Winter 12-15-2018

Intersecting Variables of Second and Foreign Language Reading: Self-Assessment and Comprehension with Adolescents and Adults Across Languages

Haley Dolosic
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Education

Dissertation Examination Committee:
Cindy Brantmeier, Chair
Joe Barcroft
John Baugh
Andrew C. Butler
Michael Strube

Intersecting Variables of Second and Foreign Language Reading: Self-Assessment and Comprehension with Adolescents and Adults Across Languages

by
Haley Dolosic

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2018
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I must acknowledge my students and participants who both inspired and supported the research presented in these pages. Without them, the initial spark of this research would not have existed! Furthermore, their drive to learn a second language will continue to motivate me for years to come. I'd also like to thank the Educational Testing Service (ETS) for their financial support in the form of a Small Grants for Doctoral Research in Second or Foreign Language Assessment. In addition, I'd like to thank my research assistant, Andrea Zuzarte, for her contributions to the work presented in these pages in the form of hours of detailed work and energizing dedication to data.

For this amazing scholarly experience, I would like to express my gratitude to Dr. Cindy Brantmeier, who invited me to become part of this club of researchers and educators in the exciting field of Applied Linguistics. I thank her for her investment in my scholarly development, her kind guidance, her thoughtful connections of myself to resources, and her encouragement to follow the questions that I had until I found the answers. I would also like to thank Dr. Michael J Strube, who offered endless support and guidance through countless approaches to these studies. In addition, I would like to thank Dr. Andrew C. Butler for his candor about writing, research, and much more. Likewise, I would like to thank Dr. Joe Barcroft, who helped me to understand the field and enabled me to implement the study of my dreams. I would also like to thank Dr. John Baugh for sharing his experiences in the field and offering full support throughout my doctoral work. Further, I'd like to thank the faculty and staff of the Education Department for helping to make this dissertation possible. I would also like to thank

the French educators in my life who assured me that I, too, could become a teacher, particularly Mme. Driesner, Mme. Uzzell, and Dr. Strange.

I would also like to acknowledge the contribution of my family, my friends, and my colleagues. I thank my parents for believing that I could make it, my sister, Dani, for sharing in the doctoral experience, and my nieces for their constant love and support. I especially thank my brother-in-law, Steven E. Loftus, for his contributions to all things technical at any hour, including his assistance in developing an online platform that I had dreamed up. Thanks also to my rock, Justin, for being there for me through each triumph and struggle. Thanks, too, to Kayla Tipsword, whose love and support were invaluable in all of my pursuits. Many thanks to all of my friends and colleagues who encouraged, inspired, and supported me through each phase of this experience especially: Ashley, Brittany, David, Emily, Erin, David, Huan, Jason, Kelly, Lyndsie, Megan, Ming, Olivia, Wei-Chieh, and Yanjie.

-Haley

Washington University in St. Louis

December 2018

For my parents, Bill & Barb, who showed me that anything is possible.

ABSTRACT OF THE DISSERTATION

Intersecting Variables of Second and Foreign Language Reading:

Self-Assessment and Comprehension with Adolescents and Adults Across Languages

by

Haley Dolosic

Doctor of Philosophy in Education

Washington University in St. Louis, 2018

Professor Cindy Brantmeier, Chair

Today, more individuals are seeking to learn a second language (L2) or foreign language (FL) than ever before. These individuals are language learners who intend to use their growing L2 knowledge to traverse linguistic and national boundaries in order to achieve their personal and professional goals (Callahan & Gándara, 2014). To realize these goals in the digital era, L2 learners must be more prepared than ever to use their knowledge with written media, making L2 reading a central skill to successful use of their L2 knowledge. Yet, when language learners attempt to use their L2 skills in their daily lives, they are often unable to diagnose their own strengths and weaknesses, making the task of comprehending a complex text challenging. In order to develop learner autonomy where learners are capable of overcoming such challenges, the learners themselves must be involved in their own assessment (Little, 2009). Therefore, with these two intersecting skills of L2/FL self-assessment and reading, at present there is a need to expand examinations to investigate the variables that shape these constructs in concert across varied contexts of language learning. The following dissertation contributes to this understanding

by (1) examining self-assessment through innovative methodological approaches and pedagogical interventions and (2) analyzing the intersecting and interacting variables that impact L2 reading comprehension in varied contexts of language learning with adolescents and adults. Specifically, these studies investigate the learners' abilities to accurately self-assess their L2 reading capabilities through three separate studies of L2 reading and self-assessment.

Building from prior examinations of these constructs, this dissertation uses quantitative methodology to investigate the connections among topic familiarity, assessment methods, text formats, self-assessment, and L2 reading comprehension (Brantmeier, 2006b; Brantmeier et al., 2012; Brantmeier, Stube, & Yu, 2014; Lee, 1996; Pulido, 2007; Wolf, 1993). In order to examine test methods, L2 reading comprehension is measured through multiple response paradigms throughout the dissertation, providing a robust understanding of L2 self-assessment as it is associated with different assessments of L2 reading comprehension (Alderson, 2000). Traditional and advanced statistical procedures are employed to robustly answer the questions presented in these studies (Plonsky, 2015).

Findings indicate that language learners (1) have variable capabilities in L2 reading comprehension, even when they are placed at the same level of instruction. Further, learners (2) may be able to accurately self-assess with and without training, particularly when afforded criterion-referenced self-assessment items and when such self-assessments are aligned with learners' prior test-taking experiences. Finally, these learners also demonstrated that (3) learners vary in their abilities to self-assess with accuracy, and individual differences in this accuracy can be accounted for with proficiency level. Implications for instruction and future research are discussed.

Chapter 1: Introduction

As more individuals around the world engage in language learning than ever before, there is an increasing interest in supporting these learners as individuals through learner-centered approaches (Butler, 2018; Yamagata, 2018). Further, there is an expressed need for these learners to develop twenty-first century competencies in their foreign language (FL) or second language (L2), including the increasingly crucial capacity to read L2 digital and print texts written for native speakers (Grabe, 2001; MLA Report, 2007). In their report, the MLA Ad-Hoc Committee on Foreign Languages called for restructuring modern language programs to meet the needs of these twenty-first century learners (MLA Report, 2007). Yet, at present, there is a need to expand current understandings of (a) the processes and products of skilled L2 reading (Bernhardt, 2011; Brantmeier, 2004) and (b) how learners may act as agents of their own learning when they are familiar with their own unique strengths and weaknesses (Brantmeier, Vanderplank, & Strube, 2012; Little, 2009; Rivers, 2001; Ziegler, 2014). Therefore, building from the theory and research that shape our current understandings of self-assessment and L2 reading, the following dissertation (1) examines self-assessment through three distinct methodological approaches and in three contexts of language learning where learners study English, Spanish, or French, and (2) analyzes the many variables that impact L2 self-assessment and reading comprehension performance at varied stages of acquisition. Specifically, the studies presented here were designed to better understand how L2 learners self-assess their L2 reading comprehension, and when and where learners are able to accurately diagnose their own strengths and weaknesses within the complex task of discourse-level L2 reading comprehension.

The following pages first present a brief review of research on the topics of reading comprehension and self-assessment as a foundation for the studies presented here. Then, each language learning context is presented in detail as is the research methodology applied in each study. Following this introduction, three separate studies conducted in varied language learning contexts are presented. In the final chapter, the findings of these studies are united to examine the overall impact of this dissertation and highlight future research that will expand upon these results.

1.1 Review of Foundational Theory and Literature

1.1.1 Reading Comprehension

1.1.1.1 First Language (L1) Reading Comprehension. Ever since the "complex, incremental ability" of reading was first examined many decades ago, researchers have sought to establish theoretical and empirically-derived models that account for the processes and products of reading (Grabe, 2009, p. 17). One of the most widely accepted models of reading was first theorized by van Dijk and Kintsch (1983) and is referred to as the Construction Integration Model (Kintsch, 1988). This model emphasizes three levels of comprehension: the surface model, the textbase model, and the situation model (Gernsbacher & Kaschak, 2013). Within this framework, the "surface model" is the literal wording of the text itself, and the "textbase model" is made of the ideas presented explicitly with the text (van Dijk & Kintsch, 1983; Kintsch, 1988). The commonly-referenced "situation model" represents the moment when learners are able to make inferences and integrate the meaning of the text into their knowledge base (Kintsch, 1988). As a result, full comprehension of a text is most often associated with this third level of comprehension, the situation model (Gernsbacher & Kaschak, 2013; Grabe, 2009). Uniting the lower-level processes of text decoding and sentence parsing with higher cognitive functions of

inferencing and organizing, this model provides an intuitive means to understand the processes and products of reading, particularly when individuals are reading to learn in academic settings (Grabe, 2009; Kintsch, 1998).

