Spring 2015

The processing of formulaic language on elicited imitation tasks by second language speakers

Yan Xun
Purdue University

Graduate School Form 30
Updated 1/15/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Xun Yan

Entitled

THE PROCESSING OF FORMULAIC LANGUAGE ON ELICITED IMITATION TASKS BY SECOND LANGUAGE SPEAKERS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

April Ginther
Chair

Okim Kang

Tony Silva

Margie Berns

Elaine Francis

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): April Ginther

Approved by: Nancy Peterson 4/17/2015

Head of the Departmental Graduate Program Date

THE PROCESSING OF FORMULAIC LANGUAGE ON ELICITED IMITATION TASKS BY SECOND LANGUAGE SPEAKERS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Xun Yan

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2015

Purdue University

West Lafayette, Indiana

For my love, 相濡以沫，不离不弃。

For my parents, 你们辛苦了。

For my sister, 勇敢寻找属于自己的人生方向！

ACKNOWLEDGEMENTS

This dissertation would not have been possible without help, advice, and support from a number of people to whom I am very much thankful.

First and foremost, I would like to express my deep gratitude to April Ginther, who has been a wonderful advisor, collaborator, and friend. She has guided me throughout my PhD study: supporting me on the training of Rasch modeling, involving me in the development of post-entry English placement test, and collaborating with me on various research projects. I have benefited tremendously as a researcher from her mentorship and friendship. My thanks also go to Nancy Kauper, for her training at the Oral English Proficiency Program (OEPP) testing office. I would not have grown as a language tester without the support from April and her.

I thank my committee members, Margie Berns, Tony Silva, Elaine Francis, and Okim Kang for their insightful comments on this dissertation and guidance throughout the research process. I also would like to acknowledge Yukiko Maeda and Jing Lv, for their substantial contribution to the systematic review in my dissertation study.

Additionally, I feel very grateful to my friends and colleagues at Purdue: especially to Cong Zhang, Chienyu Wu, and Yue Chen, for their company and support during the process of writing this dissertation. I am also very fortunate to form friendship with Mu-sheng Lin, Aung Kyi San, Yushan Fan, Lee Jung Huang, Veronika Maliborska,

Megha Anwer, and Dallas Woodburn, because of whom my PhD study at Purdue was filled with happy memories.

Last, but most importantly, I want to thank my partner, Wutthiphong Laoriandee, for taking great care of me during my PhD study. He has taught me to smile at setbacks and remain positive towards uncertainties of the future. Our four years in West Lafayette have been the best part of my life. Without his support, I would not have been able to complete my PhD degree.

"Everything goes quiet when it's you I am with." –Jason Mraz

TABLE OF CONTENTS

Page

Page

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
AWL	Academic Word List
COCA	Corpus of Contemporary American English
EAP	English for Academic Purposes
ESL	English as a Second Language
EI	Elicited Imitation
FS	Formulaic Sequence
LTM	Long-term Memory
NumSP	Number of Silent Pause
SL	Sentence Length
SLA	Second Language Acquisition
AR	Articulation Rate
SP	Silent Pause
STM	Short-term Memory
TOEFL	Test of English as a Foreign Language
WM	Working Memory

ABSTRACT

Yan, Xun. Ph.D., Purdue University, May 2015. The Processing of Formulaic Language on Elicited Imitation Tasks by Second Language Speakers. Major Professor: April Ginther.

The present study investigated the processing of formulaic language, in an effort to examine how the use of formulaic language may or may not contribute to second language (L2) fluency in speaking performance. To examine the effect of formulaic language on L2 fluency, this study utilized elicited imitation (EI) tasks designed to measure general English language proficiency in order to compare repetition of individual sentences containing formulaic sequences (FS) to repetition of sentences that do not. In addition to the presence of FS, the length of stimuli sentences was manipulated and compared to a second independent variable. Responses to EI tasks were automatically measured for articulation rate (AR) and number of silent pauses (NumSP), two important measures of L2 fluency. Repeated measures ANOVAs were conducted to examine the main and interaction effects of FS and sentence length (SL) on AR and NumSP.

Results of analyses of EI performances showed that both SL and FS had a significant effect on L2 fluency in speech production; however, these two variables had differential effects on AR and NumSP. SL had a strong effect on NumSP on EI performances: as the stimulus sentence becomes longer, NumSP on EI performances

increases. The presence of FS had a larger effect on AR than on NumSP: higher proportion of formulaic sequences in language use contributes to faster articulation rate, while the processing advantage of formulaic sequences helps reduce the number of silent pauses when the processing load is large.

Findings of this study suggest that the presence of formulaic sequences create a processing advantage for L2 speakers and that EI tasks prompt language comprehension and processing. Findings have important implications for language teaching and assessment, in particular with respect to the teaching of formulaic sequences and the use of EI as a measure of L2 proficiency. Recommendations for future research of formulaic sequences and development of EI tasks are discussed.

CHAPTER 1. INTRODUCTION

The real-time ability to process the English language plays a foundational role in academic socialization and success for second language (L2) speakers in a university context (Cho & Bridgeman, 2012; Graham, 1987; Vinke & Jochems, 1993; Wait & Gressel, 2009; Xu, 1991). Research in adult ESL education over the past three decades has shown that language proficiency is positively correlated with ESL students' academic success (Al-Musawi & Al-Ansari, 1999; Cho & Bridgeman, 2012; Graham, 1987; Sharon, 1972; Wimberley, McCloud, & Flinn, 1992). However, many L2 speakers are at a real disadvantage in both basic interpersonal and academic communications due in part to a lack of fluency or automaticity in processing language in real-life situations. Although L2 speakers may be comparable to their first language (L1) English speaking peers in terms of foundational academic aptitude or knowledge, many may not be able to communicate efficiently and, as a result, may not be well rewarded for the time and effort they invest in academic study (Johnson, 1988). Inadequate language proficiency may slow down ESL students' academic socialization and even lead to failure to fulfill graduation requirements on time (Light, Xu & Mossop, 1987).

L1 speakers often make use of formulaic language to achieve the efficiency of communication and socialization (Pawley & Syder, 1983). In terms of formulaic language, or formulaic sequences, refers to the use of preconstructed phrases or multiword strings that occur so frequently in language use that these word strings are argued to be processed as single units (Wray, 2002). The presence of formulaic language in everyday and academic conversations allows the speaker to process and produce language at faster rates and contributes to a variety of effects. For L2 speakers, mastery of formulaic language is a key aspect of high level of language proficiency (Pawley & Syder, 1983). From a cognitive perspective, the use of formulaic sequences can significantly reduce the processing load on working memory, thus enabling the speaker to produce language more fluently. Moreover, as formulaic sequences are idiomatic and fixed, i.e., shared within a speech community, mastery and use of these sequences may reduce listener effort in conversation, thereby facilitating communication efficiency and efficacy.

From a language socialization perspective, formulaic language performs important social functions in interactions in various social contexts. The use of formulaic sequences marks identity and membership within a particular speech community, facilitates new members to gain access into the community, and enhances their communication with other members.

Undergraduate English as a second language (ESL) students are frequently involved in a variety of social activities. These activities prompt them to socialize with their L1 English speaking peers and become familiar and comfortable with the new environment. Under these circumstances, the ability to process (i.e., to understand and

use) formulaic language may help ESL students quickly adjust to the new environment and become more confident in interactions with their L1 speaking peers. Therefore, the ability to process and produce formulaic sequences is a skill as important for L1 speakers as it is for L2 speakers.

Second language researchers have studied formulaic language as a phenomenon for 30 years, but the research focus on formulaic language has been shifting. In the areas of second language acquisition (SLA) and English for academic purposes (EAP), there has been a recent increase in research efforts given to the identification and instruction of formulaic language (see below). From a SLA standpoint, it is important to examine how L2 learners process formulaic language in not only the receptive mode (e.g., reading) but also in the productive mode (e.g., speaking). Based on these considerations, this dissertation study examined the effects of formulaic sequences on the L2 fluency of undergraduate ESL students enrolled in a large public university in the US.

Elicited imitation (EI) or sentence repetition is a popular psycholinguistic measure of language proficiency which has been widely used to examine both L1 and L2 proficiency and development. Despite the fact that EI has been customized to measure an array of language-related constructs, the employment of EI to investigate the processing of formulaic language has mostly been in L1 research (e.g., Tremblay, Derwing, Libben & Westbury, 2011). To date, there has not been a published study that uses EI to examine how L2 speakers process formulaic language. This study investigated the extent to which the presence of formulaic language facilitates responses of L2 speakers to elicited imitation tasks across different task conditions (i.e., length bands of stimulus sentences to be repeated). Findings of this study add to our understanding of the acquisition of

formulaic language by L2 speakers as well as the usefulness of EI tasks to measure formulaic language acquisition and L2 proficiency.

CHAPTER 2. REVIEW OF RELEVANT LITERATURE

The conceptual framework of this study has been influenced by theoretical discussions of L2 fluency, the acquisition of formulaic language, and information-processing models of SLA.

2.1 Development of Fluency in L2 Speaking Performance

As an important criterion used to describe speaking performance, fluency has been conceptualized and operationalized in various ways in the literature of first and second language acquisition (Ginther, Dimova, & Yang, 2010; Lennon, 1990; Schmitt-Gevers, 1993). In terms of conceptualization, Lennon (1990) characterized fluency into two categories: *broad* and *narrow*. In the broad sense, fluency, synonymous with the overall proficiency of a speaker, is an all-encompassing term that covers a range of speech features such as rate, accuracy, complexity, coherence, and even idiomaticity. In contrast, the narrow approach views fluency as "one, presumably isolatable, component of oral proficiency" (Lennon, 1990, p.389), i.e., speech rate and smoothness (often related to pausing). When investigating pausing patterns as a proxy for smoothness of speech, researchers often make a distinction between expected and unexpected pauses. Expected pauses are pauses that occur at predictable places, i.e., pauses that occur at syntactic or semantic boundaries (also referred to as *juncture pauses* by Hawkins (1971)).

Expected pauses mark the processes of sentence parsing and planning that occur in fluent speech. In contrast, unexpected pauses include both pauses that occur within syntactic or semantic units (or *non-juncture pauses* in Hawkins' terms) and particularly long silent pauses at syntactic or semantic boundaries. Unexpected pauses are argued to mark labored sentence processing and planning and often occur in speech produced by speakers of lower language proficiency (Anderson-Hsieh and Venkatagiri, 1994; Cenoz, 1998).

While the components of fluency in its broad sense, i.e., accuracy, and complexity are difficult to capture, the advantage of adopting the narrow approach is that fluency can be relatively easily measured. Indeed, many empirical studies (e.g., Ginther et al., 2010; Kormos & Denes, 2004; Lennon, 1990) have found strong correlations between temporal measures of fluency (the narrow sense) and the holistic ratings of fluency and overall proficiency awarded by human raters (the broad sense). These strong correlations indicate that speech rate and pausing patterns tend to co-vary with other linguistic features and can thus be regarded as reliable proxies for the overall proficiency of a speaker. The present study adopts the narrow approach to the examination of fluency, focusing on temporal measures of fluency (i.e., speech rate and silent pauses), not only because of the relative ease of quantifying and analyzing temporal measures but also due to the strong correlations between fluency measures in the broad and narrow senses.

Development of a high level of fluency is an important, albeit controversial, aspect of language proficiency and is often a common goal for language learners acquiring an additional language. However, the development of fluency in advanced L2 learners is not simply a matter of increasing speech rate (Fillmore, 1979), but rather a

matter of acquiring the formulaicity or proceduralization of linguistic knowledge that results in increased speech rate and the perception of fluency (Towell, Hawkins, & Bazergui, 1996). When discussing what they called "native-like" linguistic capacities, Pawley and Syder (1983) present two arguments that are important to understanding native-like fluency and selection, and the connection between the two: 1) native-like fluency does not mean few pauses, but is marked instead by few unexpected pauses; 2) procedural knowledge plays a role in both native-like fluency and native-like selection. To Pawley and Syder, procedural knowledge, a key aspect of native-like linguistic knowledge, involves mastery of a bank of idiomatic and formulaic expressions (they called "memorized sentences" and "lexicalized sentence stems", p. 205) that are easily selected and easily chained to create fluent idiomatic output.

Their arguments, originally proposed as a counter argument to generativist or syntactic (rule-governed) perspectives to language acquisition, offer a functional or lexical (input-based, computational or statistical) approach to the development of language proficiency and fluency. The premise of their argument is that language learning is not necessarily usage-based, but rather use-based (the lexical approach, including the information-processing or connectionist perspectives to SLA, will be further discussed in Chapter 3 in conjunction with the measurement of implicit grammatical knowledge). That is, high language proficiency is not only marked by creativity, i.e., sentence construction based on syntactic rules; but also by formulaicity (Wray, 2002), i.e., sentence construction based on lexis or the use of formulaic sequences. They argue that explaining fluency and selection requires an underlying system that is based on a combination of rate-based and lexical elements as components.

2.2 Formulaic Language and SLA

2.2.1 Definition and characteristics of formulaic language

Formulaic language as a long-recognized linguistic phenomenon represents the level of "fixedness" rather than "creativity" in language use (Wray, 2002). Formulaic language is ubiquitous in communication of all sorts and is regarded as a characteristic that marks high language proficiency or fluency.

Formulaic language has been variously referred to as *formulae* (Coulmas, 1979), *formulaic sequences* (Schmitt, 2004), *prefabricated patterns* (Hakuta, 1974), *idioms* (Lewis, 2009), *collocations* (Lewis, 2000), *lexical bundles* (Biber, Conrad, & Cortes, 2003), *multiword sequences* (Butler, 2003). There is no standard definition that encompasses all the linguistic phenomena covered under formulaic language (Wray, 2012). Perhaps the most widely cited term and definition of formulaic language is *formulaic sequences*, the one proposed by Wray (2002):

> A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (p. 9).

Wray's definition is an attempt for inclusiveness. As her definition implies, formulaic sequences cover a wide range of structures and word units, but formulaic sequence need not be a whole sentence or a set idiomatic phrase as is commonly understood (e.g., raining cats and dogs). On the contrary, a formulaic sequence could be any form that "lies

on the borderline between bound forms and words, or between words and phrases" (Bloomfield, 1993).

In spite of the terminological variation, there have been several characteristics of formulaic sequences that can be established in the extant literature (Schmitt, 2004), which include the following:

- Formulaic sequences appear to be stored as holistic units, but they may not be acquired in an all-or-nothing manner;
- Formulaic sequences can have slots that enable flexibility of use, but the slots typically have semantic constraints;
- Formulaic sequences are often additionally marked as prosodic units.

Formulaic sequences are often tied to particular conditions of use. (pp. 4-9)

These characteristics highlight two fundamental principles in the identification of formulaic sequences: fixedness in structure, and holistic storage and retrieval in processing. The fixedness of formulaic sequences has been examined through recurrence of word sequences or frames, mostly through a corpus-based approach where computational algorithms are created to identify words that tend to co-occur across utterances, contexts, time, and interlocutors. Literature in both L1 and L2 research has witnessed a fairly large number of efforts to establish the collocational patterns of words and phrases within a particular corpus (e.g., Biber, Conrad & Cortes, 2004; Hyland, 2008; Schmitt, 2004; Simpson-Vlach & Ellis, 2010).

As compared to fixedness, the examination of the holistic processing (or the processing advantages) of formulaic sequences is often faced with greater challenges largely due to the difficulty in measuring holistic processing (Schmitt, Grandage &

Adolphs, 2004). The most common approach researchers tend to adopt to measuring the processing advantages of formulaic sequences is embedding those sequences in individual sentences, paragraphs, or even longer texts and then measuring the rate and accuracy of processing of these texts in comparison with comparable texts that do not contain formulaic sequences. If the speaker can process texts with formulaic sequences at a higher rate and with greater accuracy, then researchers have inferred that the speaker processes the formulaic sequences holistically.

An examination of the literature shows that the processing of formulaic sequences has been examined more in the reception mode than in the production mode. In the reception mode, researchers have utilized self-paced timed reading tasks (e.g., Conklin & Schmitt, 2008; Tremblay et al., 2011; Reali & Christiansen, 2007) and eye-tracking techniques (e.g., Underwood, Schmitt & Galpin, 2004) to measure whether readers process formulaic sequences faster by recording reading speed and eye movement associated with formulaic sequences in text. Others have used grammaticality judgment tasks (e.g., Jiang & Nekrasova, 2007) to ask participants to rate on the acceptability of formulaic sequences as compared to nonformulaic sequences.

In the production mode, Nekrasova (2009) used gap-filling and dictation tasks to measure whether participants can complete or reproduce in writing the formulaic sequences in the stimuli. In terms of speaking, Tremblay *et al* (2011) was the only attempt to use sentence recall or EI tasks to examine whether they present a processing advantage to L1 speakers. However, of the few studies published on the processing advantages of formulaic sequences as compared to regular word sequences, the majority has focused on L1 speakers, but not L2 speakers. Moreover, there has not been any study

that examined the processing advantage of formulaic sequences for L2 speakers in the speaking mode, although a few studies have shown a positive relationship between the use of formulaic sequences and holistic ratings of oral proficiency (e.g., Boers, Brussels, Kappel, Stengers & Demecheleer, 2006; Ushigusa, 2007).

2.2.2 Significance of formulaic language acquisition

As suggested by Wray (1998), the system of human language is marked by an "uneasy" balance between formulaicity and creativity:

> Without the rule-based system, language would be limited to repertoire, clichéd, and, whilst suitable for certain types of interaction, lacking imagination and novelty. In contrast, with only a rule-based system, language would sound pedantic, unidiomatic and pedestrian (pp. 64-65).

However, maintaining the balance is challenging for L2 speakers especially if they learn the L2 mainly through explicit instruction of syntactic rules with limited opportunities to use language in authentic social contexts. Even for advanced L2 speakers at the university-level, who have threshold levels of English language skills as measured by standardized English exams, the adjustment to the idiomatic expressions prevalent in everyday and academic conversations can be very difficult (Wimberley, McCloud, & Flinn, 1992). Learners may choose to avoid acquiring or using formulaic sequences especially when they are facing semantic difficulties with the sequences (e.g, formulaic sequences that do not have counterparts in learners' L1, Dagut & Laufer; 1985; figurative phrasal verbs or idiomatic expressions, Liao & Fukuya, 2004).

The assumption underlying the acquisition of formulaic language is that lexis and grammar are not completely separated in the process of language learning and use, which entails the accumulation and processing of linguistic units larger than individual morphemes or words. Language users have available a bank of formulaic chunks; therefore, their language production is not always a process of building sentences word by word. Sinclair (1987) used the *idiom principle* and the *open choice principle* to describe the common pattern of language production, arguing that language production is characterized by frequent alternations between the two principles, and language users may apply the idiom principle before the open choice principle. In other words, language users prefer formulating utterances at the multiword level (which I refer to as making use of *the idiom principle*) and will break phrases or multiword strings down to the individual word level only when necessary (which I refer to as making use of *the open choice principle*).

We find many applications of this principle in our daily lives: military commands, aviation English, etc. Within these contexts, simple and highly formulaic phrases can effectively solve communication problems especially when multiple interlocutors are involved. Misuse of these formulaic sequences can lead to serious or even fatal consequences. Although it could be argued that our daily conversation, students' interactions at school in particular, does not always occur as a life-or-death situation, the efficiency gained from using formulaic language facilitates students' access to information and other resources. Undergraduate students are frequently exposed to a variety of social settings, all of which tend to have associated sets of registers for different communicative purposes. Registers tend to manifest through word choices and

formulaic sequences. Therefore, it is important for undergraduate ESL students to understand and fluently use such formulaic sequences in both oral and written forms across university contexts.

The significance of formulaic language cannot be understood fully outside the context of communication or interaction. Successful communication rests on meanings that are mutually intelligible to both the hearer and the speaker. According to the Relevance Theory (Sperber & Wilson, 1996, p. 474), the listener will habitually extend only the minimum processing effort necessary to comprehend and interpret an utterance, thereby freeing up limited resources for other tasks. Therefore, part of the effort on the part of the speaker should be directed toward the use of expressions to minimize listener effort or, as Wray (2002) puts it, to "corner the hearer into maximum likelihood of getting, and reacting to, the message" (p. 94). Naturally, the use of prefabricated formulae that are shared by speakers within a community is a major contribution to comprehension and communication efficiency.

The functions of formulaic language, according to Wray (2002), include "the reduction of the speaker's processing efforts, the manipulation of the hearer (including the hearer's perception of the speaker's identity), and the marking of social discourse" (p. 101). Therefore, the ability to use formulaic language is of importance to L2 speakers from both cognitive and social-cultural perspectives. On one hand, formulaic language is able to support both the speaker's and the listener's processing simultaneously (Wray, 2002, p. 93). From a cognitive perspective, in a conversation, the interlocutors need to rely on syntactic, prosodic, and pragmatic cues for the management of turns, e.g., signaling and projecting in advance the completion of a turn (Ford & Thompson, 1996;

Fox, 2001; Stivers & Robinson, 2006). The facilitative effect of linguistic resources in the management of turns in conversation coincide with the functions of formulaic language in language processing., use of formulaic language helps reduce processing load of the speaker and thus contributes to a speaker's temporal fluency and automaticity when composing (McLaughlin, Rossman, & McLeod, 1983; Pawley & Syder, 1983; Ushigusa, 2008). Likewise, formulaic language also facilitates the listener's comprehension of speech and helps them remain focused on the content rather than on the form (because the expressions are formulaic), which will "greatly enhance the success of the messages' interactional purpose" (Wray, 2002, p.99).

Furthermore, from a social-cultural perspective, formulaic language has a facilitative effect on language socialization. Formulaic language, albeit a less-frequently discussed notion in the literature of language socialization, "plays a crucial role in socializing novices to social dimensions such as politeness, hierarchy, and social identities including social roles and statuses, and relationships" (Burdelski & Cook, 2012). Formulaic language signifies the speaker's identity as an individual and/or as a social member. During social interactions, formulaic language can be used in both normative and novel ways to help the speaker build, maintain, or change various kinds of relationship with other members within a particular community.

