current study found little evidence of asymmetric cross-linguistic inhibition, possibly due to the halved group sizes resulting from each L2er group being divided by counterbalanced language testing order; follow-up study with larger group sizes is important to further investigate how the order of performance affects verbal fluency. On the other hand, especially given the surprising result of longer FRL when tested in English first, additional studies should test L1 verbal fluency in a more monolingual environment, without concurrent L2 testing, to examine the hypothesis that the expectation of L2 production was the cause of delayed L1 retrieval.

Finally, there is still much that can be learned about both bilingual and monolingual lexical retrieval from verbal fluency research. Within HB populations, the recent research concluding executive control advantages has been done with groups representing a diverse array of non-target (L1) languages, and more research with HBs who share a common L1 needs to be done to confirm the results of the extralinguistic measures (retrieval latencies, time-course, etc.). Relatedly, for both HB and L2 populations, additional language pairs ought to be investigated beyond the usual English and Spanish (and recently, Dutch, Van Assche et al. 2013 and Swedish, Ljungberg et al. 2013). For instance, Van Assche et al. (2013) looked at Chinese-English HBs and found different patterns of inhibition than Dutch-English, highlighting the importance of deliberately diversifying language pairs to assess the impact of language similarity on performance.

5.5 Conclusions

The purpose of the current study was to examine whether and how L2 learning and attainment across proficiency levels impacts productive lexical retrieval in the dominant L1. Verbal fluency tasks were employed to measure the speed and ability to retrieve words from the lexicon according to a given cue. A variety of theoretical accounts have been proposed to account for the mixed results found in a modest but growing body of literature on bilingual verbal fluency. Poorer performance on the part of bilinguals has been attributed variously to retrieval slowing, from cross-linguistic competition or accumulated frequency effects, and reduced knowledge of target language vocabulary (Bialystok et al. 2008; Gollan et al. 2002; Rosselli et al. 2000). On the other hand, equivalent or better performance by bilinguals has been attributed to executive control advantages that improve task management (Ljungberg et al. 2013; Luo et al. 2010). Beyond the total number of words produced in a given trial, additional measures include word frequency and retrieval latency, and combinations of results from these measures have been put forth as predictions distinguishing between the above theoretical accounts (Sandoval et al. 2010; Friesen et al. 2010; Luo et al. 2010).

Prior research on bilingual verbal fluency can be divided by population, with the majority examining HBs and a smaller subset looking at L2ers. Importantly, it is mainly the literature studying HBs that has compared participants with monolinguals or examined performance measures beyond word totals. The current study sought to address that relative gap in the literature by comparing the performance of English-Spanish L2 learners/bilinguals to English monolinguals in multiple verbal fluency measures, including word totals, word frequency, and retrieval latency (first-response, subsequent-response, and time-course).

HB and L2er populations are distinct in various ways that may contribute to differences in verbal fluency performance. HBs have been managing two languages for all or most of their lives, learned the L2 (or majority language co-L1 in the case of simultaneous bilingualism) naturally, mainly through immersion, and though many HBs end up dominant in the majority language, it is not inconsequential that they are being tested in their L2. Meanwhile, L2 learners have spent relatively less time bilingual, learned the L2 formally, maintain immersion in an L1 environment, and are being tested in their dominant L1. These conditions combine to make competition from the non-target language relatively less likely for L2ers compared to HBs. Simultaneously, it also means that L2ers tend to have had less lifetime practice activating and suppressing either language, the exercise of which skill is thought to be the impetus behind advantages in executive control. This work therefore adds to the existing knowledge on how bilingualism impacts the productive lexical retrieval, measured by verbal fluency, of a population with a substantially different acquisition experience than the previously studied HBs, allowing for the examination of how factors like age and format of acquisition modulate the cognitive consequences of bilingualism.

The results of the current study suggest that, yes, formally acquired L2 bilingualism can affect L1 lexical retrieval even outside of direct immersion contexts. The effect appears to hinge on L2 ability, in that only the highest proficiency group evidenced differences from the monolingual controls. In line with vocabulary-matched HBs, HVL2ers produced more words than monolinguals, suggestive of a bilingual advantage conferred by advanced L2 attainment. In contrast, production did not differ in overall latency, meaning that HVL2ers produced more words in an equivalent amount of time. While the bilingual advantage in HBs has manifested as enhanced task endurance, the ability to continue producing words longer into the trial, HVL2ers have shown signs of enhanced task efficiency. As the first study to compare monolinguals and L2ers of multiple proficiency levels and across a variety of verbal fluency measures, these results await confirmation from subsequent study; however, the findings are encouraging in their implication that advanced bilingual proficiency can result in lexical retrieval advantages even in the context of formal L2 acquisition.