1.1.1.2 Second Language (L2) Reading Comprehension. **M**any theorists initially described the comprehension processes of L2 reading as a direct transfer of L1 reading skills into the L2 reading process, including Goodman (1967) and Cummins (1981). First and second language reading are understood to differ linguistically, developmentally, and socioculturally (Grabe, 2009). In fact, counter to these initial models depicting L2 reading comprehension as nearly identical to L1 reading, research indicates key differences shape approaches to texts that are in non-native languages, particularly with adolescents and adults learning to read in a L2 in academic contexts (Grabe, 2009; Koda, 2007). For example, when a child is learning to read, they are likely to receive constant reinforcement through the world in which they live, seeing written signs and hearing information read aloud around them frequently (Grabe, 2009). Yet, adolescent and adult students in L2 learning environments are less likely to have this rich reinforcing experience, resulting in a different set of needs to acquire reading comprehension capabilities (Grabe, 2009). Further, when learning to read in a second language, individuals may also have limited access to the culture that is embedded within the texts they are reading whereas L1 readers are more likely to be surrounded by the culture (Grabe, 2009). As a result, texts with embedded cultural experiences may present differing challenges for the L2 reader (Johnson, 1981; Steffensen, Joag-Dev, & Anderson, 1979). Such developmental and sociocultural factors of L1 and L2 reading experiences may cause differences to emerge in the processes and products of L1 and L2 reading comprehension.

In addition, learners who are gaining abilities to read in a L2 as adolescents and adults are simultaneously learning to speak, listen, and write, often with great variability in their successes (Grabe, 2009). As a result, these learners will be developing their reading comprehension competencies alongside broader linguistic knowledge in the L2 (Grabe, 2009; Koda, 2005). Actual processing of the language may be different; for example, with native French speakers who self-reported that they were nearly-native in English, English reading was 30% slower than French reading (Favreau & Segalowitz, 1982). Such differences are also present for individuals who have greater distance between their first and second languages. For example, with native speakers of Chinese and Korean reading in English, Koda (2005) found that, when learners seek to read in a L2, there are cognitive consequences of having learned to read with alphabetic or non-alphabetic writing systems.

In addition to these linguistic, sociocultural, and cognitive differences in L2 reading, within his or her mind, a L2 reader is also "working with the resources of two languages" (Grabe, 2009, p. 131). Therefore, L1 reading and L2 reading development are frequently interactive and mutually supportive because they are said to share some underlying cognitive domains (Bernhardt, 2005; Genesee, Geva, Dressler, & Kamil, 2006; Grabe, 2009). For example, L2 reading comprehension is like L1 reading comprehension in that it is shaped by lower-level and upper-level processes working cohesively, involving both decoding the graphical representation of a word and assessing its meaning within a larger discourse structure in order to ultimately understand the message encoded in a text (Grabe, 2009; Koda, 2005). As Koda (2005) explains, "successful comprehension emerges from the integrative interaction of derived text information and preexisting reader knowledge" (p. 4). Grabe (2009) adds that the reading process is shaped by its interactivity, flexibility, efficiency, and it nature as a purposeful,

evaluative, strategic, and linguistic process of comprehension and learning. Furthermore, the application of this shared cognitive system is nuanced with learners reportedly using greater metalinguistic knowledge and metacognition in their L2 reading experiences (Nagy, 2007), indicating that learners will have more awareness of linguistic gaps in understanding for their L2 reading (Grabe, 2009). Bernhardt (2011) further expands this understanding with another key statement, suggesting that reading is not only a cognitive, but also a social process.

Going beyond functional literacy, reading is often where advanced language learners use texts to "synthesize, interpret, evaluate, and selectively use information" (Grabe, 2009, p. 5). Such tasks may be considered an academic literacy and are necessary for learners who seek post-secondary degrees or advanced business positions. For learners who have a purpose of using their reading abilities to cross linguistic and national borders for work or study, this academic literacy is growing in importance. Therefore, researchers have examined the variables that shape these L2 reading experiences in order to better understand them.

Bernhardt (1991) clearly delineated two approaches to examining the L2 reading experience: the text and the reader. In this definition, she honored the complexities that arise both within the text as it is situated within a broader literary tradition as well as the many variables that shape each reader's experience of the text (Bernhardt, 1991). More recently, Bernhardt has expanded these theories with her Compensatory Model of Second Language Reading. Through this model, Bernhardt (2011) asserts that there is a flexible, yet intertwined system of variables that predict a learner's ability to successfully comprehend a text. Two key predictors are L1 literacy and L2 knowledge, in terms of L2 grammar and vocabulary knowledge. In her synthesis of empirical studies, Bernhardt (2011) indicates that these variables explain 50% of the variance in performance on reading tasks, with L1 literacy skills explaining

20% of variance while L2 knowledge explains an additional 30%. Together these synthesized findings suggest that the L1 literacy skills and L2 knowledge that an individual brings to a text greatly shape his or her ability to understand it (Bernhardt, 2011).

Koda (2005) adds that L1 literacy skills may have varying impacts on the ability of an individual to comprehend the text depending upon the orthographical distance among the L1 and L2. For example, learners who are native readers of Mandarin Chinese, a logographic language, may face greater challenges in word recognition in English than a native reader of Spanish, an alphabetic language (Koda, 2005). As word recognition is an essential component of accurate L2 reading (Koda, 2005; Grabe, 2009), orthographic distance can have impacts beyond word recognition, affecting larger reading comprehension development. Further, such linguistic distance is not necessarily associated with larger L2 knowledge because the process of recognizing this unfamiliar orthography relies on practice with the system, including obvious and subtle differences in the ways in which meaning are represented such as morphological markers (Grabe, 2009; Perfetti, 2007). Yet, this L1 to L2 transfer is not necessarily automatic for learners with closer orthographic distance (Koda, 2005), meaning this area of research remains one of keen interest to gain an understanding of how learners of different linguistic backgrounds are able to become successful readers in any given L2 across stages of acquisition. While this dissertation does not specifically examine this phenomenon, the reading comprehension presented in these studies does represent different linguistic distances, with Chinese students learning English covering a broader orthographic distance than those studying Spanish and French.

Beyond these two central factors of L1 skills and L2 knowledge, Bernhardt's model (2011) highlights the remaining 50% of variance that has not yet been reliably explained,

encouraging researchers to examine other variables that may reliably shape L2 reading comprehension to build upon the current explained estimate. As researchers expand upon this model, Bernhardt (2011) contends that these variables must be examined comprehensively due to their interconnectedness, measuring many features and examining them with methodological and statistical sophistication to expand current understandings of L2 reading comprehension. Further, Brantmeier (2004) calls for increased inquiry into the factors which may explain the remaining 50% of the variance, focusing on variables which could exert large influence. Together, this theory and call provide the basis for the broad, unifying inquiry in the following dissertation.

To date more than one hundred separate articles have examined a variety of characteristics, abilities, and experiences that may shape learners' L2 reading experiences including L2 decoding, L2 vocabulary knowledge, L2 grammatical knowledge, L1 literacy, phonological awareness, orthographic knowledge, morphological awareness, working memory, and metacognitive capabilities (Jeon & Yamashita, 2014). In a meta-analysis of the common factors associated with L2 reading using fifty-nine eligible studies, Jeon and Yamashita (2014) provide further empirical evidence for many of the most prominent features of the Compensatory Model (Bernhardt, 2011), indicating that L2 vocabulary and grammar knowledge are highly associated with L2 reading, with average correlations across independent studies of $r = 0.79$ and $r = 0.85$ respectively. Further, for L1 reading comprehension and L2 reading comprehension average correlations across independent studies was $r = 0.50$ (Jeon & Yamashita, 2014). Other key correlates which provided significant average correlations across independent studies include factors that belong in the remaining 50% of variance unexplained, such as age of learner (Bernhart, 2011; Jeon & Yamashita, 2014).

There are many other learner features that shape L2 reading that could also make up the fifty percent of unexplained variance, including background knowledge or topic familiarity. L1 and L2 reading comprehension research has reliably found that topic familiarity does relate to individuals' reading comprehension (Bernhardt, 2011; Koda, 2005). For example, Brantmeier (2005b) examined the role of analogies to facilitate comprehension of L2 reading with Costa Rican and American students, studying English and Spanish respectively, yet topic familiarity emerged as a greater predictor of reading comprehension success, with analogies failing to help students overcome a lack of familiarity with a text. Recently, with a high knowledge group of Chinese EFL university students, as determined by major, Horiba and Fukaya (2015) built on these findings, demonstrating that learners with higher knowledge produced qualitatively different recalls than students who were more unfamiliar with the text topics. Yet, there is an essential need to examine the topic familiarity honoring its nuance because it is "related to far more complex cognitive processes," involving mental representations, the structure and nature of the text itself, and how the reader interacts with the text to incorporate it into their understandings (Nassaji, 2002, p. 93). Therefore, the value and role of topic familiarity across language contexts with adults and adolescents merits further investigation.