However, the acquisition of formulaic language is not an easy task for many L2 speakers. After examining the extant literature on formulaic language and second language acquisition, Wray (2002) found that L2 speakers tend to face difficulties in the acquisition and use of formulae, as storage and automatic processing of formulaic sequences is associated with a number of interrelated variables, such as exposure to

formulaic language, and language proficiency level. Some manifestations of the difficulties include over-reliance on a restricted range of formulae, the use of non-idiomatic but creative collocations, and poor control over the grammaticality of formulaic language. More effort is needed to contribute to research on the acquisition of formulaic language in order to facilitate the development of fluency and socialization for L2 speakers in the target language.

2.3 Information Processing Models of SLA

This study is also inspired by information-processing models of SLA. The idea of information processing represents the dominant approach in cognitive psychology to explaining how the brain's processing mechanisms (i.e., memory) function in the process of learning (including language acquisition). In general, the approach likens the mind to a computer information processor and assumes that complex behavior builds on simple processes.

Information-processing models investigate how memory stores, retrieves, or transforms information and how information is automatized and restructured through repeated activation (Huitt, 2003). Most information-processing models make a distinction between short-term memory (STM) and long-term memory (LTM). STM, also called working memory (WM), stores information temporarily (15 seconds) and has limited storage capacity and the processing of information in STM is more controlled. LTM, in contrast, stores information, skills, and procedural knowledge permanently that can be automatically retrieved when needed. Any information transferred from STM to LTM will gain a place in permanent storage.

There are four most widely accepted information-processing models in cognitive psychology: the stage model (Atkinson & Shriffin, 1968), the levels-of-processing model (Craik & Lockhart, 1972), the parallel-distributed processing model (Rumelhart, Hinton, & McClelland, 1986), and the connectionistic model (Rumelhart & McClelland, 1986). While these models diverge in their hypotheses of how information is stored and retrieved in the memory (for discussion of each model, see Huitt, 2003), they share a few fundamental assumptions:

- The capacity of the mental system (i.e. memory) is limited in the sense that the amount of information that can be actively processed by the system at a given point in time is constrained.
- A control mechanism is required to oversee the encoding, transformation, processing, storage, retrieval and utilization of information.
- Human beings use a two-way information process to construct meaning about the world: bottom-up processing (store information from senses) and top-down processing (retrieve information already stored in memory).
- Human organisms are genetically prepared to process and organize information in specific ways.

(Huitt, 2003)

Some influential SLA theories derived from the information-processing approach include Shiffrin & Schneider's (1977) model of automatic vs. controlled information processing, Anderson's (1983, 1985) Active Control of Thought (ACT) model, and its application in the understanding of learner strategies (O'Malley & Chamot, 1990) and

development of fluency (Towell & Hawkins, 1994), and Levelt's (1989) model of language production. These theories are built upon two fundamental assumptions, which help form the theoretical basis of the present study: 1) linguistic information is processed in either a controlled or an automatic manner; 2) practice or repetition of processes is a key component of language learning.

Learning, especially in the sense of achieving automaticity (also referred to as automatization (McLaughlin, 1987; 1990) or proceduralization (O'Malley & Chamot, 1990; Towell & Hawkins, 1994)), is seen as a transfer from controlled processing (in STM) to automatic processing (in LTM). When it comes to language production, a controlled process refers to production that is built at the level of individual words or morphemes whereas an automatic process occurs when sentences are constructed upon chunks or wordstrings without much attentional control on individual words and morphemes. During the initial stages of learning, learners must rely on controlled processing to process and produce the target linguistic structures. Such processing is constrained by the limitations of STM or WM. That is, the production is usually not automatic (fluent), and the structures are easily forgotten. Then, through repeated activation or practice, the structures become automatized or proceduralized and are stored as whole units in LTM. In this way, as the situation requires, these structures can be retrieved with little attentional control from the learner.

In the same line of reasoning, information or knowledge can be classified into declarative knowledge (knowledge of *what*) and procedural knowledge (knowledge of *how*) (Anderson, 1983; 1985). A similar attempt to define procedural knowledge, Levelt (1989), in his model of language production, uses the word *lexicon* to refer to an

independent module that stores all of the (procedural) linguistic information the speaker needs for formulating the message. The speaker can easily access this module at either the formulation or comprehension stages of communication. In spite of the terminological variation, there is much agreement in the differences between declarative knowledge and procedural knowledge. First, language production using declarative knowledge, especially at the beginning of the learning process, tends to require attentional control of linguistic information; however, once linguistic knowledge becomes proceduralized, processing of linguistic information becomes more automatic. Furthermore, when applying Anderson's model to the development of fluency, Towell and Hawkins (1994) argue that formulaic language—once learned—is usually stored as procedural knowledge in LTM, which can be either retrieved automatically to create fluent speech runs or reanalyzed to add creativity into language use under controlled processes. However, such flexibility is not possible with only declarative knowledge.

Automaticity or formulaicity in language production as a construct of language proficiency, albeit discussed in theoretical models of communicative competence, has not been well integrated with the practicalities of test construct, task characteristics, or performance measurement (van Moere, 2012). However, if fluency is a major concern in L2 learners' language proficiency, then automaticity, which contributes to fluency, should be incorporated in performance assessments of language proficiency.

CHAPTER 3. ELICITED IMITATION AS A MEASURE OF L2 PROFICIENCY

Elicited imitation is a method that usually requires participants to listen to a series of stimulus sentences (or phrases, words, sounds) and then repeat—to their best ability—the sentences verbatim (Underhill, 1987). EI features simple and economical administration procedures. In addition, when used to assess language performance, EI allows the developers and researchers to customize the target component of language proficiency and the difficulty of the tasks by manipulating or controlling the sentence stimuli (Hood & Schieffelin, 1978). The simplicity and flexibility in task development and administration makes EI adaptive to both classroom and standardized assessments and a valuable tool in exploratory research (e.g., Henning, 1983; Markman, Spilka, & Tucker, 1975; van Moere, 2012).

As a measure of language proficiency, EI has been widely used to investigate L1 development (e.g., Fraser, Bellugi, & Brown, 1963; Slobin & Welsh, 1973 for reviews of the use of EI in L1 acquisition), language disorders in children (e.g., Dailey & Boxx, 1979), and neuropsychological activities (e.g., Menyuk, 1964). The underlying assumption of testing with EI is that if the participant has acquired the grammatical features associated with or displayed in the stimuli, it should be easy to repeat the stimuli. Otherwise, repetition will be difficult (Rebuschat & Mackey, 2013).

The simple and flexible characteristics of EI have led to wide variation in the design of EI tasks. Task variation, in turn, presents challenges in applying findings of EI studies into practice in order to enhance L2 learning, teaching, and assessment; therefore, this chapter presents a systematic review of EI used in L2 research. The purpose of this review is to (1) examine the historical and current state of the development, administration, and use of EI tasks, (2) clarify the construct measured by EI, and, more importantly, (3) advance discussions toward a more principled practice of EI in L2 research. This review corresponds with Norris and Ortega (2006) who argued that research synthesis on language tasks can contribute to the appropriate use of available language testing instruments in the field of language learning and teaching. Prior to this systematic review, Zhou (2012) reported a synthesis of 24 studies using EI on L2 adult learners and concluded that EI is overall a reliable measure (internal consistency coefficient ranged from .78 to .96, p. 90). In addition, the correlation between EI scores and other measures of language proficiency was higher than .5 in the majority of the studies (p. 90) reviewed, which provides some support for the construct-related validity for EI as a measure of language proficiency.

This synthesis took a different approach to the examination of the construct validity of EI from that employed by Zhou (2012) by placing additional emphasis on the theoretical question of what EI measures and the inclusion of a discussion which I hope may enable more principled design of EI tasks. More specifically, the historical review reported in *Phase I* outlines debate on the authenticity and the construct measured by EI, serving as the theoretical basis for the statistical investigations of the meta-analysis reported in *Phase II*. In the meta-analysis, I was interested in whether scores on EI tasks

can effectively distinguish between higher and lower proficiency learners. If EI is an effective measure of language proficiency, higher and lower proficiency learners (e.g., L1 vs. L2 speakers) should be consistently distinguishable in terms of their performance on EI tasks across studies. The inability of EI tasks to distinguish speakers across proficiency levels would indicate that EI might be measuring something different from language proficiency. In addition, I examined the variation in the design of key EI task features across studies and the impact of different design of EI tasks on the sensitivity of EI to distinguish speakers with different proficiency levels (hereafter referred to as *sensitivity of EI*) because there has not been established standards or protocols to use EI for language testing.

Given the first purpose of this review, I surveyed 76 published and unpublished studies (including all the 24 studies included in Zhou [2012]), where EI was used for measuring L2 proficiency in the period of 1970-2014, to examine the status of EI to measure L2 proficiency in various settings. Specifically, I conducted both a systematic narrative synthesis in the first phase of the review (*Phase I* henceforth) and a quantitative synthesis with a meta-analysis in the second phase of the review (*Phase II* henceforth).

This review systematically investigated: (1) the use of EI with particular respect to the research/assessment context (i.e., target construct, language, and language proficiency levels); (2) variation in the design of certain key features of EI tasks discussed in previous reviews of EI (e.g., Chaudron & Russell, 1990; Vinther, 2002); (3) whether (and to what extent) EI tasks can distinguish between higher and lower proficiency speakers; and (4) whether (and to what extent) the sensitivity of EI differs when different task features are employed.

3.1 Phase I: Narrative Synthesis

The narrative synthesis addresses two specific research questions: (1) the use of EI with particular respect to the research/assessment context (i.e., target construct, language, and language proficiency levels); (2) variation in the design of certain key features of EI tasks discussed in previous reviews of EI (e.g., Chaudron & Russell, 1990; Vinther, 2002).

3.1.1 Method

3.1.1.1 Study selection criteria

The target studies for the narrative synthesis (*Phase I*) included all the studies in the period of 1970-2014 (May 2014), which discussed (in length) or used EI as a method to measure global or specific aspects of L2 proficiency. Reports that only mentioned EI but did not discuss the technique in detail were excluded from this phase (e.g., Rebuschat, 2013). The research synthesis began with Naiman (1974), the first documented application of EI in L2 research to measure linguistic competence of young L2 learners of French. The type of documents collected included journal articles, book chapters, dissertations and theses, conference proceedings, technical reports, and book reviews.

3.1.1.2 Identification of studies

The following steps were taken to locate related EI studies. First, a list of commonly used electronic databases in the fields of applied linguistics and education was

used to search for studies that fit the selection criteria mentioned above. These databases include *Academic Search Premier, Education Source, ERIC, EBSCO, JSTOR, LLBA (Linguistics and Language Behavior Abstracts), ProQest Dissertation and Theses, PsycARTICLES, PsychInfo, SSCI (Social Sciences Citation Index),* and *ScienceDirect* databases. Keywords used to search for studies were the combinations of two phrases: (a) "elicited imitation" or "sentence repetition" or "sentence recall" or "imitation" or "repetition", and (b) "second language" or "foreign language".

Second, both electronic and manual searches were performed for some widely cited journals in applied linguistics and second language acquisition (SLA), including, *Applied Linguistics, Applied Psycholinguistics, CALICO Journal, Computer Assisted Language Learning, Foreign Language Annals, Language Assessment Quarterly, Language Learning, Language Teaching Research, Language Testing, Modern Language Journal*, and *TESOL Quarterly*. Finally, reference lists of the identified reports were also used to locate additional studies that may have related to this synthesis.

The literature search process identified 76 studies that either define or use EI in a way aligned with the definition of EI mentioned above. These include 52 journal articles, 12 dissertations, five conference proceedings, five book chapters, and two book reviews. All 76 studies were used for the narrative synthesis. Thirty studies of the 76 studies were group comparison studies (see Figure 3.1 for a summary of the selection criteria and search process).

3.1.1.3 Coding

The coding process of the primary studies consisted of two stages. First, a preliminary set of coding variables were identified, based on the reviews of EI in Bley-Vroman and Chaudron (1994), Gallimore and Tharp (1981), and Vinther (2002), to represent the research /assessment context (i.e., target language, measured construct, target language proficiency levels) and task features of EI (i.e., number of items, stimuli sentence length, implementation of delayed repetition, scoring method, control of linguistics variables). These variables were used to develop a coding scheme. The coding scheme was piloted on a sample of articles from the 76 studies. These codes were then discussed among the authors of the study and unclear codes were revised. The coding scheme was finalized after three rounds of tryout of actual coding, discussion, and revision. The specific codes for each category of the final coding scheme for this synthesis are presented in Table 3.1 below.

Once the coding scheme was established, it was then made available online to the coders through Qualtrics©, a survey distribution program. All the primary studies (k = 76) were coded independently by the first and third authors of the study (who are graduate researchers in language testing and in educational psychology, respectively). The inter-coder reliability expressed in percent-agreement was 94.68%. Discrepancies between the two coders were identified and discussed.

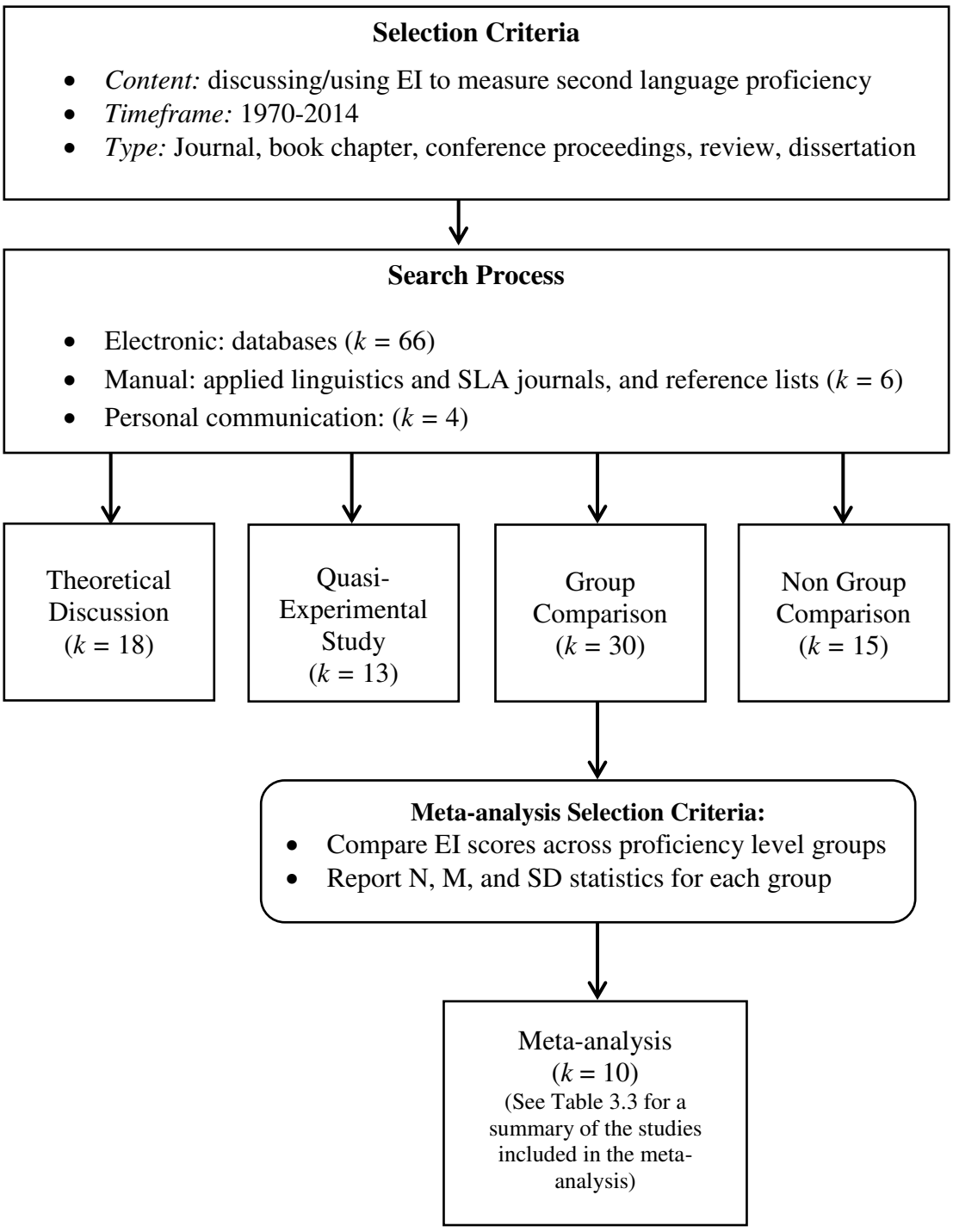


Figure 3.1 Selection and Classification of EI Studies

Table 3.1 Features of EI Tasks Coded for Phase I (k=58[a])

Features	Codes	Frequency
Target proficiency levels	advanced/high	21
	intermediate	18
	beginner/low	21
	other[b]	15
	not mentioned	6
Measured construct	Global	15
	Specific	43
	phonological	7
	syntactic and morphosyntactic	35
	Other	1
Stimuli sentence length	short (7 syllables or shorter)	3
	medium (8 to 15 syllables)	22
	long (16 syllables or longer)	3
	Varied (across two length bands)	21
	No mentioned	7
Delayed repetition	Yes	21
	No	31
	not mentioned	6
Scoring method	binary	24
	ordinal	15
	interval	15
	mixed	3
	not mentioned	1
Control of linguistic variables	grammaticality	22
	syntactic features	36
	phonological features	6
	morphological features	8
	lexical features	9

Note. [a]Articles that had only theoretical discussions were excluded in this table.
[b]Other refers to studies describing language proficiency levels in an institutional approach.

3.1.2 Results

3.1.2.1 Research and assessment contexts of EI studies

The use of EI as an L2 proficiency measure has extended across a variety of languages and linguistic constructs, targeting a range of proficiency levels. While the

majority of the studies used EI to measure performance in English (k = 34), other target languages included French (k = 7), Spanish (k = 7), Dutch (k = 3), Mandarin (k = 3), German (k = 2), and Japanese (k = 2). It should be noted that there is a lack of standardization in the characterization of participants' language proficiency levels in studies published in SLA journals (Thomas, 1994; 2006). The two most commonly used approaches are *institutional* (i.e., grouping based on their assigned curricular or course levels) and *impressionistic* (i.e., grouping based on impressionistic descriptors, e.g., beginner, intermediate, or advanced). When grouped in the institutional approach, language proficiency levels can vary even among L2 learners within the same course or program level (Tremblay, 2011). Given this limitation, this study uses the impressionistic approach; however, studies describing language proficiency levels in an institutional approach were retained and coded as *Other* on the target language proficiency level (see Table 1). Three levels of L2 proficiency were specified to classify participants in this paper: high (advanced), intermediate, and low (beginner). The frequency counts shown in Table 1 are evenly distributed across all three language proficiency levels. When distinguishing L2 speakers across proficiency levels, some studies included L1 speakers as a baseline for comparison (e.g., Erlam, 2006), while others compared EI scores of L2 speakers across proficiency levels (e.g., West, 2012; Wu & Ortega, 2013).

Among the 76 primary studies, 18 studies focused on theoretical discussions about EI with respect to the measured construct and the design of tasks features (hereafter referred to as *theoretical discussion studies*; see Figure 3.1). The other 58 studies used EI tasks to measure a variety of language-related constructs in both experimental and non-experimental settings.

There were 13 quasi-experimental studies that used EI as a learning outcome measure, testing the effect of particular interventions (hereafter referred to as *quasi-experimental studies*). The interventions included among others: form-focused instruction (Fiori-Agoren, 2004; Kim, 2012), strategies of corrective feedback (Ellis, Loewen, & Erlam, 2006; Erlam & Loewen, 2010; Faqeih, 2012; Li, 2010), explicit instruction (Akakura, 2012; Elliot, 1997), and particular types of teaching approaches (Burger & Chretien, 2001; Trofimovich, Lightbown, Halter & Song, 2009; Trofimovich, Lightbown & Halter, 2013). (See Table 3.2, for a summary of experimental studies using EI as a measure of language learning outcome).

Observational studies were classified into two types: *group comparison studies* (k=30) and *non-group comparison studies* (k=15). Group comparison studies featured comparisons of EI scores across selected proficiency levels. In contrast, non-group comparison studies mainly examined the concurrent validity of EI scores as a measure of global language proficiency (e.g., Henning, 1983) by comparing EI scores with scores on other (more established) language proficiency tests such as TOEFL iBT or IELTS (e.g., Erlam, 2006).

Table 3.2 Summary of Experimental Studies Using EI as a Measure of Language Learning Outcome, 1970-2013

Study	Sample size	Target construct	Proficiency level	Target language	Intervention	Results
Elliot, 1997	66 undergraduate students enrolled in Spanish courses in the US (43 experimental, 23 control)	Segments	Intermediate	Spanish	Explicit pronunciation instruction	Formal phonological instruction promotes more accurate Spanish pronunciation.
Burger and Chretien, 2001	30 students enrolled in content-based ESL and FSL courses for psychology in Canada (no control group)	Global proficiency	Advanced	English and French	Content-based ESL and FSL instruction	Students in content-based courses achieved significant improvement in both fluency and accuracy.
Jensen and Vinther, 2003	63 undergraduate students in Denmark (43 experimental, 20 control)	Global proficiency	Intermediate	Spanish	Exact repetition as input enhancement	Exact repetition as input enhancement shows a significant effect on learner's comprehension skills, phonological decoding strategies, and grammatical accuracy.
Fiori-Agoren, 2004	44 undergraduate students enrolled in Spanish courses (27 experimental, 17 control)	Preposition *for* (por/para) Verb form *to be* (ser/estar)	Not specified*	Spanish	Form-focused instruction	Posttest scores revealed significant statistical differences in the outcomes in favor of form-focused instruction over meaning-focused instruction.

Table 3.2 continued.