Appendix A. Multilingual Picture Vocabulary Test stimuli

#	ENGLISH TARGET	SEMANTIC	PHONEMIC	NEUTRAL
1	hand	ear	house	cup
2	dog	cat	dock	pencil
3	tree	flower	train	shirt
4	bed	stool	ball	foot
5	door	window	dollar	beach
6	sun	planet	stain	drawer
7	book	card	bell	brush
8	butterfly	centipede	button	plate
9	scissors	knife	shirt	fire
10	key	locker	koala	jacket
11	chair	table	chick	eye
12	moon	star	mop	bottle
13	plane	bus	plant	arm
14	apple	carrot	apron	frog
15	fish	pig	fist	suit
16	grapes	noodles	geese	leaves
17	horse	cow	honey	bike
18	drum	harp	duck	heart
19	glove	hat	gun	envelope
20	lightbulb	switch	leg	sink
21	cake	donut	cape	pillow
22	watch	ring	window	pen
23	bear	fox	bean	cherry
24	fork	spoon	frog	lamp
25	hat	sock	harp	bench
26	leaf	stick	lips	bee
27	tie	belt	teapot	notebook
28	candle	lantern	candy	frog
29	basket	cradle	bunny	couch
30	clown	acrobat	car	belt
31	kite	balloon	key	rooster
32	rainbow	shadow	radio	onion
33	witch	vampire	wasp	sheep
34	seesaw	tunnel	seatbelt	tub
35	flashlight	hammer	fire	rabbit
36	cloud	rain	canoe	worm
37	iron	toaster	icicle	cookie
38	feather	scale	fence	wolf

#	ENGLISH TARGET	SEMANTIC	PHONEMIC	NEUTRAL
39	peacock	ostrich	pliers	bow
40	bridge	railroad	branch	crown
41	bone	cell	bowl	swing
42	snail	oyster	sleigh	mirror
43	zipper	buckle	zebra	knob
44	lock	chain	lobster	whistle
45	whale	dolphin	wheel	lung
46	nurse	dentist	nutmeg	bee
47	cage	fence	canoe	tongue
48	arrow	bullet	arch	cheese
49	rake	shovel	rope	dart
50	saw	axe	shield	lizard
51	nest	hive	noose	swan
52	plug	socket	plunger	owl
53	wig	beard	wick	sword
54	screw	wrench	sink	ladybug
55	king	queen	kitten	spider
56	scarf	coat	scroll	vulture
57	well	dam	web	cricket
58	dustpan	vacuum	drill	whistle
59	parachute	hang glider	pendulum	footprint
60	blinds	awning	banner	stapler
61	hinge	knob	helmet	whip
62	funnel	grater	flask	splinter
63	gauge	metronome	gavel	latch
64	porthole	rudder	propeller	match
65	anvil	drill	anchor	coffin
66	mortar	beaker	mantle	trenchcoat
67	axle	spoiler	ark	easel
#	**SPANISH TARGET**	**SEMANTIC**	**PHONEMIC**	**NEUTRAL**
1	mano	oreja	manta	taza
2	perro	gato	pantalón	lápiz
3	árbol	flor	ala	camisa
4	cama	sillón	caja	moto
5	puerta	ventana	playa	fresa
6	sol	pájaro	sal	cajón
7	libro	tarjeta	leche	cepillo
8	mariposa	ciempiés	muñeco	helado
9	tijeras	cuchillo	tortuga	fuego
10	llave	taquilla	lluvia	chaqueta
11	silla	mesa	sapo	ojo