Beyond these learner features, text features have also demonstrated associations with learners' abilities to comprehend the meaning of a text. Specifically, the structure of a text is impactful because each text relies on specific discourse structures that are culturally and contextually dependent (Koda, 2005). For example, narrative texts that tell stories have been found to be easier to recall (Freedle & Halle, 1979). Yet, our current understandings could be shaped by the contexts where investigations have occurred, as discourse structures are often cultural in nature (Kaplan, 1988; Koda, 2005). In the end, these learner and text features must be

considered together and as contextualized in broader society. Rather than thinking of the features of a text and individual characteristics of learners as separate features, these bidirectional forces must be examined in concert to provide a comprehensive understanding of L2 reading processes and products. Building from the strong foundation of research spanning these many intersecting variables, the following dissertation seeks to examine the predictors of L2 reading in concert in order to expand current understandings of the dynamic processes of L2 reading comprehension.

1.1.2 Second Language Reading Assessment

There are many techniques available to assess reading in a L2 with adolescents and adults. Many teachers and researchers alike have sought to find the best practice for their particular population in order to understand learners' skills, strategies, and development when reading in a L2 for students are already literate in their L1. Across contexts, these tests of L2 reading are often built around a text and a task, with learners first being asked to read a text and then complete a task to demonstrate their comprehension. These tasks can include answering questions by writing a response, selecting a choice, matching words and phrases, filling in a number of blanks, or even recalling everything that they can about a given text (Alderson, 2000). To decide which of these many tasks to use, Alderson (2000) urges assessment creators to consider carefully which abilities are tied to successful completion of their tasks because the assessment will measure best the abilities that are key to successful completion of the task.

Further, learners' demonstration of their understanding of a text varies according to the task used to measure comprehension (Lee, 1987) and language used on that task (Brantmeier, 2006a). Studies have compared a variety of testing tasks, finding that interpretations of students' abilities to read could be explained by the test type (Alderson, Bachman, Perkins, & Cohen, 1991; Alderson, Clapham, & Wall, 1995; Riley & Lee, 1996; Brantmeier, 2005a; Carrell, 1984a,

1984b, 1985; Shohamy, 1984). For example, with beginning French students, Riley and Lee (1996) examined student responses to a free recall or a summary task after reading a short passage. Findings indicate that learners responded differently depending upon the task they were asked to complete with those completing a free recall providing more details whereas those who summarized provided more main ideas (Riley & Lee, 1996). As a result, Alderson (2000) asserts that reading assessment should always be made up of "multiple methods and techniques" in order to gain a comprehensive picture of students' L2 reading abilities (p. 270).

The following dissertation uses multiple assessments to examine these phenomena further, particularly as they interact with self-assessment across contexts, providing an opportunity to examine learners diagnosing their own strengths and weaknesses of their L2 reading comprehension across testing methodologies. Specifically, a variety of tasks were selected to offer multiple means of understanding performance, offering a broader understanding of these learners' capabilities than any one task alone (Grabe, 2017; Koda & Yamashita, 2018; Wolf, 1993). Given that discourse-level L2 reading comprehension is the focus of this dissertation, three tasks (written free recall, sentence completion, & multiple-choice) were used to provide crucial understanding of learners' capabilities to comprehend the main ideas and key details presented in the given text. As such the tasks were aligned to capture these key facets of reading comprehension: "main-idea comprehension" and "recall of relevant details" (Alderson, 2000; Grabe, 2009, p. 357; Koda, 2005; Koda & Yamashita, 2018).

Therefore, multiple approaches that centered on these components of reading comprehension were included. More precisely, responding to Bernhardt's (1991) call that each task be both appropriate and meaningful for the examination of reading comprehension, written free recall, sentence completion, and multiple-choice were selected as these tasks added valuable

information about learners' understandings of these key constructs within each text through both correct and incorrect responses (Wolf, 1993; Brantmeier, 2006b). Written free recall, which is an assessment type wherein learners are asked to recall what they have just completed reading, offers an approach that is rarely used in classroom assessment and larger standardized testing paradigms (Alderson, 2000; Brantmeier, 2005b; 2006b; Grabe, 2009). However, this form of assessment does allow learners to indicate their understanding with little interruption or interference from the researcher or instructor (Bernhardt, 1991; 2011). This "extended production response type" (Bachman & Palmer, 1996) offers a unique and nuanced picture of what the learner has understood from the text that they have just completed reading. Specifically, this assessment examines the volume of the text that the learner is able to accurately represent, providing a clear picture of what learners were able to indicate they had understood. Within this dissertation, learners' units recalled are considered equal, without the researcher and test verifiers imposing a hierarchy. As a result, no main idea or key detail is privileged above other facets of the reading. Therefore, in order to further understand learners' comprehension of specific ideas presented in the text, more traditional measures, such as sentence completion and multiple-choice test types, are also included.

Multiple-choice is the most common and most popular assessment of reading comprehension (Alderson, 2000). Multiple-choice allows the test constructor, teacher and researcher alike, to focus the attention of learners on key ideas and possible misunderstandings within a text (Alderson, 2000). In addition, it is exceptionally easy to score, and answers are either correct or incorrect with little ambiguity (Alderson, 2000; Bernhardt, 2011). However, multiple-choice test types are fallible to students who are trained to use logic to eliminate answers and make educated guesses about what the correct answer may be (Alderson, 2000;

Brantmeier, 2005b). In order to guide learners but avoid such cues to support correct answers, sentence completion is another typical measure (Alderson, 2000; Brantmeier, 2005b; 2006b). This "limited production response type" (Bachman & Palmer, 1996) avoids much of the guessing possibility found in multiple-choice, and, with well-written questions, offers a depiction of learners' understanding of main ideas and key details (Alderson, 2000). For the purposes of this dissertation, which focuses on learners' main idea and key detail comprehension, such tasks were indispensable. Further, in order to maintain external validity where the results of these studies may be the most useful to practitioners, it is vital that the tasks used within the studies represent the realities that learners face in their classrooms and beyond which often are comprised of multiple-choice and short-answer formats (Alderson, 2000; Grabe, 2009; 2017).

Therefore, test methodology that aligned with the constructs being measured and prior principled research was developed and piloted (Alderson, 2000). Within each study, individuals read L2 or FL passages for understanding and were able to read the given passage multiple times. Then, these learners were required to complete all tasks without returning to the passage in order to measure comprehension of the text without the interference of scaffolding that could come from returning to the text alongside the tasks (Wolf, 1993; Brantmeier, 2005b; 2006b). Further, the tasks were situated such that learners moved from less cued to more cued questions. In this way, learners completed tasks with less prompting information before moving to tasks that offered more scaffolding for understanding. This was done so that learners demonstrated an understanding with only the cues present on the given task (Wolf, 1993). The following dissertation, therefore, examines these test methodologies both when woven together and as independent approaches, offering implications for a broader reading comprehension capacity and the realities of test method effects on our understanding of learners' comprehension.

1.1.3 Self-Assessment

Within educational and cognitive psychology, self-assessment exists within larger frameworks of metacognition. This metacognition "involves the organization, use and monitoring of cognitive activity" (Brantmeier & Dragiyski, 2009, p. 48) and is the way in which individuals think about and process information (Leader, 2008). The overarching models of metacognition contain a variety of separate features that shape learning from the moment one considers learning a new concept to the moment the information is used through retrieval or application (Metcalfe, 1996). When students make "judgements of knowledge" and have "feelings of knowing" as they encounter information, they are making metacognitive judgements that will shape how they continue to apply their knowledge to the context (Bjork, Dunlosky, & Kornell, 2012). These judgements and feelings are understood to be inferential, cue-dependent judgements that are shaped by varied individual biases and context conditions (Bjork, Dunlosky, & Kornell, 2012; Koriat, 1997). These are not guesses about whether students think they have learned a concept or could memorize a list of words, but rather individuals' inferences based on external cues and individual beliefs and experiences (Bjork, Dunlosky, & Kornell, 2012). For example, one bias that may impact learners' beliefs about their learning is called "stability bias." In experiments examining stability bias, participants have reliably predicted that their memory of a list of words does not depend upon the number of times they are able to study the list (between 0 and 3 times) (Kornell & Bjork, 2009). Rather, they rate their ability to remember similarly regardless of the number of times they are able to study; in this way, these learners demonstrate overconfidence in their memories early on, with limited study, and under-confidence in their future abilities, following more study, while maintaining a stability to their overall judgements (Kornell & Bjork, 2009). In a review of studies of evaluative judgements such as these in more applied classroom settings, findings indicate that learners who are more advanced in a particular

field tend to be better able to self-assess, yet results also highlight the fallibility of individuals to make accurate inferential judgements of their learning and knowledge (Falchikov & Boud, 1989).