Study	Sample size	Target construct	Proficiency level	Target language	Intervention	Results
Ellis, Loewen, and Erlam, 2006	34 ESL students enrolled in a private language school in New Zealand (24 experimental, 10 control)	Past tense -ed	Low intermediate	English	Implicit and explicit corrective feedback	Results show a clear advantage for explicit feedback over implicit feedback for both the delayed imitation and grammaticality judgment posttests.
Trofimovich, Lightbown, Halter, and Song, 2009	74 francophone grade 3 students in Canada (49 experimental, 25 control)	Segments and supra -segmentals	Not specified[a]	English	Reading listening comprehension-based learning	No significant difference in terms of learning outcome was observed between comprehension-based and traditional language learning program.
Erlam and Loewen, 2010	50 undergraduate students enrolled in French courses in the US (40 experimental, 10 control)	Noun-adjective agreement	Not specified	French	Explicit and implicit recast	The type of feedback students received did not have a differential impact on learning. Moreover, the provision of feedback did not have a significant effect on learning either.

Table 3.2 continued.

Study	Sample size	Target construct	Proficiency level	Target language	Intervention	Results
Li, 2010	78 undergraduate students enrolled in Chinese courses in the US (57 experimental, 21 control)	classifiers and perfective-le	High and low	Mandarin	Corrective feedback types	Explicit feedback was more effective than implicit feedback for low-level learners, but the two types of feedback were equally effective for more advanced learners.
Akakura, 2012	94 ESL undergraduate students New Zealand (49 experimental, 45 control)	Article	Advanced and intermediate	English	Explicit instruction	Retained effects for explicit instruction were found on both implicit and explicit knowledge of articles.
Faqeih, 2012	64 EFL learners in Saudi Arabia (49 experimental, 15 control)	Modals	Intermediate	English	Corrective feedback types (metalinguistic and recast)	Results suggested that both metalinguistic information and recasts are beneficial for the development of English modals.
Kim, 2012	92 Korean EFL learners (77 experimental, 15 control)	Wh-movement	Intermediate	English	Form-focused instruction	Form-focused instruction positively affects the learning of both explicit and implicit knowledge in the long term. But learners benefit the most from a combination of form-focused and meaning-focused instruction.

Table 3.2 continued.

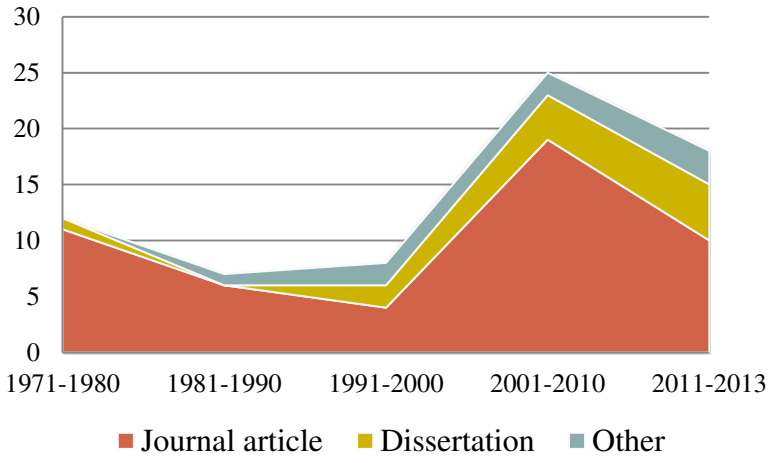
Study	Sample size	Target construct	Proficiency level	Target language	Intervention	Results
Trofimovich, Lightbown, and Halter, 2013	73 francophone grade 3 students in Canada (28 experimental, 25 control) (the same as Trofimovich et al., 2009)	Segments and supra -segmentals	Not specified	English	Comprehension-based learning	Results show an interaction effect between learner background variables and type of instruction. The comprehension-program seems to benefit a certain type of learners.
Campfield and Murphy, 2014	80 polish children with the mean age of 8 years	Supra -segmentals	Not specified	English	Exposure to rhythm-salient input	The findings established a clear link between implicit L2 acquisition and prosody.

Note. [a]Not specified means that these studies used an institutional approach to characterizing participants' language proficiency level.

3.1.2.2 The use of EI as a measure of L2 proficiency: A historical review

3.1.2.2.1 Popularity of EI in the 1970s and 1980s

The use of EI as a measure of L2 proficiency has undergone interesting shifts over the past few decades, and these shifts accompany shifts in the theoretical models of language proficiency and attendant frameworks associated with discussions of validity. Figure 3.2 shows the number of L2 EI studies during the period of 1970-2014. In the 1970s and 1980s when EI first appeared in L2 research, EI tasks were mainly used to address linguistic competence, mostly in terms of assessment of some aspect of grammar (e.g., Markman et al., 1975; Naiman, 1974).



Note. Other includes book chapters, conference proceedings, and reviews.

Figure 3.2 Publication Year Trend by Document Type

However, as structural approaches to defining language proficiency were questioned, growing interest in theoretical and empirical explorations of communicative competence, or more complex and inclusive models of language proficiency were

developed (Bachman, 1990; Bachman & Palmer 1996; Canale & Swain, 1980; Thomas, 1992). Fulcher (2000) referred to this theoretical and methodological shift as the "communicative" movement (p. 483). A concomitant move is that traditional, psycholinguistic language tasks fell out of favor among L2 researchers. An examination of the literature in L2 research reveals that, at least in the 1990s, the use of EI and similar general proficiency and psycholinguistic measures (Oller, 1973, 1976), such as dictation and cloze procedure, decreased. Instead, L2 researchers became more interested in tasks and assessments that emphasized authenticity, interactivity and performance – measures with strong face validity and appear to simulate real-life communication. Despite the former popularity of EI in language related research and its usefulness and reliability, the technique was questioned as a useful representation of language proficiency (Vinther, 2002) as authenticity, interactivity, and performance moved to center stage.

3.1.2.2.2 Debate on the authenticity and construct validity of EI: An interesting shift in the 1990s

A main criticism of EI is related to authenticity. That is, language production prompted through EI tasks is criticized as unrepresentative of natural speech or conversation. In the case of young learners, Hood and Lightbown (1978) observed that repeating the utterances of other interlocutors does not necessarily align with a child's natural speech patterns or production. At the very least, children are less likely to be asked to repeat the utterance of the caregiver. Hood and Schieffelin (1978) also stated that EI "places demands on the child that are not present in the usual interaction between child and adult" (p. 5). However, recently, van Moere (2012) argued for the authenticity

of elicited imitation tasks from the perspective of automaticity in spoken communication. When preparing or giving a response within a conversation, it is necessary that speakers draw on the language used by conversational partners and summarize, even repeat, particular statements. It is therefore reasonable to argue that natural conversation and interaction depend in part on repetition—if not verbatim, then certainly in terms of summary.

A more important criticism of EI stems from the uncertainty of what EI actually measures. In other words, the available literature has not clarified the underlying construct, and research investigating the construct-related validity of EI, with the exception of Zhou (2010), is limited. Debate over and the emphasis on the construct validity of EI reflects the change in traditional views of validity from multiple, complementary, forms of validity (e.g., content, criterion, construct validity) to a contemporary view of validity as a unified concept (e.g., Messick's (1989) Unified Theory of Construct Validity). Though not always explicitly stated, a defining characteristic of a language proficiency measure lies in its ability to assess the participants' linguistic knowledge, i.e., their ability to process linguistic information to construct meaning. However, different views exist as to whether EI can measure one's linguistic knowledge. Some scholars (e.g., Eisenstein, Bailey & Madden, 1982; Naiman, 1974) advocate that EI prompts participants to process the structure and meaning of the sentences. They argue that in order to be able to repeat a sentence, one has to comprehend the meaning of the sentence. Others suspect that EI only prompts parroting, i.e., rote repetition of the chain of acoustic information without comprehension, thus only

measuring the capacity of phonological short-term memory (Gathercole & Baddeley, 1993).

3.1.2.2.3 Resurgence of EI as a measure of implicit grammatical knowledge

Despite the criticisms of EI with respect to authenticity and construct validity, recent literature displays a resurgence of interest in using EI in L2 research; in addition, discussion of the target constructs of EI have shifted from general grammatical competence to roughly an even split between the acquisition of particular linguistic structures and/or implicit grammatical knowledge as the global construct of interest. In terms of specific linguistic constructs, EI tasks have recently been employed to examine L2 speakers' performance on a range of syntactic, morphosyntactic, lexical, and phonological structures across proficiency levels (e.g., Akakura, 2012; Schimke, 2011; Trofimovich et al., 2009; van Boxtel, Bongaerts, & Coppen, 2005; West, 2012), L1 backgrounds (e.g. Verhagen, 2011) within particular teaching or learning contexts (e.g., Akakura, 2012).

The increase in the number of EI studies on implicit grammatical knowledge (e.g., Bowles, 2011; Ellis, 2005; Erlam, 2006; Serafini, 2013) also reflects the theoretical discussions of second language acquisition (SLA) that have emerged in this era. With respect to more cognitive approaches to SLA, investigations of implicit grammatical knowledge presume a statistical, input-based, or usage-based model of language acquisition (e.g., founded on information processing theories and connectionism). The statistical approach to the development of language proficiency postulates that language learning is not different from other types of learning that can be associated with

probabilistic models (Ellis, 2005). Statistical models of language learning place an emphasis on input frequency and experience with the language and their effects on learners' mental representations of linguistic knowledge and automaticity in language comprehension and production (Ellis, 2002). That is, language comprehension and production is largely influenced by learners' lexicons as well as or instead of innate syntactic rules. Learners' mental lexicons, built on their own experiences with language input and output, stores statistical information about behavior (i.e., relative frequency, concurrence patterns, and functional contexts) of lexical items and syntactic structures in the language, allowing them to make predictions or guesses about appropriate use. Statistical information (i.e., frequency) is used to construct a model that allows learners to predict or project words and/or chunks as they comprehend or produce sentences.

These frequency effects (Ellis, 2002), to some extent, align with the socio-cognitive perspective to second language acquisition (Atkinson, 2002), where exposure to input is likened to experience with language "in the world", and implicit knowledge used to govern language processing resembles language that goes on "in the head". The transformation of knowledge from social experiences to mental representations, while approaching second language acquisition from a different perspective, is arguably related to the existence of and transfer between different types of linguistic knowledge (i.e., declarative vs. procedural knowledge, Anderson, 1983, 1985; explicit vs. implicit knowledge, Ellis, 2005) and how those different types of knowledge are stored as mental representations (i.e., in short-term memory (STM) vs. long-term memory (LTM)). According to Huitt (2003), LTM is used to store information, memories, skill sets and procedural knowledge that can be readily retrieved when needed, both voluntarily and

involuntarily. STM is designed to retain information temporarily, after which information is either forgotten or stored permanently in LTM, based on whether repeated exposure is realized and how actively it is used while an input resides in STM. During the initial stages of language learning, language production using declarative (or explicit) knowledge requires attentional control of linguistic information and thereby tends to be more labored; however, once linguistic knowledge becomes proceduralized through repeated exposure, processing of linguistic information becomes more automatic.

The distinction between how different types of knowledge are stored and retrieved can differentiate rote repetition and imitation with comprehension, and thereby inform an argument in favor of the construct validity of EI. According to Gathercole and Baddeley (1993), rote repetition only requires the sentence to be phonologically processed in STM as an acoustic image and repeated without decoding the sentence for meaning. However, the repetition of acoustic images without comprehension, that is, the repetition of meaningless strings of sounds is more difficult than the repetition of a string that is meaningful. Thus, rote repetition tends to be possible only if the sentences are short and/or are continuously rehearsed. In contrast, imitation with comprehension requires the speaker to decode the acoustic information in the sentence, map the sounds onto the corresponding structures and meanings, and eventually convert the selected structures to sounds to reconstruct the same meaning. By doing so, the speaker appeals to internalized or proceduralized linguistic knowledge, which is thought to be permanently stored in LTM and automatically retrieved when needed. In this case, even though a sentence may exceed the capacity of STM, the speaker can access (automatically) linguistic knowledge to aid in repeating the sentence. Because of the comprehension process, the speaker may

paraphrase the original sentences instead of repeating them verbatim, but the capture, access, and transformation of meaning can be assumed to occur only when adequate language resources are available.

Although it is impossible to directly observe how linguistic information is processed, the two oppositional hypotheses of what EI measures can be falsified through differential-population studies (Popham, 2003), e.g., those EI tasks that discriminate individuals with different language proficiency levels. Assuming that EI prompts language comprehension and underlying levels of language proficiency, the repetition would require the participant to decode the structural information of the sentences for meaning. That is, the participant has to map the sounds (the acoustic image) onto the corresponding phonological, lexical, and syntactic knowledge stored in his or her LTM (Naiman, 1974). Therefore, higher proficiency speakers should be able to repeat longer and linguistically more complex sentences because they tend to have internalized more sophisticated grammatical structures. On the other hand, if EI tasks only elicit rote repetition, EI may only measure the capacity of phonological STM. Then, instead of processing the meaning of the sentences, participants, regardless of language proficiency level, can rely on their STM to recall and imitate the chain of sounds and therefore should perform indistinguishably on EI tasks.

3.1.2.3 Four key task features that may affect the construct validity of EI

Along with the resurgence of EI literature in L2 research, researchers have also placed an emphasis on the design of EI tasks in relation to the sensitivity of EI. Literature on STM and EI suggests that rote repetition can happen but only under certain conditions:

(1) sentences are short (Munnich, Flynn & Martohardjono, 1994); (2) repetition takes place immediately after the stimulus (McDade, Simpson & Lamb, 1982); and (3) imitation is continually rehearsed without interruption (Gathercole & Baddeley, 1993). If EI tasks are used to measure language proficiency, these tasks features must be incorporated in a principled manner so that the tasks prompt imitation with comprehension rather than rote repetition.

Vinther (2002), building on previous reviews of EI (e.g., Bley-Vroman & Chaudron, 1994; Gallimore & Tharp, 1981), suggested four key task features that influence the validity of EI tasks as a measure of language proficiency: (a) length of sentence stimuli, (b) delayed repetition, (c) grammatical features of the stimuli, and (d) scoring methods. Control of these variables is likely to increase the sensitivity of EI to discriminate learners on the measured constructs (i.e., language proficiency) and reduce construct-irrelevant variance (i.e., STM capacity).

3.1.2.3.1 Length of sentence stimuli

Sentence length has been frequently observed as a factor that influences the difficulty of EI tasks (e.g., Miller, 1973; Perkins, Brutten & Angelis, 1986). In order to measure language comprehension, i.e., to minimize the effect of working memory, the length of sentence stimuli must exceed the learners' STM capacity. However, L2 researchers have not agreed on cutoffs for an appropriate sentence length – the length that would best discriminate learners at different levels. Regarding the limit of STM, Miller's Law (1956) states that the number of chunks (be they syllables, numbers, words, or sequences) that one can hold in STM is 7 ± 2. The "magic number seven" coincides with

Perkins et al. (1986), who suggest that the length of the sentences be set at seven to eight syllables. However, Naiman (1974) chose sentences of 15 syllables for first- and second-grade L2 learners and considered the length to be appropriate. This choice was also selected in the assessment for adult learners conducted by Eisenstein et al. (1982). Jensen and Vinther (2003) chose even longer sentences, the majority of which exceeded 16 syllables and found that most L1 speakers were capable of repeating the sentences.

In this review, I selected the following cut-offs to break down sentences into three length bands[1]: short (< 8syllables), medium (8-15 syllables), long (> 15 syllables) (see Table 1). Overall, 22 out of 58 studies used stimuli sentences of medium length while 21 studies used stimuli sentences of varying lengths, i.e., sentences across two or even three length bands.

3.1.2.3.2 Repetition delay

The insertion of delay often takes the form of a period of silence (usually three to five seconds) or an interruptive task (e.g., answering a cognitively unchallenging question) before repetition. As is discussed earlier, repetition of sentences without comprehension is possible if the learner continuously rehearses the chain of sounds before repetition; therefore, the insertion of delay should interrupt continual rehearsal. However, as Vinther (2002) argues, the insertion of delay may also interfere the processing of the structure and meaning of the sentences, especially when the sentences

[1] For studies which did not report the number of syllables per sentence or reported number of words per sentence instead of syllables, I used the examples shown in the paper for syllable counts.

are long. As Table 1 shows, only 21 out of 58 studies implemented delayed repetition in their EI tasks.

3.1.2.3.3 Grammatical features of the sentence stimuli

The difficulty of EI tasks has been shown to be influenced by linguistic features of the sentence stimuli, including among others: syntactic complexity (Ortega, 2000), lexical difficulty (Graham, McGhee & Millard, 2010), phonological structure of the words in the sentence (Menyuk, 1971), and the use of ungrammatical sentences (Erlam, 2006). As shown in Table 1, it appears that the most common ways of controlling linguistic features of the sentence stimuli are the control of the syntactic and morphosyntactic features of the sentence (k=36) and the use of ungrammatical sentences (k=22).

The relationship between features of syntactic, lexical, and phonological complexity and the resultant difficulty of EI tasks can be frequently observed across studies (Graham et al., 2010; Menyuk, 1971; Perkins et al., 1986; Ortega, 2000). However, the use of ungrammatical sentences is much debated. The rationale for using ungrammatical sentences is that these sentences naturally elicit automatic correction of grammatical errors especially when the sentence length exceeds the capacity of STM. Hamayan, Saegert and Larudee (1977) argued that failure to correct grammatical errors is evidence of inadequate implicit knowledge of the target structures. However, error correction does not necessarily occur even among L1 speakers (Markman et al., 1975), especially when the instructions do not require subjects to do so, which poses questions

on the usefulness of ungrammatical sentences in measuring the target linguistic structures.

3.1.2.3.4 Scoring method

The three most common approaches to scoring EI responses are the binary yes-no approach (k = 24), the ordinal rating scale approach (k = 15), and the interval scale approach (e.g., number or percentage of errors, or automated measures of prosodic features, k = 15). The yes-no approach only gives two possible scores for each EI response: 1 for correct repetition and 0 for incorrect repetition (e.g., Ellis, 2005; Erlam, 2006). The rating scale approach establishes a rating rubric, usually more than three score levels, to quantify the accuracy of repetition (e.g., Markman et al., 1975). The interval scale approach often utilizes error rate of particular grammatical features (e.g., West, 2012) and automated scoring tools for more complicated linguistic analysis (e.g., Longsdale & Christensen, 2011; Trofimovich & Baker, 2007). It is reasonable to speculate that the choice of scoring method may influence the reliability of EI and its ability to discriminate speakers across proficiency levels. However, the impact of the choice of scoring method remains less investigated than other task features in the literature.

3.1.3 Summary of Phase I

The narrative synthesis shows that EI has been widely used in L2 research; however, the constructs argued to be measured by EI have undergone interesting shifts

over time, with more studies focusing on specific linguistic structures and implicit grammatical knowledge. In addition, EI has been used as an outcome measure for the effectiveness of certain treatments. The resurgence of EI studies in the literature indicates that EI has regained attention from L2 researchers as a potential useful tool to measure L2 proficiency. Nevertheless, this survey of the extant empirical EI studies indicates a great degree of variation in the design of four key EI task features, all of which are associated with the construct validity of EI. The extent to which variation in the design of EI tasks has an impact on the quality of the measurement requires further investigation.

3.2 Phase II: Meta-analytic Investigation

Findings from the narrative synthesis form the theoretical basis for the quantitative meta-analysis in *Phase II*, which helps clarify the construct measured by EI and further informs whether variation in the design of the key task features have an impact on the sensitivity of EI as a measure of language proficiency. The meta-analysis presented in this section addresses two questions: (1) whether (and to what extent) EI tasks can distinguish between higher and lower proficiency speakers; and (2) whether (and to what extent) the sensitivity of EI differs across designs of the four task features discussed in *Phase I*.

3.2.1 Methods

3.2.1.1 Study selection criteria

In addition to the selection criteria for *Phase I* , studies must fit two additional conditions in order to be included in a meta-analysis: (1) the study has at least two groups of participants at two different proficiency levels (e.g., advanced vs. intermediate) to be compared quantitatively; and (2) the researchers report means, standard deviations, sample sizes and/or other statistical results (e.g., *t*-statistic, Pearson's *r*, chi-square statistic, or *F*-statistic with a degree of freedom of 1) that are required for computing a Hedges' *g* (1981) effect size.

3.2.1.2 Identification of studies

Among the 76 studies included in the narrative synthesis, thirty studies were group comparison studies. However, only 10 studies out of 30 met the additional criteria for the meta-analysis. The other 20 studies did not report sufficient statistic that enable us to compute effect sizes, and were therefore excluded from the meta-analysis. Efforts were also made requesting the statistical information necessary to compute effect sizes; however, the attempts were not successful. The process identified 10 studies that met the selection criteria, including six published journal articles and four unpublished doctoral dissertations or master theses (see Figure 3.1).

3.2.1.3 Data extraction

To examine the ability of EI scores to discriminate speakers across L2 proficiency levels, I was able to extract 13 effect sizes (representing 498 cases) from these 10 studies that indicate the differences of mean EI scores between lower and higher language speakers (e.g., either L1 or advanced L2 speakers). Because of the small number of effect sizes, using a Cohen's *d* effect size (1988) tends to overestimate the magnitude of the effect. In order to minimize the bias due to small sample size, an unbiased estimator called Hedges' *g* (Hedges, 1981) was used as the effect size. In this review, Hedges' *g* indicates the magnitude of the standardized mean differences in EI test scores of higher L2 proficiency groups and lower L2 proficiency groups. A positive effect size means that higher L2 proficiency speakers tend to score higher on EI tasks than lower L2 proficiency speakers and support the construct validity of EI as a measure of L2 proficiency; on the other hand, negative or close-to-zero effect size means that higher L2 proficiency learners perform similarly to lower L2 proficiency learners on EI tasks, suggesting that EI is not a sensitive or valid measure to distinguish speakers across different levels of L2 proficiency.