#	SPANISH TARGET	SEMANTIC	PHONEMIC	NEUTRAL
12	luna	estrella	loro	botella
13	avión	camioneta	ajedréz	brazo
14	manzana	zanahoria	maleta	ducha
15	pez	cochino	pan	traje
16	uvas	hojas	uña	aguja
17	caballo	vaca	calendario	pie
18	tambor	arpa	tiburón	corazón
19	guante	abrigo	gaviota	sobre
20	bombilla	interruptor	biberón	lavabo
21	pastel	regalo	pastilla	almohada
22	reloj	anillo	raíz	bolígrafo
23	oso	zorro	ola	cereza
24	tenedor	cuchara	trenza	hamaca
25	sombrero	calcetín	semilla	cuaderno
26	hoja	palo	hormiga	cinturón
27	corbata	falda	cuaderno	gaviota
28	vela	farol	volcán	lobo
29	canasta	maleta	cascabel	ascensor
30	payaso	acróbata	pasillo	gallo
31	papalote	globo	pastilla	oveja
32	arco iris	sombra	abanico	cebolla
33	bruja	vampiro	bigote	pata
34	balancín	arena	baúl	cabra
35	linterna	martillo	lentes	conejo
36	nube	lluvia	nariz	bañera
37	plancha	tostador	peine	mosca
38	pluma	escama	pelota	gusano
39	pavo real	avestruz	pincel	lazo
40	puente	ferrocarril	piedra	corona
41	hueso	célula	huevo	pulpo
42	caracol	ostra	codo	espejo
43	cremallera	hebilla	cuna	rana
44	candado	mirilla	caruaje	pulmón
45	ballena	delfín	baranda	pitillo
46	enfermera	jeringa	escalera	abeja
47	jaula	cerca	jabón	queso
48	flecha	bala	foca	lengua
49	rastrillo	pala	riñón	cisne
50	serrucho	hacha	sartén	araña
51	nido	colmena	nuez	dardo
52	enchufe	regleta	encías	lagarto

#	SPANISH TARGET	SEMANTIC	PHONEMIC	NEUTRAL
53	peluca	barba	palanca	espada
54	tornillo	pico	tobillo	búho
55	rey	bufón	rueda	arbusto
56	bufanda	gorro	buzón	mariquita
57	pozo	dique	pila	buitre
58	recogedor	aspiradora	resorte	grillo
59	paracaídas	globo	percha	fideos
60	persiana	toldo	paraguas	huella
61	bisagra	manija	batidor	látigo
62	embudo	rallador	esponja	astilla
63	manómetro	transportador	manivela	aldaba
64	portilla	timón	puchero	cerilla
65	yunque	taladro	yema	ataúd
66	mortero	crisol	murciélago	trinchera
67	eje	resalte	escoba	caballete

References

Aitchison, J. (2008). *The Articulate Mammal: An introduction to psycholinguistics.* New York: Routledge.

Ardila, A., Rosselli, M., & Puente, A. E. (1994). *Neuropsychological evaluation of the Spanish speaker*. New York: Springer.

Azuma, T. et al. (1997). Comparing the difficulty of letter, semantic, and name fluency tasks for normal elderly and patients with Parkinson's Disease. *Neuropsychology, 11*, 488-497.

Baayen, R. H., Piepenbrock, R., & Glikers, L. (1995). The CELEX lexical database [CD-ROM]. University of Pennsylvania, Philadelphia: Philadelphia Linguistic Data Consortium.

Baus, C., Costa, A., & Carreiras, M. (2013). On the effects of second language immersion on first language production. *Acta Psychologica, 142,* 402-409.

Berg, T., & Schade, J. (1992). The role of inhibition in a spreading-activation model of language production. I. The psycholinguistic perspective. *Journal of Psycholinguistic Research, 21*(6), 405-434.

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition, 12*(1), 3-11.

Bialystok, E., Craik, F.I.M., & Luk, G. (2008). Lexical access in bilinguals: Effects of vocabulary and executive control. *Journal of Neurolinguistics, 21*, 522-538.

Bialystok, E., & Feng, X. (2009). Language proficiency and executive control in proactive interference: Evidence from monolingual and bilingual children and adults. *Brain & Language, 109*, 93-100.

Bice, K., & Kroll, J. F. (2015). Native language change during early stages of second language learning. *NeuroReport*, *26*(16), 966-971.

Blumenfeld, H.K. & Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes, 22*(5), 633-660.

Borkowski, J.G., Brinton, A.L., & Spreen, O. (1967). Word fluency and brain damage. *Neuropsychologia, 3*, 135-140.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of verbal learning and verbal behavior*, *5*(4), 325-337.

Butman, J. Allegri, R.F., Harris, P., & Drake, M. (2000). Fluencia verbal en español: Datos normativos en Argentina. *Medicina, 60*, 561-564.

Carlson, S.M., & Meltzoff, A.N. (2008). Bilingual experience and executive functioning in young children. *Developmental science*, *11*(2), 282-298.

Cave, C. B. (1997). Very long-lasting priming in picture naming. *Psychological science*, *8*(4), 322-325.

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent? *Journal of Memory and Language, 45*, 721-736.

Coltheart, M. (1981b). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

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.