1.1.3.1 Self-Assessment of Foreign & Second Language Skills & Knowledge. Self-assessment has also been an area of research within language learning for decades. While self-assessment can be seen as a tool that outwardly demonstrates students' own beliefs and evaluations of their L2 abilities (Brantmeier, 2005; 2006) or a skill that can be learned and practiced (Sweet & Mack, 2017), for the purposes of this dissertation, self-assessment is defined as an individual's rating of his or her own abilities in the L2. Examinations within this dissertation focus on the association between the students own self-ratings and the students performed capabilities, referring to a higher association of these two constructs as "accuracy" in self-assessed ratings. These studies focus upon discovering when and where such individuals are accurately able to distinguish their own strengths and weaknesses. Such examinations of self-assessment (SA) have demonstrated a variety of SA correlations, indicating "that there is a considerable variation in the ability learners show in accurately estimating their own second language skills" (Ross, 1998, p. 5). Therefore, there is a great need to understand the causes of this variability in order to gain a better understanding of SA to better equip learners to diagnose their own strengths and weaknesses.

Since Oskarsson (1978) first examined SA, finding these instruments to be reliable metrics of student performance, researchers have examined the statistical reliability and validity of this construct. Bachman and Palmer (1989) found that learners were able provide SA ratings that aligned with their performance on a communicative activity, and LeBlanc and Painchaud (1985) discovered that a highly contextualized questionnaire was also closely associated with

learner performance. Examining the construct as it related to standardized metrics, Krausert (1991) further established that L2 SA was a reliable and valid construct by demonstrating that self-evaluations did align with students' performance and that the self-assessment instrument itself withstood tests of statistical reliability. Beyond the reliability and validity of this metric as a measure of L2 knowledge and skill, Blanche and Merino (1989) encouraged the use of SA within instructed language learning settings in order to build learner awareness and autonomy in the classroom.

Based on this pioneering research, recent investigations have continued to examine (1) the construct reliability and validity of SA in differing contexts and (2) the usefulness of SA within classroom spaces. In examining reliability and validity through a meta-analysis of prior studies of self-assessment, Ross (1998) found that when questionnaires were contextualized, providing functional skill descriptions, learners were more able to accurately self-assess across reading, writing, listening, and speaking. These findings also supported a more consistent relationship among SA and performance of L2 reading than other skills (Ross, 1998). Further examining the contexts wherein SA might be a tool that replaces or expands upon other forms of assessment, Brantmeier (2005a) sought to discover whether descriptive self-assessment could add precision to placement testing at the university. However, results in two studies indicated that SA with these descriptive items did not accurately capture students' reading performance across measures (Brantmeier, 2005a; 2006b). Little (2005) took another approach, examining L2 performance in reference to the Central European Framework (CEFR) and arguing that SA should align with concrete skills within this framework. Building on the work of the DIALANG project (https://dialangweb.lancaster.ac.uk/) and aligned with the CEFR, Brantmeier and Vanderplank (2008) utilized a criterion-referenced SA instrument that would carefully situate a

learner in a language use context for SA purposes. With university students, these criterion-referenced SA measures proved accurate, meaning that students were able to accurately self-assess (Brantmeier & Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012). Through these investigations, it became apparent that accuracy of SA relied on the situational nature of self-assessment items that students could use to self-assess (Brantmeier & Vanderplank, 2008; Brantmeier, et al., 2012).

Beyond the questionnaire items themselves, conditions for learning and self-assessing appear to shape SA abilities. For example, students who are able to experience recent linguistic performance tend to succeed at SA (Butler & Lee, 2006; Dolosic, Brantmeier, Strube, & Hoegrebe, 2016). Likewise, students with less experience using their language knowledge in real-world contexts appeared to have difficulty accurately depicting their linguistic abilities, even when provided with contextualized criterion-referenced SA items (Schultz, 2017; Suzuki, 2015). Therefore, such language use experiences may be associated with individual variation present in self-assessment accuracy (Ross, 1998).

To date, there is an evident preference for the Pearson correlation in self-assessment literature, with the vast majority of L2 self-assessment research examining the construct with a group rather than as individuals. Yet, examining self-assessment as a within-person construct, through idiographic approaches, could allow for investigation of sources of systematic variability in individual variations, providing insight into self-assessment successes and challenges (Brantmeier, et al., 2012; Ross, 1998). In order to understand and account for systematic predictors in individual variations, therefore, it is vital that researchers adapt new methodologies and go beyond the Pearson correlation when examining self-assessment.

As researchers experiment with learners' abilities to self-assess in diverse contexts with varied questionnaires and methodologies, others have sought to understand the usefulness of self-assessment activities within instructed settings. Exploring the benefits of self-assessment in small learning groups, without comparison or control groups, self-assessment was found to encourage students to take ownership for their own learning as learners noted their strengths and weaknesses and could incorporate them into their study of the language, setting goals (de Saint-Leger, 2009; Duque Mican & Cuesta Medina, 2017; Poehner, 2012). Yet, not all learners dutifully worked toward and achieved their goals (de Saint-Leger, 2009). As researchers sought to examine the possible benefits of introducing self-assessment into the classroom, quasi-experimental studies were used to examine the effectiveness of self-assessment. Although one study found no difference between learners who had and had not experienced self-assessment enriched activities (Jafarpur & Yamini, 1995), in other studies with university students in varied settings around the globe, self-assessment training consistently improved learners' abilities to self-assess and the final performance of the trained as compared to the untrained learners (Naeini, 2011; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Shahrakipour, 2012; Yoon & Lee, 2013). For example, Malzloomi and Khabiri (2016) employed a quasi-experimental study, continuing more traditional methods in one classroom while implementing a dynamic self-assessment paradigm with their English writing for half of a semester. Results indicated that those who were in the experimental group were able to (a) self-assess with accuracy and (b) generate papers that merited higher scores than the control group of more traditionally taught peers (Malzloomi & Khabiri, 2016). These results indicate the possibilities of self-assessment training when it is used to help students. Together, the theory and research presented here have

laid the foundation for our current understandings of L2 self-assessment and reading comprehension to further future understandings.

1.2 Contexts of Language Learning

Although there are many shared principles and practices of language learning in instructed settings globally, the context in which a language is learned can shape the learners' goals, motivations, expectations, and experiences in relation to the language as these experiences are nested within individuals' lives in varied countries and schooling contexts (Lantolf, Thorne, & Poehner, 2015). Therefore, this dissertation, which examines three separate language learning contexts, has selected the students and contexts carefully and interpreted findings with regard to each sample's ability to represent a larger population of language learners. Viewed together, these separate samples provide keen insight into L2 self-assessment and reading comprehension.

1.2.1 Chinese EFL University Students

The first study within this dissertation was conducted with 65 Chinese EFL university students who were enrolled in an English for Academic Purposes (EAP) course at a university in Mainland China. These students were a convenience sample of the larger EAP program. The study conducted with these students was intended to examine the intersecting features of L2 self-assessment and L2 reading comprehension comprehensively, uniting often-separated features of these complex processes within an under-researched context (Koda, 2005; Wu, 2016). Examining the products of these processes within China was crucial to understanding these students because the educational paradigms of Chinese culture are also present in the language learning environments, possibly shaping their FL self-assessment and reading comprehension in unique ways (Paris-Kidd & Barnett, 2011). For example, within China students and professors alike indicate that classrooms are a space for a transfer of knowledge from an instructor to the

students (Maoying & Aiwu, 2011), indicating an approach to language teaching that could differ from many popular methods practiced in other nations. Further, these classrooms also have a focus on grammaticality in language, spending considerable time on grammar practice even in classrooms that report task-based pedagogy (Zheng & Borge, 2013). Therefore, this context may shape how learners conceive of their language skills, affecting both SA and reading comprehension in their L2, providing insight into how, when, and where these students are able to accurately depict their own strengths and weaknesses.

1.2.2 University Students Studying Spanish in the US

The second study within this dissertation examines the implementation of an online self-assessment training program with 136 university students studying Spanish at an elite research-focused institution in the Midwestern US. These students were all enrolled in the advanced Spanish language course and were included in this study as a convenience sample. Within this context of language learning, many students will take courses in Spanish literature to complete their programs of study, focusing on the intricacies of literary traditions within such courses (Goldberg, Looney, & Lusin, 2015). However, many course instructors lament that plot deconstruction takes up a considerable amount of time in these literature courses, despite learners' advanced language proficiency. Within this particular language learning context, students took a course that covered topics in advanced uses of Spanish grammar and vocabulary, in conversations and compositions, developing their language skills through interactive activities and homework assignments. With these students in mind, a reflective self-assessment training was developed as a course assignment, providing students with an opportunity to self-assess their L2 reading comprehension across the semester. These students were compared to a control group who did not complete training. In this way, this dissertation examines a typical course in the

sequence of a Spanish major in a new way, offering findings that provide key insights into how learners may self-assess and how this knowledge may shape learners' abilities to read in Spanish.