3.2.1.4 Data analysis

I used a random-effects model as the theoretical framework for combining effect sizes because this meta-analysis may only represent a sample of all the studies that compare performance on EI tasks across L2 proficiency levels (Hedges & Vevea, 1998). The *Q* test (Hedges & Olkin, 1985) was conducted to evaluate the homogeneity of

retrieved effect sizes. An alpha level of .05 was set for statistical significance. In addition, an I^2 statistic, which indicates the ratio of the true heterogeneity (between-study variance) to the total variance across the observed effect estimates (Higgins, Thompson, Deeks, & Altman, 2003), was calculated to quantify the amount of variation in the effect sizes due to the differences between studies. Weights, calculated by taking the inverse of the variance of each effect size, were used to reflect the precision of the estimated effect sizes retrieved from each study. I used Comprehensive Meta-Analysis software (Version 2; Borenstein, Hedges, Higgins, & Rothstein, 2007) to run all the statistical analyses involved in the meta-analysis.

3.2.1.4.1 Handling multiple effect sizes

Multiple effect sizes obtained from the same study may violate the statistical assumption of independence for inferential analyses in meta-analysis. Multiple effect sizes were retrieved from eight of the 10 studies except Zhou (2012) and Iwashita (2009). Effect sizes obtained from different comparison-groups were considered independent (e.g., Serafini, 2013; Wu & Ortega, 2013); others obtained from the comparison of the same two groups were considered dependent (e.g., Bowles, 2011; Erlam, 2006; Flynn, 1986; Li, 2010; West, 2012), which require some adjustments to avoid statistical violations and to ensure the validity of the analyses. Hence, I established the following criteria to resolve the issue as described below:

- In the studies (e.g., Bowles, 2011; Erlam, 2006; Flynn, 1986; Serafini, 2013; Yoon, 2010) that reported both total scores and subsection scores, the total scores were selected for computing effect size from the study.
- In the studies (e.g., West, 2012) that used multiple comparable interval variables to score EI, effect sizes for individual variables were averaged.
- In the studies (e.g., Bowles, 2011; West, 2012) that included three proficiency level groups, the effect size for the two adjacent levels that had the smaller mean score difference was selected. Although the selection of the smaller effect sizes might underestimate the magnitude of the effect sizes, i.e., the discrimination associated with EI, I chose to be conservative as I was interested in examining the ability of EI tasks to make relatively fine distinctions between adjacent proficiency levels.
- In the quasi-experimental studies (e.g., Flynn, 1986; Serafini, 2013), only comparison on pretest scores was selected in order to avoid an intervention effect.
- In the studies (e.g., Serafini, 2013; Wu & Ortega, 2013) that included multiple independent groups for each proficiency level, the effect sizes for all independent group comparisons were used instead of the effect sizes for the combined total group comparison.

This process enabled us to retrieve 120 dependent and 13 independent effect sizes from the 10 studies.

3.2.1.4.2 Identification of potential moderators

Due to the small number of studies, an inferential test was not used for a moderator analysis that investigates how the design features of EI task may relate to the sensitivity of EI to distinguish different proficiency groups. Instead, I grouped studies by the design of different task features that I reviewed in *Phase I* and reported the average effect sizes to highlight the trend in the variation of the effect sizes as the function of different designs of EI task features. The procedure allows us to suggest potential moderators to explain the variation in the sensitivity of EI tasks. More specifically, possible moderators were suggested through discussing the weighted average effect sizes across different designs on four key task features: (1) the length of sentence stimuli, (2) the use of ungrammatical sentence stimuli, (3) the insertion of delay, and (4) the scoring method. In addition, I examined the sensitivity of EI with respect to the type of construct, by comparing weighted average effect sizes between studies that use EI to measure global language proficiency and studies that target on specific linguistic structures.

3.2.2 Results

3.2.2.1 The ability of EI to differentiate speakers with proficiency levels

Table 3.3 reports summary of studies included in the meta-analysis and statistics used to compute effect sizes. The forest plot of the 13 independent Hedges' *g* effect sizes with their 95% confidence intervals is shown in Figure 3.3. The mid-point of each line represents the point estimate of the effect size. The length of each line represents the

range of 95% chance within which the true effect size lies. The forest plot suggests the variation in the precision of effect size estimates as some of the effect sizes have larger confidence intervals while others have smaller confidence intervals.

The weighted average effect size for the 13 effect sizes was 1.42, with a standard deviation of 0.81. The large mean effect size shows that EI tasks can effectively distinguish between speakers of different proficiency levels. This adds supportive construct-related validity evidence to EI as a measure of L2 proficiency. The fact that higher proficiency speakers performed consistently better on EI tasks than did lower proficiency speakers across studies provides substantial evidence that EI is a reliable and valid measure of language proficiency. More specifically, higher proficiency speakers were more capable of repeating the sentences than were lower proficiency speakers for all studies used for meta-analysis. Therefore, it is more likely that, in order to repeat sentences, the speaker has to rely on his or her internalized linguistic knowledge to decode the structural information of the sentence and then reconstruct the meaning of the sentence. In other words, parroting, or rote repetition of a chain of sounds alone does not allow successful completion of EI tasks.

The homogeneity test of effect size indicates that the effect sizes varied significantly across studies, $Q(12)=49.71$, $p < .001$. The estimated between-study variance of effect sizes τ^2 was 0.46, which suggests a large variation in the effect sizes. The I^2 statistic was 75.86%, which indicates that a large proportion of variation in effect sizes was due to the differences across individual studies. In summary, these statistical results indicate that while, on average, the EI tasks can identify higher and lower language

proficiency groups, the sensitivity of EI differs in great extent across studies possibly depending on the way EI tasks were designed and implemented.

Table 3.3 Summary of Studies Included in Meta-analysis

ID	Study	Language	Construct	Control			Scoring	n_1	M_1	SD_1	n_2	M_2	SD_2	g	SE
				Gr	Lg	Dl									
1	Bowles, 2011	Spanish	Morpho -syntactic	Y	V	Y	Ordinal	10	78.8	15.00	10	46.70	9.80	2.43	0.58
2	Erlam, 2006	English	Morpho -syntactic and syntactic	Y	V	Y	Binary	20	0.94	0.04	95	0.51	0.17	2.74	0.30
3	Flynn, 1986	English	Syntactic	Y	M	N	Binary	14	2.28	0.90	21	1.86	0.88	0.46	0.34
4	Iwashita, 2009	Japanese	Global	N	V	Y	Interval	20	93.15	15.55	13	71.92	19.54	1.20	0.38
5	Li, 2010	Mandarin	Syntactic	Y	M	Y	Ordinal	14	5.41	2.91	14	1.93	1.52	1.46	0.42
6								14	4.82	3.01	15	1.47	1.52	1.38	0.40
7								11	5.87	2.75	10	1.98	1.75	1.60	0.49
8	Serafini, 2013	Spanish	Morpho -syntactic and syntactic	Y	L	Y	Mixed	33	0.44	0.09	23	0.31	0.11	1.30	0.30
9	West, 2012	Spanish	Morpho -syntactic	N	M	Y	Interval	16	0.35	-	16	0.41	-	0.75	0.36
10	Wu and	Mandarin	Global	N	V	Y	Binary	20	71.55	22.56	20	52.9	21.38	0.83	0.32
11	Ortega, 2013							20	61.45	23.21	20	38.2	17.12	1.12	0.33
12	Yoon, 2010	English	Phonological	N	S	N	Interval	8	82.00	7.00	18	79.00	11.00	0.29	0.41
13	Zhou, 2012	Mandarin	Global	N	V	Y	Binary	12	70.58	13.10	11	33.55	12.90	2.75	0.57

Note. S=short, M=medium, L=long, V=varied; Y=yes, N=no; Gr=use of ungrammatical sentences, Lg=sentence length, Dl=delayed imitation.

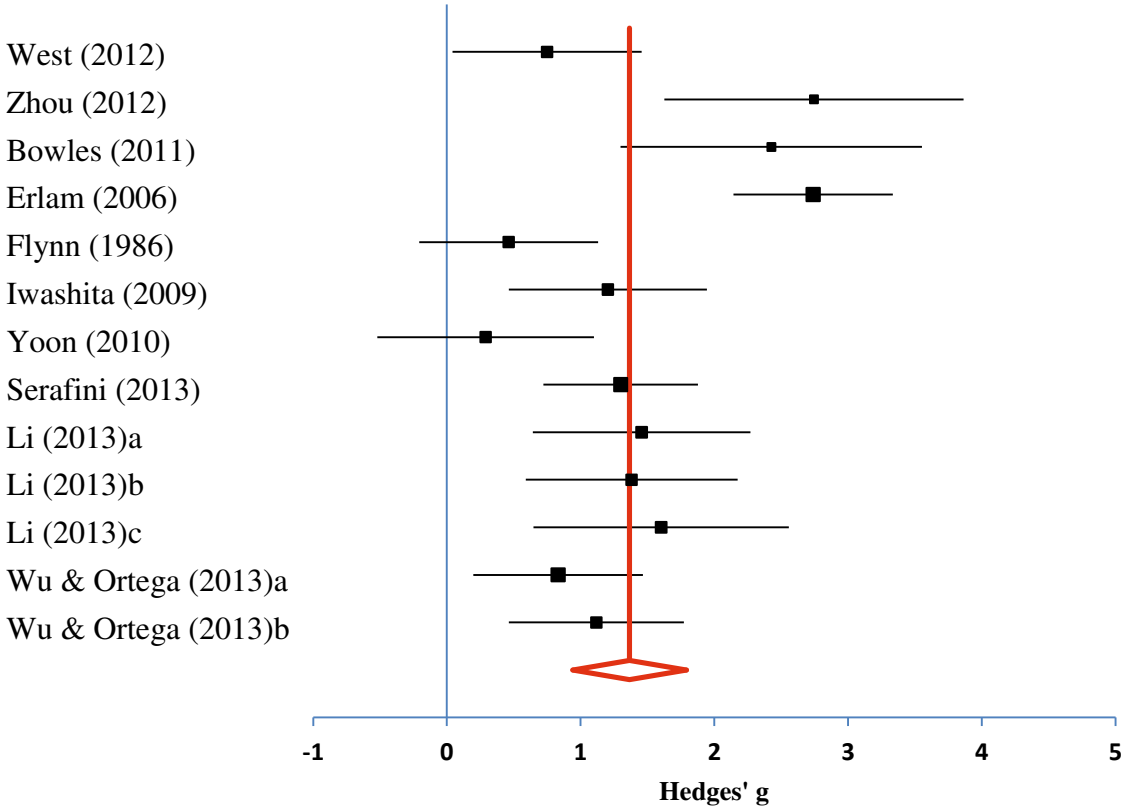


Figure 3.3 Hedges' g Effect Sizes with 95% Confidence Intervals

3.2.2.2 Task features as potential moderators for the sensitivity of EI

As previously stated, I used the descriptive statistics to explore the potential moderator variables; Table 3.4 reports the weighted average Hedges' *g* with 95% confidence intervals for EI studies grouped by the design of certain task features. It is important to note that all the confidence intervals overlapped to varying extent across different designs on the same features, although the descriptive statistics appear to be quite different.

Table 4 Comparisons of Sensitivity of EI Expressed by Hedges' *g* across Designs of Certain Task Features

Task features	**Comparison Group**	**N**		***g***	***SE***	**95% CI**
		Studies	**Effect sizes**			
Construct	Global	3	4	1.39	0.39	[0.62, 2.16]
	Specific	7	9	1.30	0.23	[0.84, 1.76]
Sentence Length	Varying length	5	6	1.79	0.32	[1.16, 2.43]
	Equal length	5	7	1.04	0.29	[0.46, 1.62]
Use of Ungrammatical Sentences	Yes	5	7	1.62	0.30	[1.03, 2.20]
	No	5	6	1.11	0.32	[0.48, 1.73]
Delayed Imitation	Yes	8	11	1.57	0.24	[1.10, 2.03]
	No	2	2	0.38	0.55	[-0.69, 1.46]
Scoring method	Binary	4	5	1.53	0.35	[0.85, 2.21]
	Ordinal	2	4	1.41	0.37	[0.69, 2.12]
	Interval & Mixed	3	4	1.10	0.44	[0.24, 1.96]

There are two possible explanations for the absence of significant differences across different designs of EI task features. One possibility is that variation in the design of these task features does not have an impact on the sensitivity of EI, and therefore, it does not matter how one may develop and administer EI tasks. The other possibility is that non-significant difference is largely due to low statistical power, which is caused by the small number of effect sizes (k) in each comparison group. Although no further evidence is available at this point, by observing the descriptive statistics and confidence interval bands for the weighted effect sizes, I can identify trends in the sensitivity of EI associated with some task design features. Therefore, I argue for consideration of these design features as potential moderators for the sensitivity of EI.

The first potential moderator was sentence length. The five studies that employed sentences of varying length ($k = 5$, $g = 1.79$, $SE = 0.32$) tended to have larger effect sizes than did studies that did not ($k = 5$, $g = 1.04$, $SE = 0.29$). This suggests that variation in the length of EI sentence stimuli may lead to heightened sensitivity in EI when used to measure different L2 proficiency levels. Another potential moderator was the insertion of delay before imitation. Although there were only two studies in the sample that did not utilize delayed imitation, the weighted average effect size for EI tasks with delayed imitation ($k = 8$, $g = 1.57$, $SE = 0.24$) appeared much larger than the studies without ($k = 2$, $g = 0.38$, $SE = 0.55$).

Regarding the use of ungrammatical sentences, EI tasks in five studies that used a combination of grammatical and ungrammatical sentences ($k = 5$, $g = 1.62$, $SE = 0.30$) appeared to have larger effect sizes than EI tasks in the other five studies that only used grammatical sentences ($k = 5$, $g = 1.11$, $SE = 0.32$) in the discrimination of higher and

lower proficiency speakers. Finally, type of construct and scoring method were not identified as potential moderators for the discriminating power of EI. The 95% confidence intervals for the two groups largely overlapped. This suggests: (1) EI tasks may be comparably suitable to measure both global language proficiency and specific linguistic constructs; and (2) The design of other task features might have contributed largely to the sensitivity of EI tasks as a measure of L2 proficiency that the scoring method does not necessarily add much variation to the sensitivity of EI scores.

Due to small sample size and possibly biased sampling, I examined the potential impact of publication bias on the validity of the statistical conclusion. The analysis indicated that studies with small sample sizes are absent in the pool of primary studies. Because studies have larger sample sizes are more likely to be published, the result might overestimate the overall sensitivity of EI although these missing studies with small sample size will have less impact on the results due to relatively small weight given to the effect size. However, the average sensitivity of EI shows strong support for its construct validity, and it is unlikely that this finding would be reversed with additional small studies.

3.2.3 Summary for Phase II

Findings of the meta-analysis suggested that, in general, EI is a sensitive measure to discriminate speakers across proficiency levels. In terms of EI task design, I found no principled or systematic ways of developing and implementing EI tasks across studies, with great variability in how EI tasks have been designed and administered. Nevertheless, a closer look at the effect sizes by EI design features suggested that the ability of EI tasks

to differentiate higher and lower proficiency speakers is likely to be strengthened by the manipulation of certain task features. This implies the importance of principled EI task development for increasing the sensitivity of the instrument.

3.3 Overall Discussion and Implications

In this two-phase systematic review, I examined the use of EI in L2 research, in an effort to clarify the construct measured by EI, and, more importantly, to advance discussions toward a more principled practice of EI in L2 research. Therefore, I further discuss the implications of these findings below from two perspectives: usefulness of EI as a measure of L2 proficiency and the design of certain key EI task features.

3.3.1 Usefulness of EI in Classroom and Standardized Assessment Contexts

Results of this review support the idea that EI is an effective measure of global language proficiency, specific linguistic structures, and the effectiveness of instructional interventions. The simple and economical administration procedures and the flexibility in the design of task features makes EI an attractive candidate for a quick and effective measure of language-related constructs in both classroom and standardized assessment contexts.

To better understand the usefulness of EI tasks in both classroom and standardized assessment contexts, future research examining the connection between psycholinguistic measures and performance-based measures is imperative to connect the two main approaches to language testing. While EI and similar psycholinguistic

measurements are argued to lack authenticity (Bachman, 1990; Morrow, 1979), they usually outperform interactive, performance-based tasks in terms of reliability (Bernstein, van Moere & Cheng, 2010; van Moere, 2012). In addition, EI tasks can facilitate classroom assessment in language classes due to its simplicity in administration and reliability of scoring. Employing a combination of psycholinguistic and performance-based measures can complement the limits of each type of measure, thus optimizing the usefulness of multiple measures of the same construct (van Moere, 2012). However, future research should go beyond simply examining correlations of holistic scores on those measurements to the analysis of alignment of specific linguistic or non-linguistic features of the tasks crucial to communication.

3.3.2 Design of Certain EI Task Features

Previous reviews of EI (see, e.g., Vinther, 2002) pointed out a number of tasks features that may affect the construct validity of EI as a measure of language proficiency. However, the findings of this review suggest that manipulation of three task features, i.e., sentence length, delayed imitation, and the use of ungrammatical sentences, may distinguish EI performances across L2 proficiency levels better. First, EI tasks using sentences with varying length appeared to be more discriminating than EI tasks using sentences with fixed length. This possibly results from the fact that sentence length tends to be positively correlated with the difficulty of EI tasks (e.g., Miller, 1973; Perkins et al., 1986). In addition, results of the meta-analysis in this study suggested that higher proficiency speakers tend to be more capable of sentence repetition than do lower proficiency speakers even when it comes to longer sentences (e.g., Serafini, 2013). Thus,

EI tasks with varied sentence length will more likely be appropriate for distinguishing learners with different proficiency levels. Second, the improved sensitivity of EI with an insertion of delay, to certain extent, is in line with the literature on STM and EI in that the insertion of delay can interrupt continual rehearsal of the sounds in the phonological loop (Gathercole & Baddeley, 1993) and thus may force the participants to rely on their internal linguistic knowledge stored in the LTM.

Finally, EI tasks using ungrammatical sentences tended to be more effective in discriminating speakers with different proficiency levels. This appears to support Hamayan et al.'s (1977) argument that ungrammatical sentence stimuli can elicit correction of grammatical errors and that higher proficiency speakers tend to be more able to automatically correct grammatical errors than lower proficiency speakers. Moreover, this trend supports the statistical approach to language comprehension and production (Ellis, 2002). Once the sentence is decoded for meaning, instead of retaining the original ungrammatical structure, the speaker reconstructs the meaning of the sentence based on the frequency patterns of relevant lexical and syntactic structures in his or her implicit grammatical knowledge and automatically corrects the grammatical errors in the repetition. In other words, the speaker is not simply imitating acoustic information from the stimuli, but rather repeating the sentence using internalized lexico-grammatical representations. Therefore, it is reasonable to argue, though indirectly, that EI prompts language comprehension and production rather than rote repetition.

3.3.3 Recommendations for Future Research and the Use of EI Tasks

Based on findings of the review, I recommend future studies using EI as a measure of L2 proficiency take the impact of the aforementioned three task features into account for designing effective EI tasks. In addition, further research effort should be made to investigate how the design of key EI task features functions under specific assessment purposes and contexts. For example, future investigations might focus on identifying optimal sentence length used in EI tasks in relation to the target proficiency levels. An important factor to consider when choosing the sentence length, though less explicitly articulated in the literature, is learners' language proficiency level. That is, higher proficiency speakers tend to repeat longer sentences as compared to lower proficiency speakers. From a measurement perspective, the more the task difficulty matches the target proficiency level, the more reliable the scores are and thus the more valid the judgment about the proficiency level of the learner can be. As is discussed previously, using a range of sentence length is likely to vary the difficulty levels of the EI tasks, which increases the potential of EI tasks to target at multiple proficiency levels. Yet, the current study indicated that the selection of the range of sentence length in relation to the appropriate difficulty levels or proficiency levels remains under explored.

In addition, future research on examining the impact of the administration and scoring procedures for the imitation of ungrammatical sentences on the sensitivity of EI is beneficial because the administration procedures may create construct-irrelevant variance in the scores (Kaplan, 1996; Munnich et al., 1994). As most EI tasks with ungrammatical sentences require participants to repeat the sentences verbatim without

mentioning the grammatical errors in the sentence stimuli, it remains unknown whether failure to correct grammatical errors in the stimulus sentence is a result of simply following the instructions. Although the purpose of using ungrammatical sentences is to elicit error correction through implicit grammatical knowledge, not all EI tasks clearly instruct the participants to correct errors in the sentence and therefore error correction is not guaranteed (see, e.g., Markman et al., 1975). However, to ameliorate construct-irrelevant variance, one should also be careful about directing too much attention from the participants to the grammatical errors as the nature of the target construct may change if error correction becomes less automatic.

Finally, to better facilitate systematic review of the sensitivity of EI, I strongly recommend that future researchers follow systematic reporting practices of empirical and statistical results. As Plonsky (2013) argues, the reporting of descriptive statistics, including sample sizes, means and standard deviations, "avails primary data to would-be meta-analysts who require such data to calculate an effect size" (p. 671). Missing or insufficient empirical information not only discounts the generalizability of findings but also inhibits readers' ability to understand and assess the results and findings of the primary studies.

3.3.4 Limitations

Although the narrative synthesis of this study was based on 76 studies, the meta-analysis was conducted on a rather small sample of effect sizes. A major difficulty I encountered during the meta-analysis was the extraction and calculation of effect sizes due to the inconsistent reporting practice of statistical information in the field of second

language studies (see also, e.g., Plonsky, 2013). As mentioned in the Methods section, there were 30 group-comparison studies identified in the literature, among which, however, 20 studies did not report sufficient statistical information that would enable computation of effect sizes. The number of independent effect sizes I extracted, while sufficient for the meta-analysis of the overall sensitivity of EI, was rather small to arrive at a conclusive understanding of the moderating effect of task design features on the sensitivity of EI. Although meaningful trends associated with potential moderators were observed, if I had enough independent effect sizes, I could perform moderator analyses to investigate the impact of the key task features on the ability of EI to discriminate speakers with different proficiency levels. Therefore, the identification of potential moderators should be interpreted with caution.