Costa, A., Santesteban, M. & Ivanova, I. (2006). How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1057-1074.

Costa, A., Miozzo, M., & Caramazza, A. (1999). Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and Language, 41*, 365-397.

Crowe, S.F. (1998). Decrease in performance on the verbal fluency test as a function of time: Evaluation in a young healthy sample. *Journal of Clinical and Experimental Neuropsychology, 20*(3), 391-401.

Davis, C.J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*(1), 65-70.

de Bot, K, & Schreuder, R. (1993). Word production and the bilingual lexicon. In R. Schreuder & B. Weltens (Eds.), *The Bilingual Lexicon* (pp.191-214). Amsterdam: John Benjamins Publishing Co.

De Houwer, A. (2013). The absolute frequency of maternal input to bilingual and monolingual children: A first comparison. In T. Grüter & J. Paradis (Eds.) *Input and Experience in Bilingual Development* (pp. 37-57). Amsterdam/Philadelphia: John Benjamins

de Picciotto, J. & Friedland, D. (2001). Verbal fluency in elderly bilingual speakers: Normative data and preliminary application to Alzheimer's Disease. *Folia Phoniatrica et Logopaedica, 53*, 145-152.

Dijkstra, A. (Ton), & Van Heuven, W.J.B. (1998). The BIA model and bilingual word recognition. In J. Grainger & A.M. Jacobs (Eds.), *Localist connectionist* *approaches to human cognition* (pp. 189-225). Mahwah, NJ: Lawrence Erlbaum Associates.

Dijkstra,A. (Ton) & Van Heuven, W.J.B. (2002). The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, *5*(3), 175-197.

Duchon, A., Perea, M. Sebastián-Gallés, N., Martí, A. & Carreiras, M. (2013). EsPal: one-stop shopping for Spanish word properties. *Behavioral Research Methods, 45*(4), 1246-1258.

Durso, F.T., & Johnson, M.K. (1979). Facilitation in naming and categorizing repeated pictures and words. *Journal of Experimental Psychology: Human Learning and Memory, 5*(5), 449-459.

Fay, D. & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry, 8*(3), 505-520.

Finkbeiner, M., Gollan, T.H., & Caramazza, A. (2006). Lexical access in bilingual speakers: What's the (hard) problem? *Bilingualism: Language and Cognition, 9*(2), 153-166.

Fox, E. (1996). Cross-language priming from ignored words: Evidence for a common representational system in bilinguals. *Journal of Memory and Language, 35*, 353-370.

Friesen, D., Luo, L., Luk, G. & Bialystok, E. (2015). Proficiency and control in verbal fluency performance across the lifespan for monolinguals and bilinguals. *Language, Cognition and Neuroscience, 30*(3), 238-250.

Gerfen, C., Tam, J., McClain, R.R., Linck. J.A., & Kroll, J.F. (in preparation). *Evidence for inhibition in native language production during immersion in the second language.*

Gollan, T.H., & Ferreira, V.S. (2009). Should I stay or should I switch? A cost-benefit analysis of voluntary language switching in young and aging bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 640-665.

Gollan, T.H., Montoya, R.I., & Werner, G.A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology, 16*, 562-576.

Gollan, T.H., Montoya, R.I., Cera, C., & Sandova, T.C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language, 58*, 787-814.

Gollan, T.H. Montoya, R.I. Fennema-Notestine, C, & Morris, S.K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & cognition, 33*, 1220-1234.

Gollan, T.H., Weissberger, G.H., Runnqvist, E. Montoya, R.I. & Cera, C.M. (2012). Self-ratings of spoken language dominance: A multilingual naming test (MiNT) and preliminary norms for young and aging Spanish-English bilinguals. *Bilingualism: Language and Cognition*, *15*(3), 594-615.

González, H.M., Mungas, D., Haan, M.N. (2005). A semantic verbal fluency test for English- and Spanish-speaking older Mexican-Americans. *Archives of Clinical Neuropsychology*, *20*, 199-208.

Green, D.W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition, 1*, 67-81.

Harley, T.A. (1999). Will one stage and no feedback suffice in lexicalization? [Peer commentary on "Lexical access in speech production" by Levelt et al.] *Behavioral and Brain Sciences, 22*(1), 45.

Hilchey, M.D. & Klein, R.M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review,18*(4), 625-658.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39,* 1-27.

Ivanova, I., Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica, 127*, 277-288.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicologica, 35*, 49-66.