1.2.3 Specialized French Immersion Program for High School Students

The third study within this dissertation examines the development of L2 self-assessment and L2 reading comprehension within a specialized short-term French immersion program designed for high school students. This program is designed to provide students with the equivalence of a year-long high school French course within four weeks. Through its intensive design, students take traditional, language-focused courses and practice their language skills as they complete typical summer camp activities, such as singing, canoeing, and crafting, entirely in French. While such programs remain a popular means to learn a language, little research has been conducted with students in these immersive environments, particularly with L2 self-assessment and L2 literacy. Yet, findings which examine L2 self-assessment and L2 oral production have offered promising findings of ability growth both for self-assessment and oral production (Dolosic, Brantmeier, Strube, & Hogrebe, 2016). Therefore, this context offers a truly enriching perspective on how, when, and where learners may be able to both develop and self-assess their L2 abilities.

1.3 Overarching Research Design and Statistical Analyses

Building from prior examinations of these constructs, the following dissertation uses quantitative methodology to unite and to investigate the connections between self-assessment and L2 reading comprehension across varied contexts of language learning (Brantmeier, 2006a; Brantmeier, 2006b; Brantmeier et al., 2012; Brantmeier & Yu, 2014; Lee, 1996; Pulido, 2007; Wolf, 1993). Through the use of robust data collected through principled practices, quality basic and advanced statistical procedures were implemented to answer the novel research questions of

this dissertation (Plonsky, 2015). All tests were conducted using *R: A language and environment for statistical computing* (2013).

In the first study of this dissertation, *An Examination of Self-Assessment and Interconnected Facets of Second Language Reading with Advanced Chinese Undergraduates Studying English*, 65 Chinese undergraduates studying English participated in an exploratory, cross-sectional examination of their second language self-assessment and reading comprehension. Coding in pausal units (Brantmeier, Strube, & Yu, 2014), reading was examined through a multifaceted assessment to capture a comprehensive picture of students' capabilities. This assessment includes free recall, sentence completion, and multiple choice reading comprehension measures (Alderson, 2000; Brantmeier, 2006b; Wolf, 1993). A 2 x 3 within-subjects ANOVA was conducted to investigate direct and indirect effects of both text and test type, as an interaction between these effects was possible but had not yet been explored. Alongside simple correlational analyses of topic familiarity and self-assessment as they related to L2 reading comprehension, the findings revealed that although these learners' topic familiarity does not relate to their reading comprehension, both text and test type do impact learner performance, and there is a statistically significant interaction among text and test type. Further, although criterion-referenced self-assessment seems to demonstrate an association between learners' self-rating and L2 reading performance, when examined by test and text type further complexities arise.

In the second study, *An Investigation of Interactive Online Self-Assessment Training and Advanced Second Language Reading of Spanish at the University*, two groups of university students were compared in a quasi-experimental design (Hatch & Lazaraton, 1991). With multiple sections of the same Advanced Spanish course, individuals completed assigned, online

self-assessment modules designed to train the students to accurately self-assess. More precisely, students taking this Advanced Spanish course in the Spring of 2017 completed the course with no intervention while students in the Spring of 2018 completed the course with the modules, training at least three times in the first twelve weeks of the semester. Both groups completed identical measures in Week 13. Following assumption tests, descriptive statistics, t-tests, and ANOVA were applied to examine the data and compare the groups to attend to the central research questions about differences in performance (Plonsky, 2015). In addition, regression-based statistics were used to examine the associations among self-assessment and reading comprehension performance. A Fisher's Z test of difference of correlations provided the opportunity to examine the differences among the correlations (Meng, Rosenthal, & Rubin, 1992). Results demonstrate that the relationship among self-ratings and L2 reading comprehension performance was altered due to the experience of the training. More precisely, findings of these tests revealed that while self-ratings and reading comprehension were considered to be stable across these groups, with only one significant difference discovered in L2 reading scores, the correlations among self-ratings and performance were significantly larger for those who were trained than the correlations for those who had not received training.

The final study of this dissertation, *An Individualized Approach to Self-Assessment with Readers in the French Village*, pairs well-known statistical techniques with novel approaches to examining self-assessment in order to gain a better understanding of the causes for the individual variability among self-ratings and accuracy of self-assessment. In this study, high schoolers completed a program of French immersion for four weeks during a summer, experiencing a rich curriculum of reading, writing, listening, and speaking in French. Analyses within this study first examined the gains that learners made during their time in the intensive language learning

environment due to a dearth of prior results in such settings. Through paired t-tests, and bootstrapping checks to overcome assumption violations, statistically significant learning was revealed (Plonsky, 2015). Further examinations first applied traditional methods of self-assessment analysis, focusing on the Pearson correlation (Ross, 1998). While these results indicated that learners were able to self-assess with accuracy both before and after their time in the immersion setting on both language knowledge and reading comprehension in French, further analyses were needed to examine the individual variability in accuracy. Therefore, hierarchical linear modeling (HLM) frameworks were employed through a no-intercept, dummy coded model. Such modeling allowed for a within-subjects design comparing accuracy of self-assessment before and after the program as well as allowing for modeling that incorporated proficiency into accounting for the variability of L2 reading comprehension self-assessment accuracy. This analysis provided an opportunity to better understand students' variable abilities to self-assess (Sheskin, 2007) and offered insight into what appears to be an under-researched language learning context.

1.4 Summary of This Dissertation

Built on this foundation of theory and research, these methods were employed in the presented contexts of language learning. The details and contributions of these studies are summarized here in Table 1.1 and presented in detail in Chapters Two, Three, and Four on the following pages. Finally, results are discussed together and future directions for research are presented in Chapter 5.

Table 1.1 *Summary of Studies within this Dissertation*

Chapter	Participants	Research Questions	Conclusions
Chapter 2: An Examination of Self-Assessment and Interconnected Facets of Second Language Reading	65 English majors studying English for Academic Purposes at a Chinese University	1) What is the relationship among text type, test method, and L2 reading performance? 2) What is the relationship between topic familiarity and L2 reading performance? 3) How do these students self-assess their L2 reading abilities? Are they able to self-assess with accuracy across text types and test methods?	Self-assessment, text types, and test types need to be examined together, as findings indicate interactive relationships. Further, topic familiarity may have a more nuanced relationship with reading comprehension at advanced levels.
Chapter 3: An Investigation of Interactive Online Self-Assessment Training and Advanced Second Language Reading at the University	136 University students studying Spanish in Midwestern USA	1) What are differences in the L2 reading comprehension self-ratings with and without self-assessment training? 2) How do self-ratings align with performance? How does the relationship between self-assessment and performance differ with and without the self-assessment training? 3) Are there significant differences in reading comprehension performance between the groups who did or did not complete self-assessment training? 4) What goal-setting behavior did learners employ?	Students are able to develop capacities to accurately self-assess and set meaningful goals through an online self-assessment tool. Yet, one semester of intervention was not enough to change learners' reading comprehension performance.
Chapter 4: An Individualized Approach to Self-Assessment with Readers in the French Village	47 High school students studying French at immersive summer program	1) To what extent do learners demonstrate gains in L2 language knowledge and reading comprehension? 2) How do learners self-assess their abilities? 3) When examining the relationship between SA and performance, do individual differences emerge? Do student characteristics such as proficiency explain these individual differences?	Learners do demonstrate gains in this context, and they are able to accurately depict their own strengths and weaknesses both before and after, in general. However, there are individual differences in their accuracy of self-assessment.

Chapter 2: An Examination of Self-Assessment and Interconnected Facets of Second Language Reading

Reading is seen as a crucial skill for the citizens of modern societies to learn information and communicate (Grabe, 2009). Recently, these skills have grown in importance as individuals around the world increasingly reach across not only national but also linguistic borders for their own academic and economic pursuits, as they strive to attain more desirable positions in society both within their homelands and abroad. As English increasingly becomes a medium of instruction and business negotiation globally, access to such success is often driven through successful performance on high stakes tests of English which frequently focus on reading skills (Baker, 2015; Cheng, 2008). In China, the importance of successful English reading is intensified for English as a foreign language (EFL) students, as this growing population is developing its reading skills earlier and with greater focus than ever before (Pan, 2007; Wu, 2016). However, despite studies which have examined L2 reading comprehension across varied dimensions, second language (L2) reading comprehension is not yet fully understood as an integrated process (Bernhardt, 2011; Grabe, 2009; Koda, 2005). Specifically, Bernhardt (2011) highlights the need for comprehensive, empirical investigations of the "array of variables" that shape L2 reading comprehension, building from prior research to examine these interacting variables (p. 136).