However, regardless of these limitations, I conclude that EI tasks have potential to effectively and reliably distinguish performance across proficiency levels. The results of this systematic review provide construct validity evidence for EI as a measure of L2 proficiency and contribute to continued and extended discussion of EI towards a more principled practice for the development of EI tasks in L2 research.

Kuhn (1962) in his discussion of the structure of scientific revolutions has argued that changes in paradigms are characterized by one (e.g., structural) being replaced by another (e.g., communicative) followed by a reassessment in which the interests of both realign. Perhaps the current interest in EI, disfavored for several decades, represents the beginning of a realignment in which the usefulness of EI can be reassessed within the broader research context emphasizing communicative activities and interaction. EI, as well as other psycholinguistic measures, can be employed in combination with, instead of

replaced by, communicative language tasks, to ultimately enhance the quality of L2 teaching and assessment.

CHAPTER 4. METHOD

The present study examined the processing of formulaic language on EI tasks by L2 speakers, as an effort to investigate the processing advantage of formulaic sequences. In other words, this study examined whether the use or production of formulaic sequences may or may not contribute to L2 fluency in speaking performance. In order to understand the effect of formulaic sequences on L2 fluency, this study utilized EI tasks designed to elicit repetition of individual sentences containing formulaic language in comparison with repetition of sentences that do not. In addition to the presence of formulaic sequences, length of stimulus sentences was included as another independent variable of interest in this study.

The development of EI tasks controlled for sentence length and insertion of delay, two task features examined in the meta-analysis that may have an impact on the sensitivity of EI as a measure of L2 proficiency. Ungrammatical sentence stimuli were not employed as it remains uncertain whether these sentences will naturally elicit error correction in the repetition. In addition, formulaic sequences were embedded in half of the EI sentences to examine whether formulaic sequences have processing advantages for L2 speakers. Finally, all the sentence stimuli were controlled on a set of linguistic variables suggested in the literature as contributing factors in the difficulty of EI tasks (discussed in 4.3 Elicited Imitation Tasks).

Responses to EI tasks were automatically measured on articulation rate and number of silent pauses using PRAAT, Version 5.4.05, (Boersma & Weenink, 2015), a free computer software package for analysis of speech in phonetics. These two variables constituted the dependent variables of interest in this study. A two-way repeated measures ANOVA model was employed to examine the main and interaction effects of the two independent variables on the two dependent variables.

4.1 Variables of Interest and Research Questions

In this study, presence of formulaic sequences (FS) and stimulus sentence length (SL) were the independent variables of interest. FS was a within-subjects factor with two levels: sentences with FS (FS-F) and sentences without FS (FS-NF), which means that each participant repeated both sentences with formulaic sequences and sentences without formulaic sequences. SL was also a within-subjects factor but with three levels: short (SL-S), medium (SL-M), and long (SL-L).

The dependent variables of interest, both of which were about performance characteristics, consist of articulation rate (AR) and number of silent pauses (NumSP), both measured automatically through PRAAT.

Based on the aforementioned variables, three research questions can be formulated for this dissertation:

1. Does presence of FS have a significant effect on AR of EI responses? Does SL have a significant effect on AR of EI responses? Is there any interaction between FS and SL?

2. Does presence of FS have a significant effect on NumSP in EI responses? Does SL have a significant effect on NumSP in EI responses? Is there any interaction between FS and SL?
3. Is there a significant correlation between AR and NumSP?

4.2 Participants

Participants of this study consisted of 194 undergraduate ESL students at Purdue University. These students were enrolled in an EAP course designed to improve students' English language skills in order to help them take full advantage of a range of educational opportunities available at Purdue University. The students' English proficiency level ranged from low intermediate to high intermediate. This group of students was targeted because they represent the majority of the undergraduate ESL students in terms of English proficiency level and the need of additional language support for academic performance at English-medium universities. Descriptive statistics of their TOEFL iBT scores are included in Table 4.1.

Table 4.1 Descriptive Statistics of Participants' TOEFL iBT Scores

	N	**Mean**	**SD**	**Minimum**	**Maximum**
Reading	194	24.63	2.962	19	30
Listening	194	23.46	3.049	17	30
Speaking	194	20.41	1.708	18	27
Writing	194	22.68	2.328	18	28
Total	**194**	**90.91**	**5.124**	**79**	**103**

Prior to the data collection, power analysis (set at the power level of 0.8) was conducted using GPower (version 3.1.6) to inform the number of participants needed to

investigate the main and interaction effects of FS and SL in a two-way within-subjects (or two-way repeated measures) ANOVA design (see Figure 4.1). Because two ANOVA tests were performed on the same group of participants, the Bonferroni adjusted significance level of 0.025 was specified in the power analysis. Results of the power analysis indicated that at least 42 participants are needed to detect a medium eta-squared effect size ($\eta^2 = .25$). Therefore, the sample of 194 participants in this study provides enough power for the statistical tests performed in the repeated measures ANOVA design.

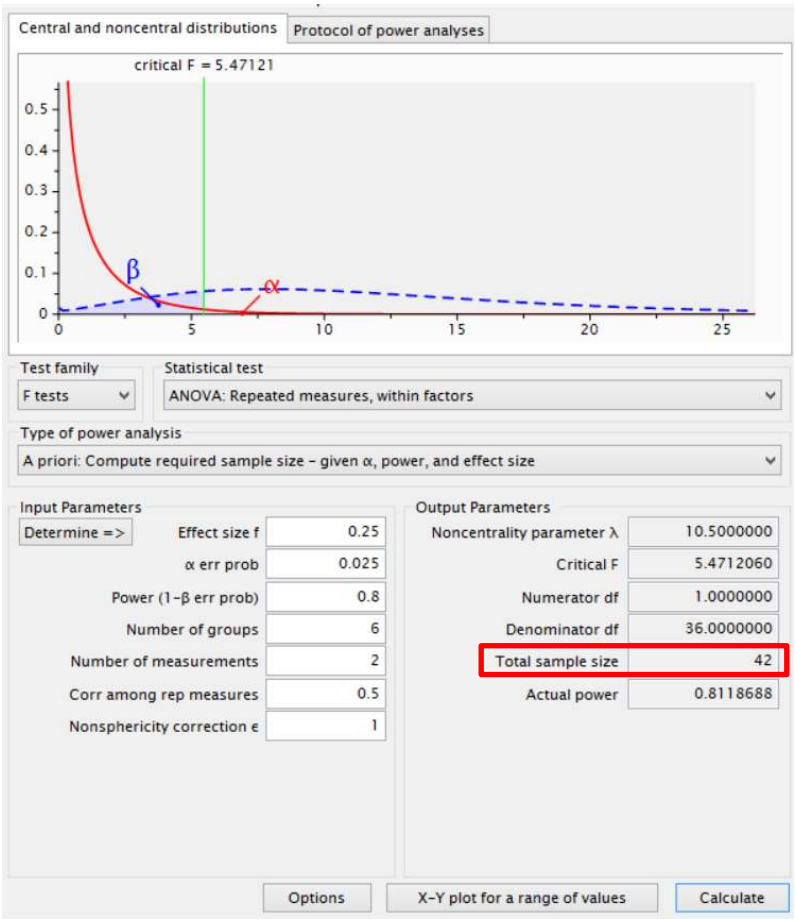


Figure 4.1 Results of Power Analysis for the Required Sample Size

4.3 Elicited Imitation Tasks

Four forms of EI tasks were used in this study, each comprising a set of 24 sentence stimuli for participants to hear and repeat. Situated in a university setting, all the sentence stimuli used in EI tasks were formulated such that the sentences either convey important information pertaining to everyday college life or are sentences that college students may frequently hear or say on campus. The topics covered in the sentences included, but were not limited to, health insurance, academic context, student clubs and activities, lifestyle on campus and in Midwestern US.

As suggested in the literature review, to avoid parroting on the EI tasks, a delay and interruption was inserted, in the form of a question task, between each sentence stimulus and response. Participants had to choose a word that is mentioned in the sentence before repeating it. In addition, as literature suggests, the difficulty of EI tasks are associated with a number of linguistic variables. In order to better investigate effects of the independent variables of interest, all sentence stimuli in this study were carefully controlled on phonological, lexical, and syntactic complexity. Levels of each factor, related literature, and specific control procedures are provided in Table 4.1. Finally, of all the 24 sentences, 12 sentences contained formulaic sequences and the other 12 sentences did not. In addition, the 24 sentences were evenly distributed across three levels of sentence lengths, with eight short sentences (eight to nine syllables), eight medium-length sentences (15 to 16 syllables), and eight long sentences (20 to 21 syllables). The combinations of the two variables (i.e., FS and SL) and the number of sentences in each category are illustrated in Table 4.2.

Table 4.1 Control Procedures of Linguistic Factors in EI Sentence Stimuli

Factor	Level	References	Control procedures
Sentence length	8-9 syllables 15-16 syllables 20-21 syllables	Bailey, Eisenstein, & Madden (1976) Jensen & Vinther (2003) Naiman (1973)	• SendGrid Syllable Counter (Poetry Soup, 2013), a free web-based program was used to count the number of syllables in each sentence.
Presence of formulaic sequence	Formulaic Non-formulaic	Schmidt (2004) Simpson-Vlach & Ellis (2010) Tremblay et al. (2011)	• Half of the sentence stimuli include formulaic sequences whereas the other half do not. • Both formulaic and non-formulaic sequences are defined in terms of corpus-based frequency and mutual information index. • Formulaic sequences are first selected from the academic formulas list (commonly used in speaking only) created by Simpson-Vlach & Ellis (2010), and then matched with Corpus of Contemporary American English (COCA) spoken corpus for frequency and mutual information index. Non-formulaic sequences are checked in the COCA spoken corpus for frequency and mutual information index. • Specific selection criteria: o Formulaic sequence: (Frequency >10 per million, Mutual Information>3) o Non-formulaic sequence: (frequency< 5 per million)

Table 4.1 continued.

Factor	Level	References	Control procedures
Syntactic complexity	Relative subordination Noun subordination Adverbial subordination	Perkins, Brutten, & Angelis (1986) Tracy-Ventura, McManus, Norris, & Ortega (to appear)	• Half of the sentences have subordinate clauses while the other half do not. • Subordinate clauses comprise three types: adjective, noun, and adverbial subordinations.
Lexical difficulty	-	Graham, McGhee, & Millard (2010) West (2012)	• No words contain more than three syllables. • No words contain more than two morphemes. • The K1 (the most frequent 1000 words), K2 (the most frequent 2000 words), and AWL (academic word list) words lists were consulted for keeping sentences lexically comparable (around 90% of the words are on the K1 and K2 word lists) through Compleat Lexical Tutor (version 6.2; Heatley & Nation, 1994), a free web-based program.

Table 4.2 Design of EI Sentences

		Sentence length (SL)			**Total**
		Short (SL-S)	Medium (SL-M)	Long (SL-L)	
Formulaic sequences (FS)	Sentences with formulaic sequences (FS-F)	4	4	4	**12**
	Sentences without formulaic sequences (FS-NF)	4	4	4	**12**
Total		**8**	**8**	**8**	**24**

4.4 Procedures

Participants were asked to listen to 24 sentences and then repeat the sentences verbatim. Each sentence was played only once. After each sentence was played on the computer, the screen would change and two words would appear. One of the two words was mentioned in the sentence. Participants had to click on the word that was mentioned in the sentence and then repeat the sentence exactly as it was stated (See APPENDIX for directions and sample EI tasks).

The EI tasks were timed. Participants had eight seconds to choose the word and 20 seconds to repeat each sentence. Meanwhile, a small timer clock would appear at the top right corner of the screen to indicate the remaining response time for each task. When the response time is up, the screen will automatically switch to the next task. If the participant finishes the task before the time constraints, he or she can click on the "continue" button on the bottom right corner of the screen to move on to the next task.

Responses to the EI tasks were analyzed using a PRAAT script developed by de Jong and Wempe (2009) to automatically extract values associated with the two temporal measures of interest in this study: articulation rate and number of silent pauses. Articulation rate is the total number of syllables divided by the sum of speech time and total filled pause time; this value was multiplied by 60 to obtain the articulation rate per minute. In this study, silent pauses were defined as pauses of 0.25 seconds or longer; number of silent pauses refers to the total number of silent pauses per speech sample (i.e., responses to each of the 24 EI tasks). After the two values were extracted for all the EI tasks, these values were then grouped by condition (i.e., combinations of the levels of the two factors: FS and SL). AR of all sentences within each condition was averaged; NumSPs were summed to obtain the total number of silent pauses within each condition (see Table 4.2, for the design of EI tasks). These values in different cells represented scores on the repeated measures across conditions for the ANOVA tests.

4.5 Data Analysis

Participants' responses to EI tasks were analyzed on two measures: AR and NumSP. Data analyses for this study comprised two stages: 1) investigation of the main and interaction effects of the two independent variables (i.e., FS and SL) on the temporal measures of L2 fluency (i.e., AR and NumSP); 2) examination of the correlations between between AR and NumSP. Therefore, the first stage addresses the first two research questions, and the second stage addresses the third research question (see Section 4.1).

During the first stage, I conducted two two-way repeated measures ANOVA tests to analyze the main and interaction effects of FS and SL on the variance of AR and NumSP. Since effects on two dependent variables were tested, both ANOVA tests were Bonferroni-adjusted to the significance level of .025. Next, Pearson *r* correlation coefficients were computed to estimate the relationship between AR and NumSP.

All the statistical analyses were performed on the Statistical Package for the Social Sciences (SPSS), Version 21.0 (IBM Corp., 2012). Prior to the ANOVA tests, screening procedures were performed to check whether the data met the statistical assumptions for two-way within-subjects ANOVA. First, for a two-way within-subjects ANOVA design, the dependent variables must be quantitative, and the independent variables must be categorical. In addition, there are three assumptions that the data should satisfy in order to generate reliable and valid results from the analyses.

- Normality: The distribution of the dependent variables should be approximately normal; the distributions of the repeated measures variables should be multivariate normal.
- Linearity: Relationships among repeated measures should be linear.
- Sphericity: the variances of the differences between all combinations of related groups must be equal.

(Warner, 2013, p. 984)

In this study, descriptive statistics and bivariate scatter plot matrices of the dependent variables were used to assess the assumption of univariate and multivariate normality respectively. If the absolute value of skewness and kurtosis of a dependent variable is smaller than 2, the distribution of that variable can be regarded as univariate normal.

Although multivariate normality can be tested statistically, in most cases, bivariate scatter plot matrices are used to visualize the bivariate distribution of the dependent variables. If all the cells in the bivariate scatter plot are in an oval shape, the distribution of all the dependent or repeated measures variables can be regarded as multivariate normal. However, it should be noted that ANOVA and other general linear models are robust again violations of normality assumption.

Pooled within-groups correlation matrices (using Pearson *r* coefficients) were used to assess the linearity assumption. Usually, if the Pearson *r* coefficients, which model a linear relationship, are statistically significant, the relationship between repeated measures can be regarded as linear; if the Pearson *r* coefficients are not significant, the linearity assumption might be violated.

Sphericity is an important assumption in repeated measures ANOVA designs. Repeated measures ANOVA requires that the differences of paired scores (on the repeated measures, i.e., scores on the dependent variables) in all combinations (or cells) of the independent variable levels (also referred to as *treatment*) have equal or similar variances. The sphericity assumption can be regarded as an extension of the homogeneity of variance assumption (or the homoscedasticity assumption) in between-subjects ANOVA. That is, in a between-subjects ANOVA, we expect samples (or groups) that we draw in the statistical analyses to have similar characteristics to the populations being sampled. When all the groups in the analyses share equal or similar variances on the dependent variable, we can infer a significant treatment effect with a lower level of uncertainty if there are any significant differences across the groups. The same logic applies to the assumption of sphericity in that we expect all the groups to share equal or

similar variances to make inferences about the main and interaction effects of the independent variables. Violations of this assumption can result in an inflated *F*-value, which may then lead to a smaller *p*-value and rejection of the null hypothesis. In this case, the Type I error rate in the significance tests is likely to increase, i.e., observing a significant effect or difference when there is not. However, it should be pointed out that the sphericity assumption does not apply to within-subjects ANOVAs that have only two levels.

The Mauchly's sphericity test is a Chi-square (χ^2) test that assesses the sphericity assumption. If the χ^2-value has an associated *p*-value of less than .05, then the sphericity assumption is violated. The degree of violation of sphericity, or the degree to which the sample variance/covariance matrix departs from sphericity, is measured by the epsilon (ε) statistic. "When ε = 1, there is no departure from sphericity. The closer ε is to 1, the less the sample variance/vocariance matrix departs from sphericity (Warner, 2013, p. 988). "

However, if Mauchly's test is significant, there are two common procedures that can be performed to correct for violations of sphericity: the Greenhouse-Geisser correction (1959) and the Huynh-Feldt correction (1976). Both corrections attempt to make downward adjustments to the degrees of freedom (*df*) in the ANOVA test in order to produce a more accurate (or higher) significance (*p*) value and a reduced Type I error rate; however, the F-ratio remains unadjusted. The difference between the Greenhouse-Geisser ε and the Huynh-Feldt ε is that the Greenhouse-Geisser ε is more conservative (i.e., tends to over-correct) whereas the Huynh-Feldt ε is less conservative (i.e., tends to under-correct). There are different opinions in terms of the choice of correction; however, a commonly adopted solution is the criterion recommended by Girden (1992), which

states that when epsilon is > .75, the Huynh-Feldt correction should be applied; and when epsilon is < .75 or nothing is known about sphericity, the Greenhouse-Geisser correction should be applied.

CHAPTER 5. RESULTS AND DISCUSSION

5.1 Statistical Assumptions

5.1.1 Normality

5.1.1.1 Univariate normality

Prior to the repeated measures ANOVA, the data was screened for the statistical assumptions of normality, linearity, and sphericity. Descriptive statistics of AR and NumSP of responses to EI tasks grouped by SL and FS are provided in Table 5.1 and Table 5.2 respectively.

Table 5.1 shows that the distribution of the average articulation rate on the 24 EI tasks was approximately normal (*skewness = 0.04, kurtosis = – 0.02*), with a mean average articulation rate of about 239 syllables per minute ($M_{AR\_AV} = 238.88$, $SD_{AR\_AV} = 19.20$). In addition, the average articulation rate for all conditions (i.e., the six combinations of the levels within SL and FS) was approximately normally distributed, with skewness and kurtosis values within the range between – 2 and 2 (See the last two columns).

With respect to number of silent pause, Table 5.2 shows that the distribution of the total number of silent pause across all EI tasks was approximately normal (*skewness = 0.72, kurtosis = 0.68*), with a mean total number of about 29 pauses per test (M_{NumSP_TT} *= 28.97,* SD_{NumSP_TT} *= 12.22*). In addition, the total number of silent pause for most conditions was approximately normally distributed except for responses on short sentences without formulaic sequences (i.e., NumSP_S_NF; *skewness = 2.08, kurtosis = 7.86*). The distribution of NumSP for short sentences with formulaic sequences (i.e., NumSP_S_F) also slightly deviated from normality, as the absolute values of skewness and kurtosis were close to 2 (*skewness = 1.28, kurtosis = 1.71*). However, ANOVA and other variations of general linear models are robust against violations of the normality assumption as long as the sample size is large (Warner, 2013).

Table 5.1 Descriptive Statistics of AR by SL and FS

	N	***M***	***SD***	***Min***	***Max***	***Skewness***	***Kurtosis***
AR_S_F	194	244.52	30.86	167.7	336.75	-0.05	-0.02
AR_S_NF	194	232.37	27.51	152.55	301.2	0.00	-0.38
AR_M_F	194	239.66	25.79	168.6	301.8	0.12	-0.28
AR_M_NF	194	239.46	24.72	154.65	301.2	-0.36	0.54
AR_L_F	194	240.27	25.81	182.55	339.45	0.34	0.45
AR_L_NF	194	236.98	27.72	175.65	364.05	0.47	1.38
AR_AV	**194**	**238.88**	**19.20**	**188.65**	**305.7**	**0.04**	**-0.02**

Table 5.2 Descriptive Statistics of NumSP by SL and FS

	N	***M***	***SD***	***Min***	***Max***	***Skewness***	***Kurtosis***
NumSP_S_F	194	2.24	2.27	0	11	1.28	1.71
NumSP_S_NF	194	2.35	2.36	0	17	2.08	7.86
NumSP_M_F	194	5.01	2.73	0	15	0.74	0.50
NumSP_M_NF	194	5.44	3.25	0	16	0.67	0.22
NumSP_L_F	194	6.4	3.70	0	20	0.77	0.42
NumSP_L_NF	194	7.54	3.82	0	21	0.89	1.20
NumSP_TT	**194**	**28.97**	**12.22**	**6**	**72**	**0.72**	**0.68**

5.1.1.2 Multivariate normality

Figures 5.1 and 5.2 present the bivariate scatter plot matrix among the repeated measures of AR and NumSP respectively. In terms of AR, the shape of the scatter plot in all cells was approximately oval, indicating that the distributions of repeated measures of AR are multivariate normal. As to NumSP, the shape of the scatter plot in most cells was approximately oval; however, the cells that involved NumSP in responses to short sentences (both with and without formulaic sequences, i.e., NumSP_S_F and NumSP_S_NF) were not in an oval shape, suggesting distributions of repeated measures of NumSP were not multivariate normal when it comes to the repetition of short sentences. The violation of multivariate normality by the distributions of NumSP on short sentences can be attributed to their deviation from univariate normality as, when univariate normality is violated, multivariate normality will be violated as well. However, as ANOVA is robust against violations of normality assumption, the data can still be used to yield reliable results regarding the main and interaction effects of SL and FS on AR and NumSP.