Kaushanskaya, M., Gross, M., & Buac, M. (2014). Effects of classroom bilingualism on task-shifting, verbal memory, and word learning in children. *Developmental science*, *17*(4), 564-583.

Kempler, D., Teng, E. L., Dick, M., Taussig, I. M., & Davis, D. S. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of the International Neuropsychological Society*, *4*(06), 531-538.

Kleinman, D. & Gollan, T.H. (2016). Speaking two languages for the price of one: Bypassing language control mechanisms via accessibility-driven switches. *Psychological Science, 27(5),* 700-714.

Kormi-Nouri, R., Morado, A., Morado, S., Akbari-Zardkhaneh, S., & Zahedian, H. (2012). The effect of bilingualism on letter and category fluency tasks in primary school children: Advantage or disadvantage? *Bilingualism: Language and Cognition*, *15*(2), 351-364.

Kroll, J.F., Bogulski, C.A., & McClain, R. (2012). Psycholinguistic perspectives on second language learning and bilingualism. *Linguistic Approaches to Bilingualism*, *2*(1), 1-24.

Kroll, J.F., & Dijkstra, A. (Ton). (2002). The bilingual lexicon. In R.B. Kaplan (Ed.) *The Oxford Handbook of Applied Linguistics.* (pp. 301-321). Oxford: Oxford University Press.

Kroll, J.F. & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149-174.

Kroll, J.F., van Hell, J.G., Tokowicz, N., & Green, D.W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism (Cambridge, England), 13*(3), 373-381.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44,* 325-343.

Levelt, W.J.M. (1992). Accessing words in speech production: Stages, processes and representations. *Cognition, 42*(1), 1-22.

Levelt, W.J.M. (1993). Lexical access in speech production. In E. Reuland & W. Abraham (eds.) *Knowledge and Language, Volume 1* (pp. 241-251). Netherlands: Kluwer Academic Publishers.

Levelt, W.J.M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*(23), 13464-13471.

Levelt, W.J.M., Roelofs, A. & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22, 1-75.*

Lezak, M.D. (1983). *Neuropsychological assessment,* 2nd ed. New York: Oxford University Press.

Linck, J.A., Kroll, J.F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language evidence for the role of inhibition in second-language learning. *Psychol Sci, 20,* 1507-1515.

Ljungberg, J.K., Hansson, P., Andrés, P., Josefsson, M., Nilsson, L. (2013). A longitudinal study of memory advantages in bilinguals. *PLoS ONE, 8*(9), e73029.

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition, 114,* 29-41.

Macnamara, J., & Kushnir, S.L. (1971). Linguistic independence of bilinguals: The input switch. *Journal of Verbal Learning and Verbal Behavior*, *10*(5), 480-487.

Mahon, B.Z., Costa, A., Peterson, R., Vargas, K.A., & Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503-535.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*(2), 97-115.

Martin, A., Wiggs, C.L., Lalonde, F., & Mack, C. (1994). Word retrieval to letter and semantic cues: A double dissociation in normal subjects using interference tasks. *Neuropsychologia, 32*(12), 1487-1494.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological review, 88*(5), 375-407.

McGill, W.J. (1963). Stochastic latency mechanisms. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.) *Handbook of mathematical psychology, Vol. 1* (pp. 309-360). New York: Wiley.

Meuter, R.F.I., & Allport, A. (1999). Bilingual language switching in naming: Asymmetrical costs of language selection. *Journal of Memory and Language, 40*, 25-40.

Misra, M., Guo, T., Bobb, S.C., & Kroll, J.F. (2012). When bilinguals choose a single word to speak: Electrophysiological evidence for inhibition of the native language. *Journal of Memory and Language, 67*(1), 224-237.

Mitchell, D.B., & Brown, A.S. (1988). Persistent repetition priming in picture naming and its dissociation from recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 213-222.

Neuman, E., McCloskey, M.S., & Felio, A.C. (1999). Cross-language positive priming disappears, negative priming does not: Evidence for two sources of selective inhibition. *Memory & Cognition, 27*(6), 1051-1063.

O'Seaghdha, P.G. (1999). Parsimonious feedback. [Peer commentary on "Lexical access in speech production" by Levelt et al.] *Behavioral and Brain Sciences, 22*(1), 51-52.

Paap, K.R., & Greenberg, Z.I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive psychology*, *66*(2), 232-258.

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

Penfield, & Roberts, (1959). *Speech and Brain.* Princeton, NJ: Princeton University Press.