Furthermore, when learners encounter tests of L2 reading, they may also not be aware of their own strengths and weaknesses (Brantmeier, 2006; Schultz, 2017). Such an awareness, or ability to self-assess, could prove central to learners' capabilities to become autonomous, life-long L2 readers (Little, 2009). Although prior research examines self-assessment as it relates to

general reading comprehension tests (Ross, 1998) and in reference to its ability to accurately place students (Brantmeier, 2005b; 2006b), few studies examine L2 reading self-assessment in relation to the variables that shape L2 reading comprehension. These comparisons could provide crucial insight on understanding L2 self-assessment and reading comprehension, offering implications for supporting learners in their quests for success both on tests of English and throughout their lives as L2 readers. Therefore, this study unites some of these often-separated features of self-assessment and L2 reading, seeking a comprehensive understanding of the interactive relationships among key variables such as text type, test method, topic familiarity, self-assessment, and L2 reading comprehension with Chinese EFL university students.

2.1 Literature Review

2.1.1 Second Language Reading

Although reading comprehension competence can be described in many ways, reading is generally understood to be the process by which one perceives, decodes, and seeks to gain meaning from text on a page or screen by combining the extracted information with the knowledge that the reader already possesses (Koda, 2005). Yet, Grabe (2009) highlights that reading is much more than the process of decoding and incorporating information into pre-existing knowledge systems. Rather, he contends that reading is an efficient, purposeful, interactive, strategic, and fundamentally linguistic process, calling attention to the multifaceted and multipurposed nature of reading (Grabe, 2009; Perfetti, Landi, & Oakhill, 2005). When this process takes place in a L2, it is further complicated by orthography, grammar, text conventions, and other knowledge that is key to processing and understanding text in the other language (Koda, 2005; 2007). Readers must overcome these complexities in order to develop a mental representation of the text, often referred to as a situation model (Kintsch & van Dijk, 1978).

Bernhardt (2011) examines this complex process in her synthesis of research on the topic of advanced L2 reading, developing upon her Compensatory Model of Advanced L2 Reading (Bernhardt, 1991; Bernhardt, 2011). Unifying research in the field of L2 reading, this Compensatory Model (Bernhardt, 2011) confirms that first language (L1) literacy can account for approximately 20% of variance in performance on reading assessment tasks, whereas L2 linguistic knowledge, in terms of grammar and vocabulary knowledge, explains an additional 30% with advanced L2 readers. This model also accounts for the fact that individuals may compensate for weakness in one category, such as L1 literacy, with strength from another, such as L2 vocabulary knowledge. Yet, there is a remaining 50% of variance in reading comprehension performance that has not yet been reliably explained (Bernhardt, 2011). Brantmeier (2004) contends that further investigations should examine factors that are found within this remaining variance, seeking factors which may be more influential in yielding better L2 reading comprehension. At present, test method, text features, and topic familiarity are recognized by many researchers for their nuanced effects on L2 reading (Brantmeier, 2005b; 2006b; Hammadou, 1991; Riley & Lee, 1996). However, further research uniting these features is necessary to determine their place within the Compensatory Model (Bernhardt, 2011). A great deal of the remaining variance could be explained with the relationships among these key factors that have been shown to shape L2 reading, expanding current theoretical understandings of L2 reading and providing insights that may be applied in classrooms.

2.1.1.1 Testing Methods. Researchers and educators alike use various types of tests and assessments to build an understanding of language learners' reading abilities. Within these tests students are often asked to read a passage appropriate to their stage of acquisition or course level. Then, they complete a task with the text that they have been provided. These tasks can be quite

different, asking the learner to select an answer, match two words or ideas, complete a passage by filling in blanks, or simply recall everything that they possibly can from the passage they have just read (Alderson, 2000). Researchers frequently use three different types of assessment tasks in a single study of reading comprehension: free recall, sentence completion, and multiple choice (Brantmeier, 2005b; Brantmeier, 2006b), and research has also examined differences by testing methodology (Brantmeier, 2006b; Shohamy, 1984). In their examinations, Bachman (1990) and Wolf (1993) divided these testing tasks, often referred to as test methods, into two categories where the learner is either selecting or constructing a response. Studies have compared findings and correlations across testing tasks, demonstrating differences among results based on type of test (Alderson, Bachman, Perkins, & Cohen, 1991; Alderson, Clapham, & Wall, 1995; Riley & Lee, 1996; Brantmeier, 2006b; Carrell, 1984a, 1984b, 1985; Shohamy, 1984). Results to date indicate that students consistently achieve different scores depending upon the type of testing task they encounter (Brantmeier, 2006b; Wolf, 1993).

When synthesizing research on the topic of second language reading assessment, Bernhardt (1991, 2011) supported the use of free recall to measure individuals' abilities to comprehend a passage, suggesting that assessment of learners' comprehension is best demonstrated when the assessor does not interfere with the frame of understanding. Yet, Alderson (2000) supports test methods that respond more specifically to the information that the test constructor is trying to gain. Further, Alderson (2000) recommends that a good assessment incorporates multiple methods of assessment, creating a comprehensive and statistically reliable picture of an individual's comprehension abilities. In addition, both Bernhardt (2011) and Alderson (2000) highlight the variety of factors outside of the type of assessment that could shape an individual's performance on a given method of reading assessment such as the learner's

proficiency or topic familiarity. Drawing from Bernhardt's Compensatory Model (2011), it is possible that these factors interact, shaping learners' abilities to successfully comprehend advanced L2 texts.

Lee (1987) suggested that learners' ability to communicate what they understand varies according to the task used to measure comprehension. While Wolf (1993) proposed that difficulty in completing a task relies only on the skills required to complete it, the novelty of the task may also need to be considered. For example, according to the theoretical framework Transfer Appropriate Processing (TAP) (Morris, Bransford, & Franks, 1977), students may be more able to succeed on a task that aligns closely with how they were instructed. Recently, in a first language study endeavoring to grow learner autonomy, Thomas and McDaniel (2007) found that where the learning task was consistently closely aligned with the assessment task, students consistently performed better, as TAP would predict. The TAP framework is also consistent with L2 frameworks of vocabulary acquisition such as the TOPRA (Type of Processing—Resource Allocation) model (Barcroft, 2002, 2003, 2004, 2013), which emphasizes that tasks must engage specific types processing in order for students to learn certain aspects of form and meaning mapping for novel L2 words. Therefore, prior results that demonstrate greater success when students select their responses could also be associated with instructional methods, such as test preparation, that emphasize selecting the correct response.

Although these examinations do provide an understanding of L2 reading assessment, there is a need for more empirical work examining varied types of assessment, especially as these test types may interact with other key variables in the reading process (Brantmeier, 2005a). Further, as tests of language proficiency have become embedded in the larger culture of testing and access to higher status in society within China (Cheng, 2008), it is vital that these test types

be understood within the often under-researched context of Chinese EFL learners in China (Wu, 2016). Chinese EFL university students may have different approaches to such text-based tasks, based on their language learning contexts and motivations that differ greatly from many more closely examined contexts of language learning (Li & Cutting, 2011). Therefore, there is a need for a study which examines test methods in concert with other variables of L2 reading, particularly within the context of Chinese EFL learners.

2.1.1.2 Text Types. Knowing and being able to predict the structure of a text can also be beneficial on a test of reading comprehension (Alderson, 2000) and has been seen as a possible variable to include as part of the Compensatory Model. Within the broader field of applied linguistics, the structure or organization of a text has been examined largely from three different vantage points: (1) the cultural conventions present or not present in a text, (2) the coherence of the passage, or (3) the narrative and expository spectrum of texts. In examining the role of text structure in terms of cultural coherence among 240 Taiwanese university students majoring in English, Chu, Swaffar, and Charney (2002) asked students to read and recall passages that were tailored either toward traditional Taiwanese writing conventions or more English writing conventions. Findings indicated that both experience with the language and reading a text with English conventions were predictive of performance (Chu et al., 2002). Greater experience with the English language led to better recall whereas a text that was tailored toward English or Western conventions yielded a less complete recall (Chu et al., 2002). When Riley (1993) examined text features in terms of coherence, findings indicated that with university students studying French that coherence breakdowns in a story were most impactful for intermediate level students. Riley (1993) argued that the students of lowest proficiency may not have enough processing capacity to comprehend this change of coherence when working to understand the

text in their second language. At the same time, he suggested, advanced students may have enough mental processing available to overcome such problems in coherence even after using their attentional resources on language processing, leaving intermediates as the most sensitive group to such coherence issues (Riley, 1993). Horiba (1996) found similarly that L2 speakers of Japanese and English were sensitive to the coherence of a text only when they were proficient in the L2, indicating that the factors of coherence are tied to key facets of the L2 reading experience such as L2 proficiency.