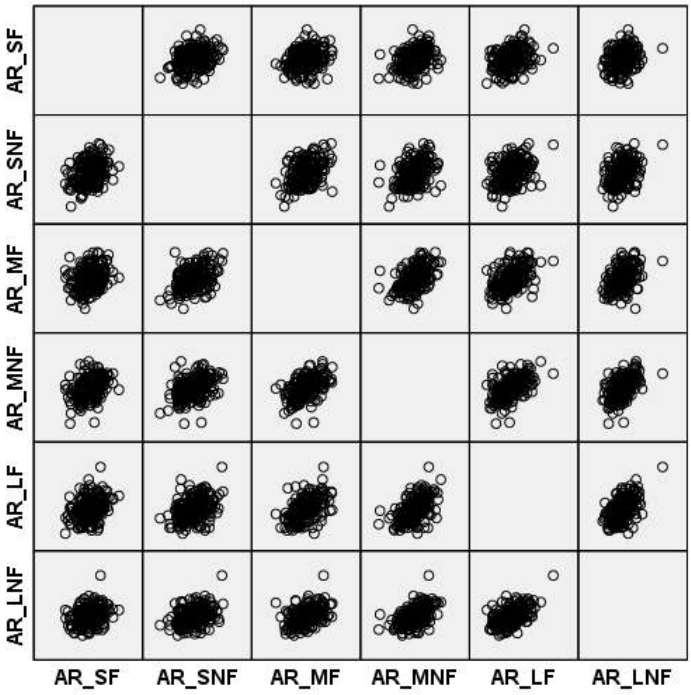


Figure 5.1 Scatter Plot for the Repeated Measures of Articulation Rate

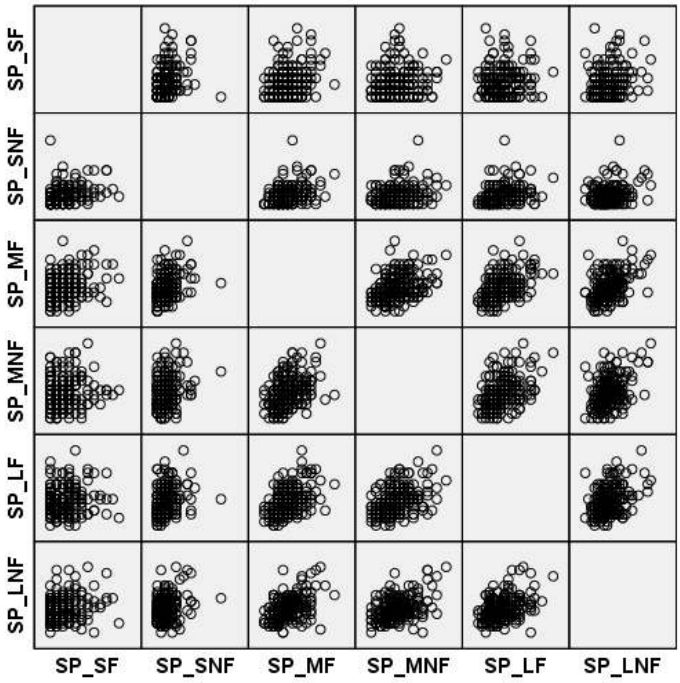


Figure 5.2 Scatter Plot for the Repeated Measures of Number of Silent Pauses

5.1.2 Linearity

A pooled within-groups correlation matrix was computed for both AR and NumSP of responses across the six conditions (see Tables 5.3 and 5.4). In terms of articulation rate, Table 5.3 shows that the average AR across all the six conditions was significantly correlated with each other, with the Pearson *r* coefficient ranging from .26 to .54. Regarding the number of silent pause, Table 5.3 shows that NumSP across all the six conditions was also significantly correlated with each other, with the Pearson *r* coefficient ranging from .19 to .48. The strong and significant linear correlation coefficients among repeated measures for AR and NumSP suggest that the linearity assumption is satisfied.

Table 5.3 Pearson *r* Correlation Coefficients among AR across all Conditions

	AR_S_F	***AR_S_NF***	***AR_M_F***	***AR_M_NF***	***AR_L_F***	***AR_L_NF***
AR_S_F	-	.34**	.27**	.35**	.33**	.26**
AR_S_NF		-	.41**	.41**	.36**	.37**
AR_M_F			-	.50**	.49**	.46**
AR_M_NF				-	.53**	.53**
AR_L_F					-	.54**
AR_L_NF						-

Note. ** $p < .01$.

Table 5.4 Pearson *r* Correlation Coefficients among NumSP across all Conditions

	NumSP_S_F	***NumSP_S_NF***	***NumSP_M_F***	***NumSP_M_NF***	***NumSP_L_F***	***NumSP_L_NF***
NumSP_S_F	-	.34**	.30**	.22**	.22**	.19**
NumSP_S_NF		-	.36**	.25**	.29**	.28**
NumSP_M_F			-	.40**	.48**	.47**
NumSP_M_NF				-	.42**	.46**
NumSP_L_F					-	.41**
NumSP_L_NF						-

Note. ** $p < .01$.

5.1.3 Sphericity

Mauchly's test was performed on AR and NumSP separately to test the assumption of sphericity. For each test, chi-square statistics were used for the main effect of SL and the interaction effect between SL and FS. Results of the Mauchly's test for AR indicated that the assumption of sphericity was violated for both the main effect of SL, $W = .91$, $\chi^2(2) = 18.76$, $p < .001$; and the interaction effect between SL and FS, $W = .94$, $\chi^2(2) = 11.81$, $p < .01$. These results suggest that the observed matrix does not have approximately equal variances and equal covariances, and thus using an uncorrected repeated measures ANOVA F-test would result in a likely inflation of Type I Error. Therefore, degrees of freedom of the repeated measures ANOVA were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon_{SL} = .92$, $\varepsilon_{SL\text{-}FS} = .95$). Similarly, results of the Mauchly's test for NumSP indicated that the assumption of sphericity had also been violated for both the main effect of SL, $W = .90$, $\chi^2(2) = 20.95$, $p < .001$; and the interaction effect between SL and FS, $W = .93$, $\chi^2(2) = 13.81$, $p < .01$. Therefore, degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon_{SL} = .91$, $\varepsilon_{SL\text{-}FS} = .94$).

5.1.4 Form Effect

In addition to the statistical assumptions for repeated measures ANOVA, potential form effect was examined for both AR and NumSP. Results of one-way ANOVAs (see Tables 5.5 and 5.6) indicated that, although participants took different forms of EI tasks,

form did not have a significant effect on the average AR, $F(3, 190) = 1.05$, $p = .37$; and NumSP, $F(3, 190) = 0.85$, $p = .47$.

Table 5.5 One-Way ANOVA for Form Effect on Average AR

Source	***df***	***SS***	***MS***	***F***	***p***
Between groups	3	1160.93	386.98	1.05	.37
Within groups	190	69955.73	368.19		
Total	193	71116.67			

Table 5.6 One-Way ANOVA for Form Effect on Total NumSP

Source	***df***	***SS***	***MS***	***F***	***p***
Between groups	3	379.37	126.46	0.85	.47
Within groups	190	28432.45	149.64		
Total	193	28811.81			

5.2 Significances Tests for Repeated Measures ANOVA

5.2.1 Main and Interaction Effects on AR

Table 5.7 presents results of the F significance tests for the main and interaction effects of SL and FS on AR, with degrees of freedom adjusted using Huynh-Feldt estimates. As shown in the table, the overall F for differences in mean AR across the three levels of SL was not statistically significant, $F(1.85, 356.28) = 0.27$, $p = .74$. The non-significant effect of SL suggests that the increase in sentence length, which is supposed to increase the processing load, did not lead to a decrease of articulation rate in repetition. In other words, when repeating longer sentences, the participants did not speak slower.

In comparison, the overall differences in mean AR across the two levels of FS was statistically significant, $F(1, 193) = 19.32$, $p < .01$. This suggests that participants repeated sentences that contained formulaic sequences at a higher articulation rate than sentences that do not. The corresponding effect size for FS was a partial η^2 of .09. In other words, after stable individual differences in AR of the speakers are taken into account, the production of formulaic sequences accounts for 9% of the variance of AR within the participants' performance on EI tasks.

Nonetheless, there was a significant interaction effect between SL and FS on AR, $F(1.91, 367.76) = 9.29$, $p < .01$, with a partial η^2 of .05. This suggests that, with both individual differences and the main effect of FS considered, the interaction effect accounts for an additional 5% of the variance in AR within the participants' EI performance. However, in total, the main effect of FS and the interaction effect of SL and FS only account for 14% of the variance in AR, suggesting that the processing advantage of FS across different SL bands, albeit statistically significant, does not contribute much to the AR in speech production on EI tasks. These main and interaction effects will be discussed in detail in the following sections.

Table 5.7 Test of Within-subjects Main and Interaction Effects on AR

Source	***SS***	***df₁***	***df₂***	***MS***	***F***	***p***	**Partial η^2**	**Power**
SL	276.72	1.85	356.28	149.90	0.27	.74	.00	.04
Error (SL)	192451.71	1.85	356.28	540.16				
FS	7910.81	1	193	7910.81	19.32	.00	.09	
Error (FS)	79019.82	1	193	409.42				
SL × FS	7460.57	1.91	367.76	3915.32	9.29	.00	.05	
Error (SL × FS)	154915.71	1.91	367.76	421.24				

Note. Degrees of freedom were adjusted using Huynh-Feldt estimates.

5.2.1.1 Main effects of SL and FS on AR

To further discuss the main effects of SL and FS on AR, descriptive statistics and planned contrasts were obtained to compare mean AR from each of the three levels of SL and each of the two levels of FS. Descriptive statistics of mean AR across different levels of SL and FS are provided in Table 5.8; Results of post hoc F tests for the planned contrasts are presented in Table 5.9.

In terms of the main effect of SL, mean AR on short sentences ($M_{SL\text{-}S} = 238.44$, $SE_{SL\text{-}S} = 1.71$) was not significantly different from mean AR on medium-length sentences ($M_{SL\text{-}M} = 239.55$, $SE_{SL\text{-}M} = 1.57$), $F(1, 193) = 0.45$, $p = .50$; neither was mean AR on long sentences ($M_{SL\text{-}L} = 238.62$, $SE_{SL\text{-}L} = 1.68$) significantly different from the mean AR on short and medium-length sentences, $F(1, 193) = 0.08$, $p = .77$. The non-significant effect of SL on AR is counter-intuitive as one would expect that increase of processing load should lead to a decrease of fluency in speech production. In that sense, articulation rate, as an important component of L2 fluency, should be lower on longer sentences than on shorter sentences. On contrary, the observation of similar mean AR on sentences of different length bands suggests that, when processing longer sentences on EI tasks, the participants did not (need to) lower their articulation rate.

As to the main effect of formulaic sequences, mean AR on sentences with formulaic sequences ($M_{FS\text{-}F} = 241.48$, $SE_{FS\text{-}F} = 1.49$) was significantly higher than mean AR on sentences without formulaic sequences ($M_{FS\text{-}NF} = 236.27$, $SE_{FS\text{-}NF} = 1.51$), $F(1, 193) = 19.32$, $p < .01$, with a partial η^2 of .09. This suggests that, on average, when producing sentences with a three- to five-word formulaic sequence, the participants'

articulation rate increased by about five syllables per minute. Five syllables per minute seem to be a small increase of articulation rate at the sentence level; however, the contribution of formulaic language to L2 fluency is substantial given that other domains of linguistic knowledge, e.g., syntactic, semantic, as well as pragmatic knowledge, also contribute to the automaticity of a speaker's general language proficiency. More importantly, the cumulative impact of the presence of formulaic sequence on AR can be large, especially when the speech activity lasts longer than five minutes.

The significant effect of FS on AR is in line with findings of previous studies on the processing advantage of formulaic language. That is, the participants processed the formulaic sequence as holistic units, thereby contributing to the increase of L2 fluency in speech production.

Table 5.8 Descriptive Statistics of AR by Level of SL and FS

		FS-F		FS-NF		Total	
Level of SL	***n***	***M (SE)***	**97. 5% CI**	***M (SE)***	**97.5% CI**	***M (SE)***	**97.5% CI**
S	194	244.52 (2.22)	[239.19, 249.86]	232.37 (1.98)	[227.62, 237.13]	238.44 (1.71)	[234.30, 242.58]
M	194	239.66 (1.85)	[235.20, 244.12]	239.46 (1.65)	[235.18, 243.73]	239.55 (1.57)	[235.77, 243.33]
L	194	240.27 (1.85)	[235.81, 244.73]	236.98 (1.99)	[232.19, 241.77]	238.62 (1.68)	[234.56, 242.68]
Total	194	241.48 (1.49)	[237.89, 245.07]	236.27 (1.51)	[232.63, 239.91]		

Note. CI = confidence interval.

Table 5.9 F Tests for the Planned Contrasts of AR within Levels of SL and FS

Source	Contrast	*SS*	df_1	df_2	*MS*	*F*	*p*	Partial η^2	Power
SL	M vs. S	239.32	1	193	239.32	0.45	.50	.00	.10
	L vs. M and S	28.04	1	193	28.04	0.08	.77	.00	.06
FS	F vs. NF	5273.87	1	193	5273.87	19.32	.00	.09	-

5.2.1.2 Interaction effect of SL and FS on AR

As the interaction effect of SL and FS was also significant, the main effect of FS should be analyzed and interpreted by different levels of SL. According to Table 5.10, mean AR on short sentences (SL-S) with formulaic sequences (FS-F) was significantly higher than mean AR on short sentences without formulaic sequences, $M_{diff\_SL\_S} = 12.15$, $t(193) = 5.05$, $p < .01$, $d = 0.73$; however, the mean differences of AR associated with presence of formulaic sequence were not statistically significant in medium-length sentences, $M_{diff\_SL\_M} = 0.20$, $t(193) = 0.11$, $p = .91$; and long sentences, $M_{diff\_SL\_L} = 3.29$, $t(193) = 1.78$, $p = .08$. The differential effect of FS on AR across levels of SL is also illustrated in the interaction plot shown in Figure 5.3. The horizontal axis represents the two levels of FS, and each of the separated lines represents a corresponding level of SL. As shown in the figure, the blue solid line, which represents the mean AR on short sentences, shows a large drop from short sentences with formulaic sequences to short sentences without formulaic sequences. By contrast, the differences for medium-length sentences (green dotted line) and long sentences (yellow dotted line) were rather small or negligible. Interesting, the mean AR on short sentences without formulaic sequences was lower than the mean AR on medium-length and long sentences without formulaic sequences. This might be partly attributable to the effect of the EI task, which places a

limit on the response time. When repeating longer sentences, participants were pressured to finish repeating within the response time; in contrast, when repeating short sentences, participants needed not rush to repeat the sentence, thereby instinctively slowing down. The task effect necessitates examining the facilitation of formulaic sequences within the same length bands.

Table 5.10 Post Hoc Paired t-tests of AR by Level of FS and SL

Compare	**Condition**	**Mean difference**	***t***	***df***	***p***	***d***	**97.5% CI**
FS-F vs. FS-NF	SL-S	12.15	5.05	193	.00	0.73	[7.40, 16.90]
	SL-M	0.20	0.11	193	.91	-	[-3.38, 3.79]
	SL-L	3.29	1.78	193	.08	-	[-0.35, 6.92]

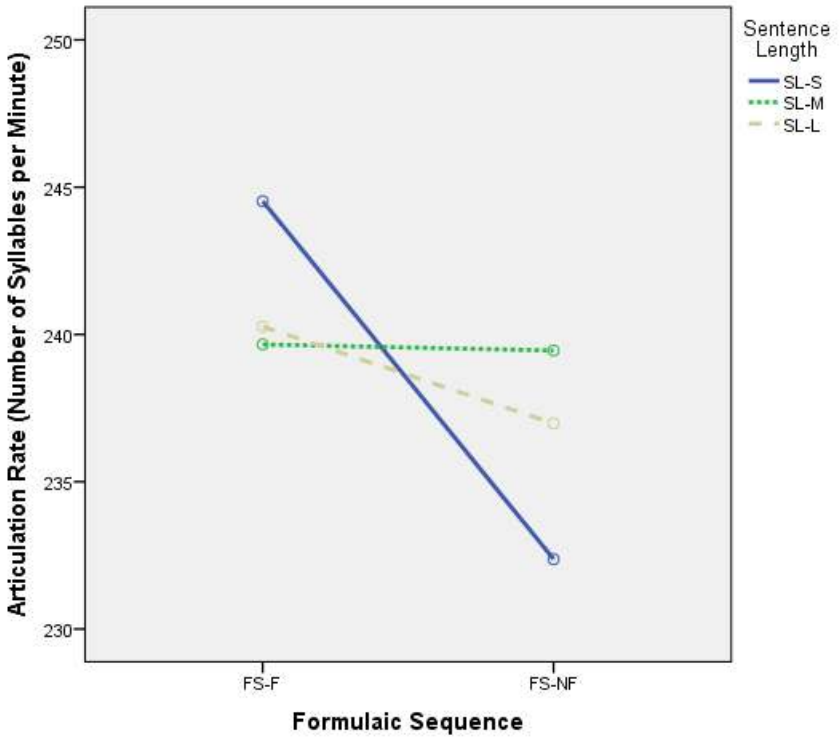


Figure 5.3 Interaction of FS and SL on Average AR

The results of the post hoc paired t-tests indicate that the presence of formulaic sequence only had a significant effect on repetition or production of short sentences, but not medium-length or long sentences. A reasonable explanation for the differential effect of formulaic sequences on AR across sentence length bands is that the increase of articulation rate is attributed to the proportion of formulaic sequences (or the degree of formulaicity) rather than the appearance of FS in the sentences. That is, in short sentences (about eight syllables), a formulaic sequence (typically three-word sequence) can account for almost half of the sentence; therefore, if these formulaic sequences were fixed and processed holistically, i.e., with a faster articulation rate, the average AR of the whole sentence is likely to increase. In this study, the processing advantage of formulaic sequences has led to a substantial increase of 12.15 syllables per minute in the mean AR on repetition of short sentences (see the second row of Table 5.10).

To better illustrate the benefit of fixedness of formulaic sequences on articulation rate, a subsample of 20 responses to two short EI sentences (one with FS and the other without) were analyzed manually through PRAAT. AR for both formulaic sequences and comparable non-formulaic sequences were extracted for comparison. These two sentences, taken from Form 3 of the EI tasks, have similar sentence structures. However, differently, in Sentence (1), the four-syllable phrase *"have a question"* was identified as a formulaic sequence, whereas Sentence (2) did not contain any formulaic sequence. Therefore, the four-syllable phrase "*will take this course*" in Sentence (2) was selected for contrast (referred to as "contrast phrase" hereafter). Additionally, both contrast phrases were followed by an adverbial phrase that had four syllables, i.e., *"about homework"* in Sentence (1) and *"next semester"* in Sentence (2).

(1) I *have a question* about homework.

(2) I will take this course next semester.

Articulation rate was taken for both the contrast phrases and the adverbial phrases for all the 20 speech samples. Descriptive statistics of AR are provided in Table 5.11 below. As shown in the table, the mean AR of Sentence (1) ($M_{FS\text{-}F}$ = *252.73,* $SD_{FS\text{-}F}$ = *32.14*; see the last two columns), which contained the formulaic sequence, was higher than that of Sentence (2) ($M_{FS\text{-}F}$ = *232.27,* $SD_{FS\text{-}F}$ = *25.31*). In particular with respect to the contrast phrases (see the third and fourth columns), the mean AR of the formulaic sequence (i.e., have a question) ($M_{FS\text{-}F}$ = *291.52,* $SD_{FS\text{-}F}$ = *33.44*) was higher than that of the non-formulaic sequence ($M_{FS\text{-}NF}$ = *227.92,* $SD_{FS\text{-}NF}$ = *27.48*).

Table 5.11 Descriptive Statistics of AR on Formulaic vs. Non-formulaic Sequences

		Contrast phrase		**Adverbial phrase**		**Whole sentence**	
	n	***M***	***SD***	***M***	***SD***	***M***	***SD***
Sentence (1)	20	291.52	33.44	229.23	24.62	252.73	32.14
Sentence (2)	20	227.92	27.48	240.63	26.45	232.27	25.31

These findings suggest that the fixedness of formulaic sequences leads to a faster articulation of these sequences. Moreover, the facilitative effect of formulaic sequences is manifested through the proportion of formulaic sequence or the degree of formulaicity in the sentence. That is, when embedded in short sentences, the contribution of these sequences to the AR of the whole sentence is significant. However, when formulaic sequences are embedded in longer sentences, these sequences can account for only a small part of the sentence; in that case, even if the formulaic sequences were processed as holistic units, the contribution of fixedness of formulaic sequence to AR may be lessened by the processing of non-formulaic sequences in the same sentence.

5.2.2 Main and Interaction Effects on NumSP

Table 5.11 presents results of the F significance tests for the main and interaction effects of SL and FS on NumSP, with degrees of freedom adjusted using Huynh-Feldt estimates. As shown in the table, the overall F for differences in mean total number of NumSP across the three levels of SL was statistically significant, $F(1.82, 352.97) = 296.29$, $p < .01$, with a partial η^2 of .61 (see the second row of Table 5.11). Such a strong effect of SL on the total NumSP suggests that the increase in sentence length (and thereby the processing load) led to a substantial increase of the number of silent pauses in speech production. In other words, when repeating longer sentences, the participants became less fluent, as was reflected in the increased number of pauses.

Additionally, the overall differences in mean total NumSP across the two levels of FS was statistically significant, $F(1, 193) = 15.42$, $p < .01$ (see the third row of Table 5.11). This suggests that participants repeated sentences that contained formulaic sequences with fewer pauses than sentences that do not. The corresponding effect size for FS was a partial η^2 of .08. In other words, after the effects of SL and interaction between SL and FS are taken into account, the processing advantage of formulaic sequences accounts for 8% of the variance of NumSP within the participants' performance on EI tasks.