Piatt, A.L., Fields, J.A., Paolo, A.M., & Tröster, A.I. (1999a). Action (verb naming) fluency as an executive function measure: convergent and divergent evidence of validity. *Neuropsychologia, 37*, 1499-1503.

Piatt, A.L., Fields, J.A., Paolo, A.M., Koller, W.C., & Tröster, A.I. (1999b). Lexical, semantic, and action verbal fluency in Parkinson's Disease with and without dementia. *Journal of Clinical and Experimental Neuropsychology*, *21*, 435-443.

Poarch, G.J., & van Hell, J.G. (2012). Executive functions and inhibitory control in multilingual children: Evidence from second-language learners, bilinguals, and trilinguals. *Journal of Experimental Child Psychology*, *113*(4), 535-551.

Portocarrero, J.S., Burright, R.G., & Donovick, P.J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology, 22* 415-422.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*(1), 36-57.

Prior, A. & MacWhinney B. (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, *13*(2), 253-262.

Ransdell, S.E., & Fischler, I. (1987). Memory in a monolingual mode: When are bilinguals at a disadvantage? *Journal of Memory and Language*, *26*, 392-405.

Roberts, P.M., & Le Dorze, G. (1997). Semantic organization, strategy use, and productivity in bilingual semantic verbal fluency. *Brain and Language, 59*, 412-449.

Rohrer, D., & Wixted, J.T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*(5), 511-524.

Rohrer, D., Wixted, J.T., Salmon, D.P. & Butters, N. (1995). Retrieval from semantic memory and its implications for Alzheimer's disease. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1127-1139.

Rosselli, M., Ardila, A., Araujo, K., Weekes, V.A., Caracciolo, V., Padilla, M. & Ostrosky-Solis, F. (2000). Verbal fluency and repetition skills in healthy older Spanish-English bilinguals. *Applied Neuropsychology, 7*(1), 17-24.

Rosselli, M., Ardila, A., Savatierra, J., Marquez, M., Matos, L., & Weekes, V.A. (2002). A cross-linguistic comparison of verbal fluency tests. *International Journal of Neuroscience, 112*, 759-776.

Sadat, J., Martin, C.D., Alario, F.X., & Costa, A. (2012). Characterizing the bilingual disadvantage in noun phrase production. *Journal of Psycholinguistic Research, 41*, 159-179.

Sagarra, N., & Ellis, N.C. (2013). From seeing adverbs to seeing verbal morphology. *Studies in Second Language Acquisition, 35*, 261-290.

Sandoval, T.C., Gollan, T.H., Ferreira, V.S. & Salmon, D.P. (2010). What causes the bilingual disadvantage in verbal fluency? The dual-task analogy. *Bilingualism: Language and Cognition, 13*(2), 231-252.

Schmid, M.S., & Köpke, B. (2009). L1 attrition and the mental lexicon. In A Pavlenko (Ed.) *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 209-238). Brisol, UK: Multilingual Matters.

Schwieter, J.W., & Sunderman, G. (2008). Language switching in bilingual speech production: In search of the language-specific selection mechanism. *The Mental Lexicon*, *3*(2), 214-238.

Sebastián-Gallés, N., Martí, M.A., Cuetos, F., & Carreiras, M. (2000). LEXESP: Léxico informatizado del español. Barcelona: Edicions de la Universitat de Barcelona.

Shao, Z., Janse, E., Visser, K., & Meyer, A.S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology, 5,* 1-10.

Seibert Hanson, A., & Carlson, M.T. (2014). Global effects: The roles of first language and proficiency in L2 processing of Spanish clitics. *Language Learning, 64*(2), 310-342.

Soares, C. & Grosjean, F. (1984). Bilinguals in a monolingual and a bilingual speech mode: The effect of lexical access. *Memory & Cognition*, *12*(4), 380-386.

Snodgrass, J.G. & Tsivkin, S. (1995). Organization of the bilingual lexicon: Categorical versus alphabetic cuing in Russian-English bilinguals. *Journal of Psycholinguistic Research, 24*(2), 145-163.

Spivey, M.J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281-284.

Tombaugh, T.N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology, 14*, 167-177.

Van Assche, E., Duyck, W., & Gollan, T.H. (2013). Whole-language and item-specific control in bilingual language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1781-1792.

Van Hell, J. & De Groot, A.M.B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition, 1*(3), 193-211.

Vorberg, D. & Ulrich, R. (1987). Random search with unequal search rates: Serial and parallel generalizations of McGill's Model. *Journal of Mathematical Psychology, 31*, 1-23.

Wixted, J.T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*(1), 89-106.