Yet, in general, when examining these textual factors researchers tend to discuss the differences in comprehension for narrative and expository passages (Koda, 2005). Typically, these are defined such that narratives tell stories with a causal chain of events whereas expository passages report facts and information, often in hierarchies (Koda, 2005). DuBravac and Dalle (2002) took up these differences, examining university students' abilities to comprehend varied texts in their L2, French. Findings indicated that students were better able to comprehend the narrative text, having poorer comprehension when reading an expository text. Donin, Graves, and Goyette (2004) similarly found that military officers studying French recalled more from the narrative than the expository texts. Further, with Japanese EFL university students, Yoshida (2012) found that more was recalled for narratives than expository texts, and that proficiency was likewise a key predictor of students' success. Despite the differing foci of these studies, together their findings indicate that the structure of a text shapes readers' comprehension. However, these differences may be affected by the reader's L2 proficiency or familiarity with the test method. Despite a need to better understand Chinese EFL learners' approaches to such varied texts both within and outside of reading comprehension tests, no prior study has examined

the role of text organization with advanced EFL university students in China, particularly as text features relate to other key features of second language reading.

2.1.1.3 Topic. In addition to test and text type, first and second language research have reliably found that a readers' familiarity with the topic of the passage that they are reading impacts their understanding of the text (Koda, 2005). For example, Barry and Lazarte (1998) found with 48 university students studying Spanish that participants who demonstrated "high knowledge" were able to generate more inferences, fewer incorrect inferences, recall more text information, and generate qualitatively better representations of the text than "low knowledge" participants on a three-passage free recall comprehension assessment. Further, Brantmeier (2005a) examined the role of analogies to facilitate comprehension in L2 reading with Costa Rican and American students studying English and Spanish respectively, yet topic familiarity variance came to the forefront as the greater predictor of students' successes, with analogies failing to help students who had little knowledge of the topics of the passage to comprehend the text. Uso-Juan (2006) found similar results with Spanish-speaking EFL students. Dividing students into high and low knowledge groups based upon their majors in the university setting, Uso-Juan (2006) found that background knowledge was integral to students' comprehension. More recently, Horiba and Fukaya (2015) also found that a high knowledge group of Chinese EFL university students, as determined by major, were more successful in comprehending reading passages with qualitatively different recalls being evident between the high and low knowledge groups. Together these findings clearly indicate that there are strong effects associated with the topic of the reading passage, particularly with high and low knowledge groups. Together these findings appear to support the claim that topic knowledge is central to comprehension.

However, as Nassaji (2002) cautions, background or topic knowledge is often found to be more complex than it is straightforward. Prior knowledge is not a simple mechanism that clearly and directly predicts success on reading comprehension. Rather, Nassaji (2002) contends, it is "related to far more complex cognitive processes," involving the readers' previous mental representation, the structure and nature of the text, and how the reader is able to incorporate it into their memory (p. 93). Such dynamic representations have been found in studies which examine proficiency and topic familiarity together. For example, Chen and Donin (1997) examined reading with graduate students who were L2 speakers of English, finding that, for these individuals, language proficiency and background knowledge could compensate for one another in order for learners to successfully comprehend the text. Therefore, these students could use language proficiency to overcome a lack of background knowledge (Chen & Donin, 1997). Further, few studies examine topic knowledge when it is less extreme, as the topics of standardized assessment might be. Research is needed to understand the exact role of topic familiarity in L2 reading, particularly in China where few studies to date have taken up this area of inquiry and many high-stakes English reading tests occur.

2.1.2 Second Language Self-Assessment

Within this study, self-assessment (SA) is defined as one's own evaluation of one's performance or capabilities. SA has been studied in the field of language learning since the late 1970s both to understand the mechanisms and instruments of self-assessment themselves and to assess their impact on students' learning. At the institutional level, Oskarsson (1978) examined correlations between written test scores and students' self-ratings, finding a positive relationship between university students' own views and their actual scores on language assessments. Seeking to better understand these self-evaluative techniques, LeBlanc and Painchaud (1985)

found that, with a highly-contextualized SA instrument, students were able to accurately represent their own L2 abilities through self-assessment, meaning that their self-ratings were closely associated with their actual performance. Bachman and Palmer (1989) continued this trend by examining the validity of SA as a reliable construct, finding that it was statistically reliable with their students. From these early stages, authors have argued for the use of self-assessment to develop learner autonomy (Blanche & Merino, 1989; Harris, 1997; McNamara & Deane, 1995).

As Little (2009) argued, students must be involved in "planning, monitoring, and evaluating" their language learning in order to become autonomous, life-long language learners (p. 224). Specifically, an accurate understanding of one's own strengths and weaknesses could provide language learners with opportunities to set realistic, challenging goals which are central to such life-long language learning (Sweet & Mack, 2017; Ziegler, 2014). Therefore, self-assessment may be an important indicator of students who are developing the skills they need to become these life-long language learners. For example, Sweet and Mack (2017) investigated the effects of reflective, collaborative speaking self-assessment tasks with 367 university students studying Spanish in the United States. For this program, learners were provided with speaking tasks to complete and reflect upon, both individually and with a peer. Findings demonstrated that these learners were able to develop an awareness of their capabilities, the process of learning language, and an ability to monitor their own progress (Sweet & Mack, 2017). Self-assessment itself may also become more accurate, with learners being better able to represent their own strengths and weaknesses following other types of training. For example, with 91 Vietnamese university students studying EFL, Nguyen and Gu (2013) provided half of the students with a metacognitive, reflective training for English writing. The students who received the training

gained significantly greater accuracy in their self-assessment than students who did not experience the training. Further, these students also produced English writing that was scored more highly (Nguyen & Gu, 2013). Similarly, regular self-assessment training has demonstrated improvement across the semester in terms of linguistic development for beginning and intermediate reading, writing, and listening (Mazloomi & Khabiri, 2016; Shahrakipour, 2012; Yoon & Lee, 2013). Such results may indicate that understanding when and where students are aware of their own strengths and weaknesses, both with and without specific self-assessment training, could be vital in developing courses and programs that support life-long language learning.

However, nuances of self-assessment have also been discovered through closer examination of L2 self-assessment as a measurement and placement tool. For example, in his meta-analysis of self-assessment, Ross (1998) found that SA is more accurate when the items used to self-assess are more functional and less abstract. Brantmeier (2005b), in seeking a fine-tuned placement tool, continued this line of inquiry, finding that with adults studying Spanish in the United States, descriptive SA did not correlate with multiple choice assessment, yet it did correlate with free recall measures of reading comprehension. In another study, with university students studying Spanish, Brantmeier (2006b) found that descriptive SA did not correlate with students' performance on a computer-based proficiency exam. Yet, when Brantmeier and Vanderplank (2008) introduced and utilized criterion-referenced items that aligned with DIALANG proficiency standards (https://dialangweb.lancaster.ac.uk/) and situated the learner within a specific context of language use, students were able to accurately self-assess their reading comprehension as measured on multiple choice items. These items were statements of actual concrete skills that the learner rated his or her ability to complete on a five-point Likert

scale. When these criterion-referenced items were expanded to include other skills, Brantmeier, Vanderplank, and Strube (2012) discovered that advanced language learners were able to represent their abilities by answering these items across listening, reading, and writing. Their self-evaluations were significantly correlated with their performance (Brantmeier, et al., 2012).

Yet, it appears that the format of the questions alone cannot account for students' responses. For example, with EFL learners, Butler and Lee (2006) found that elementary students who completed a series of language tasks before completing a self-assessment demonstrated associations between their self-ratings and performance. However, students who did not have this "on-task" condition did not demonstrate the same associations, meaning that having the recent experience may have led to different results (Butler & Lee, 2006). In another study, with Chinese learners of Japanese at a university, Suzuki (2015) found that experiential factors were heavily associated with students' ability to self-assess. Those with greater experiences in the L2 seemed to underrate their skills whereas those with little experience appeared to overrate their skills (Suzuki, 2015). In addition, Dolosic, Brantmeier, Strube, and Hogrebe (2016) found that adolescents studying French in an immersive setting were unable to self-assess their speaking abilities on a criterion-referenced instrument when they arrived, but they were able to accurately self-assess on a criterion-referenced measure after spending four weeks in the immersive environment, demonstrating the power of experience for successful self-assessment. Further, for adult students studying English in China who lacked these immersive experiences, criterion-referenced self-assessment scores did not share a relationship across performance scores (Schultz, 2017). Yet, Ding and Stapleton (2016) also found through qualitative inquiry that Chinese university students studying in an English medium university in Hong Kong were able to develop self-assessment abilities from exposure to an immersive

context of learning. These findings taken together indicate that cultural and educational experiences as well as contextualized items are key for successful SA. With the possible positive benefits of SA in terms of learner autonomy, it is vital to better understand self-assessment, particularly with this growing population of EFL students in China.