There was also a significant interaction effect between SL and FS on AR, $F(1.88, 364.36) = 4.63$, $p < .017$, with a partial η^2 of .03 (see the third row of Table 5.12). This suggests that, with both the main effects of SL and FS considered, the interaction effect of SL and FS accounts for an additional 3% of the variance in NumSP within the

participants' EI performance. In total, the main and interaction effects of SL and FS account for 72% of the variance in NumSP. These main and interaction effects will be discussed in detail in the following sections.

Table 5.12 Test of Within-subjects Main and Interaction Effects on SP

Source	***SS***	***df₁***	***df₂***	***MS***	***F***	***p***	**Partial η^2**
SL	4327.07	1.82	352.97	2365.99	296.29	.00	.61
Error (SL)	2818.59	1.82	352.97	7.98			
FS	92.42	1	193	92.42	15.42	.00	.08
Error (FS)	1156.58	1	193	5.99			
SL × FS	53.32	1.88	364.36	28.24	4.63	.01	.03
Error (SL × FS)	2219.67	1.88	364.36	6.09			

Note. Degrees of freedom were adjusted using Huynh-Feldt estimates.

5.2.2.1 Main effects of SL and FS on NumSP

To further discuss the main effects of SL and FS on NumSP, descriptive statistics and planned contrasts were obtained to compare mean total NumSP from each of the three levels of SL and each of the two levels of FS. Descriptive statistics of mean total NumSP across different levels of SL and FS are provided in Table 5.13; Results of post hoc F tests for the planned contrasts are presented in Table 5.14.

In terms of the main effect of SL, mean total NumSP on short sentences ($M_{SL\text{-}S}$ = *2.29,* $SE_{SL\text{-}S}$ = *0.13*) was significantly lower than mean NumSP on medium-length sentences ($M_{SL\text{-}M}$ = *5.22,* $SE_{SL\text{-}M}$ = *0.18*), $F(1, 193) = 262.61$, $p < .01$, with a partial η^2 of .58 (see the second row of Table 5.14). Moreover, mean NumSP on long sentences

($M_{SL\text{-}L} = 6.96$, $SE_{SL\text{-}L} = 0.22$) was statistically different from the mean total NumSP on short and medium-length sentences, $F(1, 193) = 322.15$, $p < .01$, with a partial η^2 of .63 (see the third row of Table 5.14). The strong effect of SL on NumSP suggests that increase of processing load by lengthening the sentences leads to a decrease of fluency in speech production, which is reflected in an increased number of silent pauses. On the EI tasks in this study, when the sentence length increased from seven-eight syllables to 20-21 syllables, the total NumSP increased by about five silent pauses, on average more than one pause in each sentence. This substantial increase of pausing with SL indicates that participants spend longer time processing longer sentences, as longer sentences contain more information than do shorter sentences. More importantly, the strong effect of SL on NumSP aligns with arguments for EI as a measure of language proficiency that prompts language comprehension, i.e., processing of syntactic structures and lexical items in the sentence.

As to the main effect of formulaic sequences, mean total NumSP on sentences with formulaic sequences ($M_{FS\text{-}F} = 4.54$, $SE_{FS\text{-}F} = 0.15$) was significantly higher than mean total NumSP on sentences without formulaic sequences ($M_{FS\text{-}NF} = 5.11$, $SE_{FS\text{-}NF} = 0.17$), $F(1, 193) = 15.42$, $p < .01$, with a partial η^2 of .08 (see the last row of Table 5.14). This suggests that, on average, when participants' repeat sentences with a three- to five-word formulaic sequence, the total NumSP decreased by about one pause. The decrease of silent pauses is negligible overall; however, as there was a significant interaction between SL and FS on NumSP, the facilitative effect of formulaic sequences on speech production should be dissected and interpreted separately by level of SL.

Table 5.13 Descriptive Statistics of NumSP by Level of SL and FS

		FS-F		FS-NF		Total	
Level of SL	***n***	***M (SE)***	**97.5% CI**	***M (SE)***	**97.5% CI**	***M (SE)***	**97.5% CI**
S	194	2.24 (0.16)	[1.85, 2.63]	2.35 (0.17)	[1.94, 2.76]	2.29 (0.13)	[1.96, 2.62]
M	194	5.01 (0.19)	[4.53, 5.48]	5.44 (0.23)	[4.88, 6.01]	5.22 (0.18)	[4.79, 5.65]
L	194	6.40 (0.26)	[5.76, 7.04]	7.54 (0.27)	[6.88, 8.20]	6.96 (0.22)	[6.42, 7.51]
Total	194	4.54 (0.15)	[4.17, 4.91]	5.11 (0.17)	[4.69, 5.52]		

Note. CI = confidence interval.

Table 5.14 F Tests for the Planned Contrasts of NumSP within Levels of SL and FS

Source	Contrast	*SS*	df_1	df_2	*MS*	*F*	*p*	Partial η^2
SL	M vs. S	1665.93	1	193	1665.93	262.61	.00	.58
	L vs. M and S	1995.85	1	193	1995.85	322.15	.00	.63
FS	F vs. NF	61.61	1	193	61.61	15.42	.00	.08

5.2.2.2 Interaction effect SL and FS on NumSP

Table 5.15 provides *post hoc* paired t-tests of NumSP between sentences with and without formulaic sequences, separated by SL band. As shown in the table, the mean differences of NumSP associated with presence of formulaic sequence were not statistically significant for short sentences, $M_{diff\_SL\_S}$ = -0.11, $t(193)$ = -0.59, p = .56; or medium-length sentences, $M_{diff\_SL\_M}$ = -0.44, $t(193)$ = -1.84, p = .07 (see the first and second rows of Table 5.15). However, the presence of formulaic sequence made a significant difference in NumSP on long sentences, $M_{diff\_SL\_L}$ = -1.44, $t(193)$ = -3.86, p < .01, d = 0.56. These results suggest that the presence of formulaic sequences can significantly lessen the number of silent pauses (thereby contributing to L2 fluency of sentence repetition or speech production) only on long sentences. More specifically, participants had about one and a half fewer pauses when repeating long sentences with formulaic sequences, i.e., about one less silent pause in every other sentence (since there are eight long sentences, see Table 4.2).

The differential effect of FS on NumSP across levels of SL is also illustrated in the interaction plot shown in Figure 5.4. The horizontal axis represents the three levels of SL, and each of the separated lines represents a corresponding level of FS. As shown in

the figure, the green solid line, which represents the mean total NumSP on sentences without formulaic sequences, shows a steady and substantial increase as the sentence length increases. A similar trend in the mean total NumSP is also present on sentences with formulaic sequences, represented by the blue dotted line. The increase of NumSP with sentence length is also reflected in the significant differences observed from all the paired t-tests of NumSP by level of SL in Table 5.15 (from the fifth to the last row). In addition, the facilitative effect of formulaic sequences on sentence processing and speech production is reflected in the gap of NumSP between the two lines on long sentences.

Table 5.15 Post Hoc Paired t-tests of NumSP by Level of SL and FS

Condition	**Compare**	**Mean difference**	***t***	***df***	***p***	***d***	**97.5% CI**
SL-S	FS-F vs. FS-NF	-0.11	-0.59	193	.56	-	[-0.49, 0.27]
SL-M	FS-F vs. FS-NF	-0.44	-1.84	193	.07	-	[-0.91, 0.03]
SL-L	FS-F vs. FS-NF	-1.14	-3.86	193	.00	0.56	[-1.72, -0.56]
FS-F	SL-S vs. SL-M	-2.77	-12.91	193	.00	1.86	[-3.19, -2.35]
	SL-M vs. SL-L	-1.39	-5.71	193	.00	0.82	[-1.87, -0.91]
	SL-S vs. SL-L	-4.16	-14.13	193	.00	0.60	[-4.74, -3.58]
FS-NL	SL-S vs. SL-M	-3.09	-12.30	193	.00	0.44	[-3.59, -2.60]
	SL-M vs. SL-L	-2.09	-7.88	193	.00	0.30	[-2.62, -1.57]
	SL-S vs. SL-L	-5.19	-18.58	193	.00	0.75	[-5.74, -4.64]

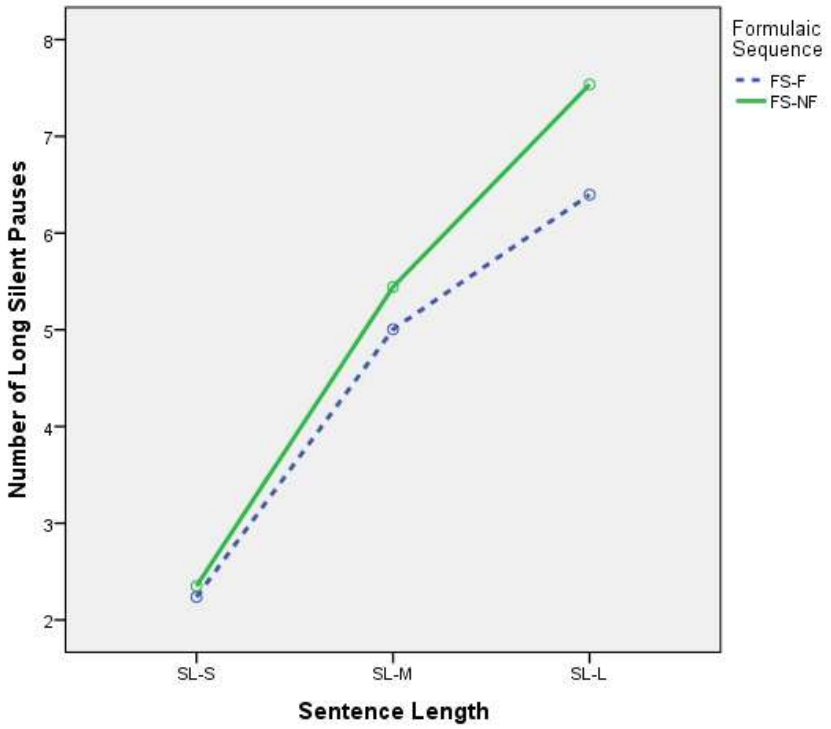


Figure 5.4 Interaction of FS and SL on Total NumSP

A reasonable explanation for the differential facilitation of FS on long sentences is that the presence of formulaic sequences in a sentence helps reduce the processing load in speech production, resulting in fewer silent pauses; however, the facilitation will only take effect when the processing load is high. Due to the holistic processing advantage of formulaic sequences, these sequences are processed as single lexical items (i.e., words). This means, if a 20-word sentence contains a five-word formulaic sequence, then, in terms of processing load, this sentence would practically contain only 16 words, making it easier to process and reconstruct the meaning of the sentence. In contrast, if the sentence is short, i.e., does not require much effort in processing, the facilitation of formulaic sequences may not be triggered.

Another interesting question regarding the facilitation of formulaic sequences in sentence processing lies in when and where the facilitation occurs. In the case of EI tasks, participants have to comprehend the meaning of the sentence first and then reconstruct the sentence. Thus, it is likely that, through these two rounds of language processing, the facilitation of formulaic sequences in speech production occurs not only immediately following but also preceding the formulaic sequence. Intuitively, we may expect that the holistic processing advantage of formulaic sequence facilitates the processing of linguistic information after the formulaic sequence. This expectation is reasonable; however, the facilitation from formulaic sequences may occur in both directions, i.e., holistically. In order to repeat the sentence verbatim, participants have to comprehend the meaning of the sentence as a whole. This means, the reduced processing load due to formulaic sequences can release space in the working memory to help process linguistic information of the whole sentence. Therefore, when comparing repetition of long sentences with and without formulaic sequences, we would expect repetition of long sentences with formulaic sequences to have fewer pauses both before and after the formulaic sequence.

To further examine the possibility of the above explanations, a subsample of 20 responses to two long EI sentences (one with FS and the other without) were analyzed manually through PRAAT to count the number of silent pauses. Sentence (3) and (4) were taken from Form 2 of the EI tasks, and had comparable lexical difficulty. Both sentences had a fronted adverbial clause, but Sentence (3) had an additional complement clause *"that you have a place to live"*. In Sentence (3), the three-syllable phrase *"make*

sure that" was identified as a formulaic sequence, whereas Sentence (4) did not contain any formulaic sequence.

(3) Before you arrive on campus, you need to *make sure that* you have a place to live.

(4) After he worked on the project all evening, the student went directly to bed.

When analyzing the number of silent pauses, both sentences were split in half at the clause boundary. For Sentence (3), the first half was *"before you arrive on campus,"* while the second half was *"you need to make sure that you have a place to live."* For Sentence (4), the first half is *"after he worked on the project all evening"* and the second half is *"the student went directly to bed."* Two NumSP were recorded for each sentence, one for the first half of the sentence, and the other for the second half of the sentence. The second half of Sentence (3) was the clause that contained the formulaic sequence.

Table 5.16 Descriptive Statistics of NumSP on Sentences with vs. without Formulaic Sequences

		First half		**Second half**		**Whole sentence**	
	n	***M***	***SD***	***M***	***SD***	***M***	***SD***
Sentence (3)	20	0.37	0.35	0.53	0.96	1.10	1.22
Sentence (4)	20	0.93	0.97	1.33	0.95	2.29	1.93

Table 5.16 presents descriptive statistics of NumSP on Sentences (3) and (4). As shown in the table, the mean NumSP on Sentence (4) ($M_{FS\text{-}F} = 2.29$, $SD_{FS\text{-}F} = 1.93$; see the last two columns) appeared to be larger than the mean NumSP on Sentence (3) ($M_{FS\text{-}F} = 1.10$, $SD_{FS\text{-}F} = 1.22$). In both sentences, the mean NumSP for the second half appeared to be higher than the mean NumSP for the first half, which seems to suggest that it is more difficult to process the second half of the sentence than the first half. In addition,

the difference of the mean NumSP for the first half (M_{diff} = *0.87,* SD_{diff} = *1.13*) was similar with the difference of the mean NumSP for the second half (M_{diff} = *0.93,* SD_{diff} = *1.33*), suggesting the facilitation effect carries over to other parts as well. The carryover effect of formulaic sequences on NumSP is interesting in that formulaic sequence shows a global effect on NumSP across the whole sentence, but a local effect on AR within the FS region.

However, the differences on NumSP may not be reliable and should be interpreted with caution, as the distribution of NumSP was highly negatively skewed, as can be seen from the large standard deviations, which were as large as the means. The highly skewed distribution of NumSP was a result of many responses that did not have any silent pauses, especially on the first sentence. Therefore, to further examine this explanation, carefully designed experiments should be conducted where the two comparison sentences are the same except for the word(s) within the formulaic sequences (e.g., *the concept of* vs. *the theory of*).

Overall, the results of the main and interaction effects of SL and FS on NumSP indicate that as the length of EI sentence stimuli increases, the processing load involved in sentence repetition increases, resulting in an increase of the number of silent pauses in speech production. In addition, the presence of formulaic sequences helps reduces the number of pauses in speech production, but only on long sentence stimuli. The main and interaction effects of SL and FS account 72% of the variance in NumSP within the participants' performances on the EI tasks.

5.3 Correlation between NumSP and AR

For the 194 participants in this study, the correlation between the average AR (*M = 238.88, SD = 19.20*) and the total NumSP (*M = 80.89, SD = 6.90*) on the EI tasks was significant, r (194) = -.26, $p < .01$. According to Cohen's (1988) guidelines for interpreting effect sizes, a threshold for a medium effect is .3, and .5 for a large effect. Although the correlation between AR and NumSP on EI tasks is smaller than .3, the relationship is substantial as the participants in this study, albeit a representative sample of undergraduate ESL students in US universities, only represented a restricted range of language proficiency level among all L2 speakers. However, the correlation was not extremely strong so that the two variables were functioning as one. Therefore, the correlation coefficient suggests that both variables should be considered when evaluating performances on EI tasks, the processing of formulaic sequences, or L2 fluency in general.

Table 5.15 Descriptive Statistics with Pearson *r* Coefficient for AR and NumSP

Variable	***n***	***M***	***SD***	***r***	
AR_AV	194	238.88	19.20	-	
NUMSP_TT	194	28.97	12.22	-.26**	-

5.4 Summary and Discussion of Findings from Repeated Measures ANOVA

Results of analyses of EI performances showed that both SL and FS had a significant effect on L2 fluency in speech production; however, these two variables had differential effects on different components of L2 fluency, i.e., AR and NumSP. Overall,

FS helped more with AR than with NumSP; but the SL only made a difference in NumSP. Specifically, SL had a strong effect on NumSP on EI performances; the presence of FS had a smaller but substantial effect on the processing of individual sentences. Moreover, this effect is cumulative and can result in a big difference in L2 fluency when the speech production lasts for a longer duration of time.

The strong effect of SL on sentence processing and L2 fluency was only reflected in the increase of NumSP as SL increased. On the EI tasks, longer sentences contain more syntactic and lexical information, which requires a higher level of automaticity or fluency from the speaker to process the linguistic information. When the speaker has a relatively low level of L2 fluency or general L2 proficiency, longer sentences tend to be more difficult to process, thus resulting in more silent pauses. Therefore, the effect of SL on NumSP is an indication that EI tasks prompt language comprehension, i.e., the processing of linguistic information in the sentences.

The presence of FS had a smaller effect on L2 fluency compared with SL, but this variable had more interesting interaction effects with SL on both AR and NumSP. First, the proportion of FS helped increase AR of speech production. In other words, as the language production and use becomes more formulaic, the faster the articulation rate will be. Second, the presence of FS facilitated the processing of sentences, by reducing the number of silent pauses. This facilitative effect became strong and significant when it came to long sentences. This means, the presence of formulaic sequences helps lessen the number of silent pauses, but the effect is only significant at the long sentences. Thus, FS is more likely to help maintain the level of L2 fluency when the processing load is large. Both the faster articulation rate on formulaic sequences and the fewer number of silent

pauses on sentences with formulaic sequences align with the processing advantage of formulaic sequences, which allows these sequences to be articulated faster and helps the speaker to maintain the level of L2 fluency when the processing load is high. More interestingly, with respect to the sentence-level processing, formulaic sequence appeared to have a global effect on NumSP across the whole sentence, but a local effect on AR within the FS region.

Finally, the correlation between AR and NumSP confirms that rate and pausing are distinct but related components of L2 fluency. These two features of L2 fluency can be applied equally well to the evaluation of condition language tasks as they are in language tasks that involve less conditioned speech production. In addition, these two variables can serve as outcome variables in the investigations of performances on EI tasks and the processing of formulaic sequences.

CHAPTER 6. CONCLUSIONS AND IMPLICATIONS

The present study investigated the processing of formulaic language, as an effort to examine how the use of formulaic language may or may not contribute to L2 fluency in speaking performance. To examine the effect of formulaic language on L2 fluency, this study utilized EI tasks designed to measure general English language proficiency to elicit repetition of individual sentences containing formulaic language in comparison with repetition of sentences that do not. In addition to the presence of formulaic language, length of stimulus sentences was included as the other independent variable of interest in this study. Responses to EI tasks were automatically measured on articulation rate and the number of silent pauses. Repeated measures ANOVAs were conducted to examine the effects of formulaic sequences and sentence length on two measures of L2 fluency, i.e., articulation rate and the number of silent pauses.

Findings of this study suggest that formulaic sequences and sentence length have differential effects on L2 fluency in speaking performance. In terms of sentence length, as the stimulus sentence becomes longer, thereby increased processing load, the number of silent pauses on EI performances increases. With respect to formulaic sequences, increase of the proportion of formulaic sequences in language use contributes to faster articulation rate, while the processing advantage of formulaic sequences helps reduce the number of silent pauses when the processing load is large.

6.1 Processing of Formulaic Sequences

Although the effect of formulaic sequences is smaller than the effect of sentence length on fluency features of EI performances, the contribution of formulaic language use to L2 fluency in speaking may be more important that observed. Because the present study examined performance on EI tasks, a more conditioned type of language task, participants were only required process and reconstruct the sentences rather than construct sentences anew in terms of content. Therefore, the facilitative effect of formulaic language use is expected to increase, as, in free speech production where the speech tends to be longer and more complex, the processing load is larger. Therefore, findings regarding the processing advantage of formulaic language bear important implications for language teaching and learning.

The acquisition of formulaic sequences is believed to facilitate the development of L2 fluency. Previous research has shown that L1 speakers process formulaic sequences faster than non-formulaic sequences. However, this study has indicated that the processing advantage also applies to L2 speakers. Based on findings of this study, the teaching of formulaic sequences is recommended in language classes, especially those with an emphasis on speaking skills. Language teachers can benefit from a variety of published lists of formulaic sequences (including collocations, lexical bundles), mostly identified in a corpus-based approach. However, the teaching of formulaic sequences should not simply focus on the speech. That is, the facilitation of formulaic sequences on fluency cannot be separated from appropriate use of these formulaic sequences.

Inappropriately used formulaic sequences will result in ineffective communication as much as lack of fluency in speech production.

6.2 EI as a Measure of L2 Proficiency

Another important finding of this study relates to the appropriateness of EI as a measure of L2 proficiency. A main criticism of EI lies in whether or not EI tasks prompt language comprehension and processing. Results of both the meta-analysis of EI studies and repeated measures ANOVA on EI performances contribute to clarifying the underlying construct measured by EI, i.e., the processing of linguistic information in the sentences.

The quantitative meta-analysis of 10 studies suggests that EI tasks can be used to effectively distinguish performances of higher and lower proficiency speakers. Additionally, the EI tasks used in this study has demonstrated their potential to examine general language proficiency and the processing of formulaic sequences. The economic development and administration procedures for EI tasks make EI a desirable complement to more interactive or less conditioned performance measures, which tend to be more time-consuming, expensive, and less reliable than psycholinguistic measures.

In addition, both the meta-analysis and repeated measures ANOVAs pointed to sentence length as a strong predictor of the difficulty of EI tasks. In the meta-analysis, sentence length was identified as one of the three potential moderators for the sensitivity of EI as a measure of L2 proficiency. In the repeated measures ANOVA, the strong effect of sentence length on the number of silent pauses in speech production aligns with findings of the meta-analysis in that the longer the sentence is, the more difficult it is to

process and repeat the sentence. These findings confirm previous research of EI, which showed sentence length as the strongest predictor of EI task difficulty. Future development of EI tasks should consider tailoring EI tasks on sentence length to align with the target proficiency levels of L2 speakers.

6.3 Recommendations for Future Research & Test Development

Based on findings and limitations of this study, a few recommendations can be drawn, in particular with respect to the investigation of the processing of formulaic sequences and the development and use of EI tasks to measure L2 proficiency.

First, the results of this study were based on participants who represent a restricted range of proficiency level in the population of L2 speakers. As the participants were enrolled in a university-level EAP course, their English language proficiency levels can be regarded as intermediate or lower intermediate. The unique characteristics of formulaic sequences (i.e., fixedness and holistic storage) indicate that the speaker has to become automatic at using these sequences so that they can be stored and accessed as holistic units. High proficiency speakers tend to have a high level of automaticity or L2 fluency, and thereby have acquired a larger pool of formulaic sequences than low proficiency speakers. Therefore, to fully examine the processing advantage of formulaic sequences, EI performances of low and advanced L2 speakers should also be included.

Second, sentences selected in this study were controlled on lexical, phonological, and syntactic complexity, which helped reduce a number of potential confounding variables; however, to further examine the facilitative effect of formulaic sequences on pausing, future researchers should consider comparing sentences that are identical except

for the slot where the formulaic sequence is inserted. A possible way to test holistic processing of formulaic sequences is to change one word in a formulaic sequence that would result in a non-formulaic sequence while keeping similarities in the semantic elements of the sentence, e.g., *the concept of* vs. *the theory of*.

Third, future research can benefit from investigating the sociocultural functions of formulaic sequences in relation to their cognitive functions (i.e., effects on fluency features). Current research of formulaic sequences has mostly focused on the cognitive functions of formulaic language, with few studies looking into the sociocultural functions of formulaic language use. More research is needed to investigate the extent to which the use of formulaic language can facilitate ESL students' adjustment and socialization, e.g., perception of membership and identity, in the speech communities they subscribe to.

Fourth, although this study suggests that EI tasks prompt language comprehension and language processing, other measures should be explored to provide further evidence that EI measures language comprehension, not parroting. Different forms of advanced technology, such as eye-tracking techniques, can be incorporated in the design of EI tasks, to better demonstrate how L2 speakers of different proficiency levels process and reconstruct the meaning of the sentences.

Finally, future research should examine fluency features in relation to accuracy features of performances on EI tasks. Research efforts in this direct are beneficial to the feasibility of developing an automated scoring system for EI tasks. If automated features can be used to predict accuracy features of EI performances, EI tasks with an automated scoring system will be a desirable candidate for an efficient and effective measure of L2 proficiency.

REFERENCES

REFERENCES

*Akakura, M. (2012). Evaluating the effectiveness of explicit instruction on implicit and explicit L2 knowledge. *Language Teaching Research*, *16*(1), 9–37. doi: 10.117/1362168811423339.

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1985). *Cognitive Psychology and Its Implications (2nd edition)*. New York, NY: Freeman.

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly, 28*(4), 807–812.

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *The Modern Language Journal, 86*(4), 525–545.

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In K Spence & J Spence (Eds.). *The psychology of learning and motivation: Advances in research and theory (Vol. 2)*. New York: Academic Press. Baddeley & Gathercole.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bailey, N., Eisenstein, M., & Madden, C. (1976). The development of wh-questions in adult second-language learners. *On TESOL*, *76*, 350–362.

Bernstein J., van Moere A., & Cheng J. (2010). Validating automated speaking tests. *Language Testing, 27*(3): 355-377.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*, 371–405.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence, in (eds) *Research methodology in second-language acquisition* (pp.245-261). illsdale, NJ: L. Erlbaum Associates.

Bloomfield, L. (1993). *Language.* London: Allen & Unwin.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein H. (2007). Comprehensive meta-analysis: A computer program for meta-analysis [Computer software]. Englewood, NJ: Biostat Inc.

Boers, F., Brussels, J. E., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.05, retrieved 24 January 2014 from http://www.praat.org/.

*Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition, 33,* 247– 271. doi: 10.1017/S0272263110000756.

Burdelski, M., & Cook, H. M. (2012). Formulaic language in language socialization. *Annual Review of Applied Linguistics*, 32, 173-188.

*Burger, S., & Chretien, M. (2001). The development of oral production in content-based second language courses at the University of Ottawa. *The Canadian Modern Language Review / La revue canadienne des langues vivantes, 58*(1), 84-102. doi: 10.3138/cmlr.58.1.84.

Butler, C. (2003). Multi-word sequences and their relevance for recent models of functional grammar. *Functions of Language 10*, 179-208.

*Campfield, D. E., & Murphy, V. A. (2014). Elicited imitation in search of the influence of linguistic rhythm on child L2 acquisition. *System, 42*, 207–219. doi: 10.1016/j.system.2013.12.002.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*: 1–47.

Cenoz, J. (1998). *Pauses and communication strategies in second language speech.* ERIC Document ED 426630. Rockville, MD: Educational Resources Information Center.

Chaudron, C., & Russel, G. (1990). *The validity of elicited imitation as a measure of second language competence.* Paper presented at the ninth World Congress of Applied Linguistics, Thessaloniki, Greece.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing, 29*(3), 421-442.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral sciences.* Hove and London: Lawrence Erlbaum Associates.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72-89.

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics, 3,* 239-266.

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Thinking and Verbal Behavior, 11*, 671-684.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs: A case for contrastive analysis. *Studies in Second Language Acquisition, 7,* 73–79.

Dailey, K., & Boxx, J. (1979). A comparison of three imitative tests of expressive language and a spontaneous language sample. *Language, Speech, and Hearing Services in Schools, 10*, 6-13.

de Jong, N., & Wempe, T. (2009). Praatscript to detect syllable nuclei and measure speech rate automatically. *Behavior research methods, 41*(2),385-390.

Eisenstein, M., Bailey, N., & Madden, C. (1982). It take two: Contrasting tasks and contrasting structures. *TESOL Quarterly, 16*(3), 381-393.

*Elliot, R. A. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania, 80*(1), 95-108.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *SSLA, 24*, 143–188.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*, 141-172. doi: 10.1017/S0272263105050096.

*Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*, 339-368. doi: 10.1017/S0272263106060141.

*Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied linguistics, 27*(3), 464-491. doi: 10.1093/applin/aml001.

*Erlam, R., & Loewen, S. (2010). Implicit and explicit recasts in L2 oral French interaction. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, *66*(6), 887-916. doi: 10.3138/cmlr.66.6.877.

*Faqeih, H. I. (2012). *The effectiveness of error correction during oral interaction: Experimental studies with English L2 learners in the United Kingdom and Saudi Arabia.* Unpublished dissertation: University of York.

Fillmore, C. (1979). On fluency. In C. Fillmore, D. Kempler, & W. Wang (Eds.), *Individual Differences in Language Ability and Language Behavior (pp. 85-102).* New York: Academic Press.

*Fiori-Agoren, M. L. (2004). *The development of grammatical competence through synchonous computer mediated communication.* Unpublished dissertation: Pennsylvania State University.

*Flynn, S. (1986). Production vs. comprehension: Differences in underlying competences. *Studies in Second Language Acquisition*, *8,* 135-164.

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar (pp. 134-184).* Cambridge: Cambridge University Press.

Fox, B. A. (2001). An exploration of prosody and turn projection in English conversation. In M. Selting, & E. Couper-Kuhlen, (Eds.), *Studies in interactional linguistics (pp. 287-215).* Amsterdam: John Benjamins.

Fraser, C., Bellugi, U., & Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of verbal learning and verbal behavior*, *2*(2), 121-135.

Fulcher, G. (2000). The 'communicative' legacy in language testing. *System*, *28*(4), 483-497. doi: 10.1016/S0346-251X(00)00033-6.

Gallimore, R., & Tharp, R. G. (1981). The interpretation of elicited sentence imitation in a standardized context. *Language Learning*, *31*(2), 369-392.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove, England: Lawrence Erlbaum Associates Ltd.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing, 27*(3), 379–399.

Girden, E. (1992). *ANOVA: Repeated measures*. Newbury Park, CA: Sage.

Graham, J.G. (1987). English language proficiency and the prediction of academic success. *TESOL Quarterly, 21*(3), 505-521.

Graham, R., McGhee, J., & Millard, B. (2010). The role of lexical choice in elicited imitation item difficulty. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.). *Selected Proceedings of the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices (pp. 57-72).* Somerville, MA: Cascadilla Proceedings Project.

Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95–112.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning, 24*(2), *287-297.*

Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language & Speech, 20*(1), 86-97.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech, 14,* 277–288.

Heatley, A., & Nation, P. (1994). Range. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Hedges, L. V. (1981). Distribution theory for Class's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107-128.

Hedges, L. V., & Olkin, I. (1985). *Statistical model of meta-analysis.* New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods. 3*(4), 468-504. doi: 10.1037/1082-989X.3.4.486.

Henning G. (1983). Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language learning*, *33*(3), 315-332.

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal, 327,* 557-560. doi: 10.1136/bmi.327.7414.557.

Hood, L., & Lightbown, P. (1978). What children do when asked to "say what I say"--Does elicited imitation measure linguistic knowledge? *Allied Health and Behavioral Sciences, 1*(2), 195-219.

Huitt, W. (2003). The information processing approach to cognition. *Educational Psychology Interactive*. Valdosta, GA: Valdosta State University. Retrieved on February 09, 2014, from, http://www.edpsycinteractive.org/topics/cognition/infoproc.html

Huynh, H., & Feldt, L.S. (1976). Estimation of the Box correction for degrees of freedom from sample data in randomized block and split=plot designs. *Journal of Educational Statistics, 1*, 69–82.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27,* 4–21.

IBM Corp. (2012). IBM SPSS Statistics for Windows (Version 21.0) [computer software]. Armonk, NY: IBM Corp.

*Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly, 3*, 151-169. doi:10.1207/s15434311laq0302_4.

*Jensen, E. D., & Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Language Learning, 53*(3), 373-428. doi: 10.1111/1467-9922.00230.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*(3), 433-445.

Johnson, P. (1988). English language proficiency and academic performance of undergraduate international students. *TESOL Quarterly 22*, 164-168.

Kaplan, T. I. (1996). *Elicited imitation in L2 research: lessons from native speakers.* Unpublished manuscript, University of Iowa.

*Kim. J. (2012). *The optimal conditions for form-focused instruction: Method, target complexity, and types of knowledge*. Unpublished dissertation: Georgetown University.

Kormos, J., & Denes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32,* 145-164.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Larsen-Freeman, D. E. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly*, 9(4), 409-419.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Levelt, W. (1989). *Speaking: From Intention to Articulation.* Cambridge, MA: MIT Press.

Lewis, M. (2000). *Teaching collocation: Further development in the lexical approach*. Hove, England: Heinle & Heinle.

Lewis, M. (2009). *The idiom principle in L2 English: Assessing elusive formulaic sequences as indicators of idiomaticity, fluency, and proficiency*. Saarbrucken, Germany: VDM Verlag Dr. Muller.

**Li, S. (2010). *Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability.* Unpublished dissertation: Michigan State University.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning, 54*(2), 193-226.

Lonsdale, D., & Christensen, C. (2011). *Automating the scoring of elicited imitation.* Proceedings of the ACL-HLT/ICML/ISCA Joint Symposium on Machine Learning in Speech and Language Processing.

Markman, B. R., Spilka, I. V., & Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Language Learning, 25*(1), 31-41.

McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The Use of Elicited Imitation as a Measure of Expressive Grammar: A Question of Validity. *Journal of Speech and Hearing Disorders*, *47*(1), 19-24.

McLaughlin, B. (1987). *Theories of Second Language Learning.* London: Edward Arnold.

McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11,* 113-128.

McLaughlin, B., Rossman, T., & McLeod, B. (1983). Second language learning: An information processing perspective. *Language Learning*, *33*, 135-157.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement (pp. 13-103)*. New York: American Council on Education/Macmillan.

Menyuk, P. (1964). Comparison of grammar of children with functionally deviant and normal speech. *Journal of Speech, Language, and Hearing Research*, *7*(2), 109-121.

Menyuk, P. (1971). *The acquisition and development of language.* Englewood Cliffs. N. J.: Prentice-Hall.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81-97.

Miller, J. F. (1973). Sentence imitation in pre-school children. *Language and Speech, 16*, 1-14.

Morrow, K., 1979. Communicative language testing: revolution of evolution? In C.K. Brumfit K. Johnson (Eds.), *The communicative approach to language teaching (pp. 143-159)*. Oxford: Oxford University Press.

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. E. Tarone, S. Gass & A. D. Cohen (Eds), *Research methodology in second-language acquisition (pp. 227-243)*. Hover: Lawrence Erlbaum.

Naiman, N. (1974). *Imitation, comprehension and production of certain syntactic forms by young children acquiring a second language.* Unpublished dissertation: University of Toronto.

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris & L. Ortega (Eds), *Synthesizing research on language learning and teaching (pp. 3-50).* Amsterdam: John Benjamins.

O'Malley, J., & Chamot, A. (1990). *Learning Strategies in Second Language Acquisition.* Cambridge: Cambridge University Press.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*(1), 105-118.

Oller, J. W. (1976). Evidence for a general language proficiency factor: An expectancy grammar. *Die neueren sprachen, 75*(2), 165-174.

Ortega Alvarez-Ossorio, L. (2000). Understanding syntactic complexity: The measurement of change in the syntax instructed L2 Spanish learners. Unpublished dissertation: University of Hawi'i.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication (pp. 191-226)*. London: Longman.

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language learning, 36*(2), 125-141.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*, 655-687.

Poetry Soup. (2013). *SendGrid Syllable Counter*. Retrieved from http://www.poetrysoup.com/haiku_syllable_counter/.

Popham, W. J. (2003) *Classroom assessment: What teachers need to know*. Boston, MA: Pearson.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57,* 1-23.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning, 63*(3), 595–626. doi: 10.1111/lang.12010.

Rebuschat, P., & Mackey, A. (2013). Prompted production. In C. A. Chappelle (Ed.). *The encyclopedia of applied linguistics (vol. 5)*. Oxford: Wiley-Blackwell.

Richards, D. R. (1980). Problems in elicited unmonitored speech in a second language. *The Interlanguage Studies Bulletin – Utrecht*, 5(2), 63-98.

Rumelhart, D.E., Hinton, G.E., & McClelland, J.L. (1986). A General Framework for Parallel Distributed Processing. In Rumelhart, D.E., & McClelland, J.L. and the PDP Research Group (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 1: Foundations. MIT Press: Cambridge, MA.

Rumelhart, D., & McClelland, J. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use (pp. 127–152).* Amsterdam: Benjamins.

Schmitt-Gevers, H. (1993). La Notion d'aisance dans la production et la reception orales en langue etrangere. Melanges – *Centre de Recherches et d'Applications pedagogiques en Langues, 21*, 129-148.

Schimke, S. (2011). Variable verb placement in second-language German and French: Evidence from production and elicited imitation of finite and nonfinite negated sentences. *Applied Psycholinguistics, 32*, 635-685.

*Serafini, E. (2013). *Cognitive and psychosocial factors in the long-term development of implicit and explicit second language knowledge in adult learners of Spanish at increasing proficiency*. Unpublished dissertation: Georgetown University.

Shiffrin, R. M. & Schneider, W. (1977). Controlled and automatic human information processing II: Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*, 127-190.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Sinclair, J. (1987). Collocation: a progress report. In R. Steele & T. Treadgold (eds.) *Essays in honor of Michael Halliday (pp. 319-331).* Amsterdam: John Benjamins.

Slobin, D., & Welsh, C. A. (1973). Elicited imitation as aresearch tool in developmental psycholinguistics. In C. Ferguson & D. Slobin (Eds.), *Studies of child language development (pp. 485-489).* New York: Holt, Rinehart and Winston Inc.

Sperber, D., & Wilson, D. (1996). Précis of *Relevance: communication and cognition*. In H. Geirsson & M. Losonsky (Eds.). *Reading in Language and Mind (pp. 460-486)*. Cambridge, MA: Blackwell.

Spitze, K. & Fischer, S. D. (1981). Short-term memory as a test of language proficiency. *TESL talk, Quarterly for Teachers of English as a Second Language*, 12(4), 32-41.

Stivers, T., & Robinson, J. D. (2006). A preference for progressivity in interaction. *Language in Society, 35*, 367-392.

Thomas, M. (1992). *What do elicited imitation data reveal about comprehension?* Proceedings of the twelfth second language research forum, April 2-5, 1992, Michigan State University.

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning, 44*, 307-336. doi: 10.1111/j.1467-1770.1994.tb01104.x.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching (pp. 3-50).* Amsterdam: John Benjamins.

Towell, R. & Hawkins, R. (1994). *Approaches to Second Language Acquisition.* Clevedon: Multilingual Matters.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advance learners of French. *Applied Linguistics, 17*, 84-119.

Tracy-Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (in press). "Repeat as much as you can": Elicited imitation as a measure of global proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Multilingual Matters*.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: “Clozing” the gap. *Studies in Second Language Acquisition, 33*, 339-372.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language learning*, 61(2), 569-613.

Trofimovich, P., & Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Applied Psycholinguistics, 28*(2), 251-276. doi: http://dx.doi.org/10.1017/S0142716407070130 .

*Trofimovich, P., Lightbown, P. M., Halter, R. H. & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, *31*, 609-639. doi: 10.1017/S0272263109990040.

* Trofimovich, P., Lightbown, P. M., & Halter, R. (2013). Are certain types of instruction better for certain learners?. *System*, *41*(4), 914-922. doi: 10.1016/j.system.2013.09.004.

Underhill, N. (1987). *Testing Spoken Language: A Handbook of Oral Testing Techniques.* Cambridge: Cambridge University Press.

Underwood, G., & Schmitt, N., & Galpin, A. (2004). The eye have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use (pp. 127–152).* Amsterdam: Benjamins.

Ushigusa, S. (2008). *The relationships between oral fluency, multiword units, and proficiency scores.* Unpublished doctoral dissertation: Purdue University.

van Boxtel, S., Bongaerts, T., & Coppen, P. (2005). Native-like attainment of dummy subjects in Dutch and the role of the L1. *IRAL, 43,* 355-380. doi: 10.1515/iral.2005.43.4.355.

van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing,* 29(3), 325–344. doi:10.1177/0265532211424478.

Verhagen, J. (2011). Verb placement in second language acquisition: Experimental evidence for the different behavior of auxiliary and lexical verbs. *Applied Psycholinguistics, 32*, 821-858.

Vinke, A. A., & Jochems, W. M. G. (1993). English proficiency and academic success in international postgraduate education. *Higher Education, 26*(3), 275-285.

Vinther, T. (2002). Elicited imitation: a brief overview. *International Journal of Applied Linguistics, 12*(1), 54-73. doi: 10.1111/1473-4192.00024.

Wait, I. W., & Gressel, J. W. (2009). Relationship between TOEFL scores and academic success for international engineering students. *Journal of Engineering Education, 98*(4), 389-398.

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques (2nd edition)*. New York, NY: SAGE Publications.

*West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition, 4*(3), 203-222. doi: 10.1515/langcog-2012-0011.

Wimberley, D. W., McCloud, D. G., & Flinn, W. L. (1992). Predicting success of Indonesian graduate students in the United States. *Comparative Education Review, 36*(4), 487-508.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication, 18,* 47-67.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.

*Wu, S. & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annuals, 46*(4), 680-704. doi: 10.1111/flan.12063.

Xu, M. (1991). The impact of English-language proficiency on international graduate students' perceived academic difficulty. *Research in Higher Education, 32*(5), 557-570.

*Yoon, S. (2010). *English syllable confusion and imitation in Korean bilingual and monolingual children and adults.* Unpublished dissertation: University of Illinois at Urbana-Champaign.

*Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language.* Unpublished dissertation: University of Hawaii at Manoa.

APPENDIX

APPENDIX

Directions and sample elicited imitation tasks

Introduction. In this section, you will hear 12 sentences. Each sentence will be played once. After each sentence, the screen will change and two words will appear. One of the two words was mentioned in the sentence.

Task. your task is to (1) identify the word that was mentioned in the sentence, then (2) repeat the sentence that you heard. Try to repeat the sentence exactly as it was stated.

Preparing for your response. Listen to each sentence carefully. You will have 5 seconds to choose the word and 20 seconds to repeat each sentence.

Sample Items:

You will hear the following sentence:

Parking on campus is free on Sunday. (AUDIO ONLY)

Click on the word below that you heard the sentence? *(CLICK ON WORD)*

Parking	Swimming

The word mentioned in the example sentence was *Parking*. So you should have clicked on *Parking.*

OK, now **repeat the sentence you heard after you hear a voice that states, "recording now":**

Is parking on campus free on Sunday?

Sample sentence stimuli and words

1. Purdue University was founded in 1869. (founded // wanted)
2. Purdue offers both undergraduate and graduate programs. (programs // letters)
3. All students at Purdue have access to computer lab printers. (time // access)
4. There are a number of options for on and off campus housing for students. (were // and)
5. Students living in the undergraduate residence halls are required to purchase a meal plan. (living // visiting)

VITA

VITA

Xun Yan

Oral English Proficiency Program

Second Language Studies, Department of English

Purdue University

West Lafayette, IN

xyanacademic@gmail.com

Education

- PhD, English: Language Testing and Assessment. Purdue University, 2015.
- MA, TESOL. The Ohio State University, 2009.
- BA, English Language and Literature. Wuhan University, China, 2008.

Work

- Instructor, Oral English Proficiency Program. Purdue University. August 2013 – present.
- Graduate Testing Coordinator, Oral English Proficiency Program. Purdue University. July 2011 – July 2013.
- Composition Instructor, Department of English, Purdue University. August 2010 – May 2011.