2.2 Research Questions

In the modern, globalized world it is vital that researchers continue to examine features and facets that make up L2 reading in order to gain a full understanding. Despite a need to examine the constructs of topic familiarity, text type, test method, self-assessment, and L2 reading performance comprehensively in a single study in order to gain a more complete understanding of the multivariate nature of reading, to date no study examines these constructs in a single investigation, particularly with the growing and developing population of Chinese EFL university students. Motivated by this need, the following study was guided by the following research questions:

1. What is the relationship among text type, test method, and L2 reading performance with Chinese EFL university students studying English in China?
2. With these students, what is the relationship between topic familiarity and L2 reading performance?
3. How do these students self-assess their L2 reading abilities? Are they able to accurately self-assess their L2 reading abilities as measured by an L2 reading task across varied text types and test methods?

2.3 Methods

2.2.1 Context & Procedure

This study was conducted at a medium-sized university in Northern China specializing in the training of teachers. All students (N= 77) were English majors, enrolled in an English for Academic Purposes (EAP) course during the third year of study at the university. Many students planned to become English teachers, continue for advanced degrees, or use their English language skills to succeed in business. Only 64 students out of 77 completed all measures successfully. As is typical for the context, 59 students self-identified as female; five identified as male. These students were brought together in a single lecture hall to complete all instruments at one time. All students came to the lecture hall and completed all measures within one paper packet with the researcher and course instructor present. Students progressed linearly, from start to finish, over the course of a two-hour session.

2.2.2 Instruments

During the two-hour session, students completed multiple measures, including a demographic questionnaire, criterion-referenced self-assessment questionnaire, and a reading comprehension assessment. The criterion-referenced self-assessment was tailored from prior research, having been validated with larger samples (Brantmeier, Vanderplank, & Strube, 2012). Developed from the DIALANG framework, this questionnaire was made up of criterion-referenced items that situated the learner in a specific language use context, asking learners to rate their ability to complete the task on a scale of one to five, with one meaning that the learner "Strongly disagreed" that they would be able to complete the task while five meant that the learner "Strongly agreed" that they would be able to complete the task (Brantmeier et al., 2012). The reading comprehension assessment was carefully constructed according to the guidelines

established by Wolf (1993) and Alderson (2000). It included three forms of assessment: Free Recall, Sentence Completion, and Multiple Choice. Four passages were carefully selected to be of equal length and difficulty, differing only in topic and text type (with two fitting the description of "narrative" and two fitting the description of "expository"). The four passages that were tested were qualitatively checked for agreement that they were narrative or expository in nature among two researchers and the author. Instructions for all tasks were presented bilingually (in Mandarin and English). Topic familiarity was measured through a five-point Likert scale question following each passage where the learner indicated their familiarity with the topic of the passage on a scale of one to five. These items were also presented bilingually (in Mandarin and English). Free recall was scored on pausal units. (See Brantmeier, Strube, & Yu, 2014 for full discussion.) Sentence completion and multiple choice were scored as correct or incorrect based on pre-determined answers drawn from the texts by the researcher, with one point given for each correct response. For further information about the number of items on each section and possible number of items for which an individual participant could have been given a higher score, see Table 2.2.

2.2.3 Analysis

All analyses were conducted with complete observations of all variables in the open-source statistical program, R version 3.3.3 (R Core Team, 2017), using additional packages including psych (Revelle, 2017), psychometric (Fletcher, 2010), and car (Fox & Weisberg, 2011). Figures were generated using GGPlot2 (Wickham, 2009). Although data were transformed to run analyses that corresponded to the research questions outlined here, final results are presented with un-transformed data because all results and interpretations were equal across Box-Cox transformed data and the original data (Box & Cox, 1964). Data were cleaned

by removing outliers and incomplete questionnaires. Removed outliers were determined to be significantly different from the rest of the sample when examined through univariate and multivariate techniques. Analyses comparing different measures were done with percentages in order to account for the number of possible points one could obtain on any given measure.

2.4 Results

Descriptive statistics were examined, and correlations, regression, and ANOVA were conducted to answer the research questions. Descriptive statistics are provided in Table 2.1 to demonstrate the means and variability of the data collected.

Table 2.1. *Descriptive Statistics of All Key Variables*

	Min.	Max.	Mean	Standard Deviation
Topic Familiarity	6	21	10.44	3.12
Self-Assessment	47	73	58.45	6.05
Free Recall	4	50	26.73	11.55
Sentence Completion	9	27	19.69	3.97
Multiple Choice	8	28	22.61	3.28
Narrative	12	56	34.88	9.78
Expository	17	47	34.16	7.23
Composite	33	98	69.03	15.97

1. *What is the relationship among text type, test method, and L2 reading performance with Chinese EFL university students studying English in China?*

Descriptive statistics indicate a great variability to students reading performance as measured through composite scores of all test types with a standard deviation of 15.97 (Figure 2.1). Scores

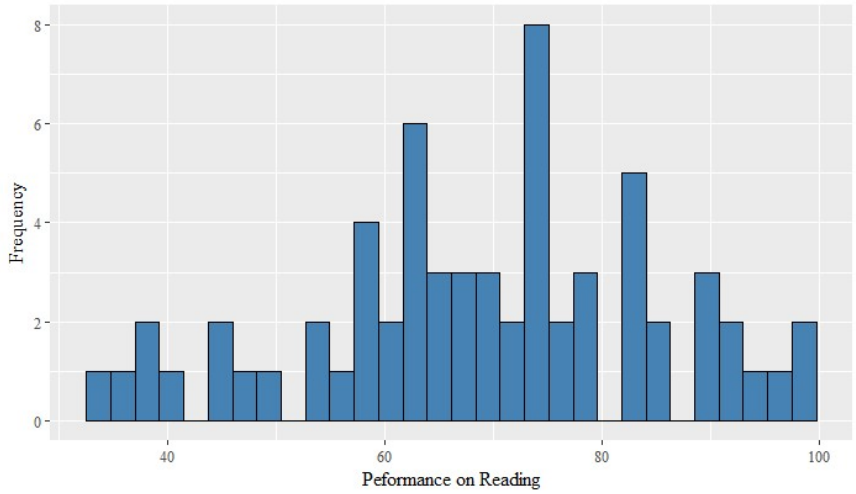

Figure 2.1. A Histogram of Overall Reading Comprehension Performance

were examined together as percent scored correct because percentage allowed for all text and test types to be compared on the same metric. Using percentage of correct answers as a metric, there are evident differences among these types of texts and tests. Table 2.2 presents the composite and the sub-sections of reading comprehension scores. The columns are organized to provide a clear understanding of the average raw score, called "Mean Correct" compared to both the total number of points possible on any given section, called "Possible Correct," and the variability of students' performance or "Standard Deviation." In addition, there is a clear metric to make some comparisons, such as expository compared to narrative texts, through the column "Mean Percent

Table 2.2. *L2 Reading Performance Measures*

	Mean Correct	Possible Correct	Standard Deviation	Mean Percent Correct
Composite	69.03	177	15.97	39%
Narrative Texts	34.88	97	9.78	36%
Expository Texts	34.16	82	7.23	42%
Free Recall	26.73	122	11.55	22%
Sentence Completion	19.69	28	3.97	70%
Multiple Choice	22.61	28	3.28	81%

Correct" where students' average scores are converted to percentages based on the number of possible points to obtain on a given section. Together these columns should provide an understanding of the spread and scoring of the data. In order to gain further understanding of these variables, a 2 x 3 within-subjects ANOVA was conducted to investigate direct and indirect effects of both text and test type, as an interaction between these effects was possible. (Results of this analysis are presented in Table 2.3.)

These results indicate that scores, when separated by test method, had statistically significant differences. This association had an effect size (η^2) of 0.162, meaning that a small yet significant effect is related to the test type. Follow-up analyses indicated that students performed

Table 2.3. *ANOVA Results*

	F statistic	Statistical Significance	Eta Squared (η^2)
Test Method	921.02	$p < 0.001$	0.162
Text Type	37.68	$p < 0.001$	0.003
Text & Test Type Interaction	26.82	$p < 0.001$	0.002

significantly better on multiple choice items than sentence completion or free recall and were more successful on sentence completion than free recall (Figure 2.2). All such test method differences were statistically significant ($p < 0.05$), accounting for Bonferroni corrections, and held considerable effect sizes ($d = 0.58$-6.39). It is evident that test type had a clear relationship with students' performance as students performed significantly better on tests with discrete question types.

Statistically significant differences were also evident between the two text types (Figure 2.2); however, such differences in terms of practical significance were much smaller, with an eta squared of only 0.03 ($\eta^2 = 0.03$). In addition, there is statistically significant interaction between text and test type, indicating that the differences among test scores is inconsistent across text types. However, this effect is small, with an eta squared of only 0.02 ($\eta^2 = 0.02$). This can be seen in Figure 2 by the relative column heights within each grouping that demonstrate slight differences. Follow-up tests of simple main effects using a Bonferroni correction were conducted

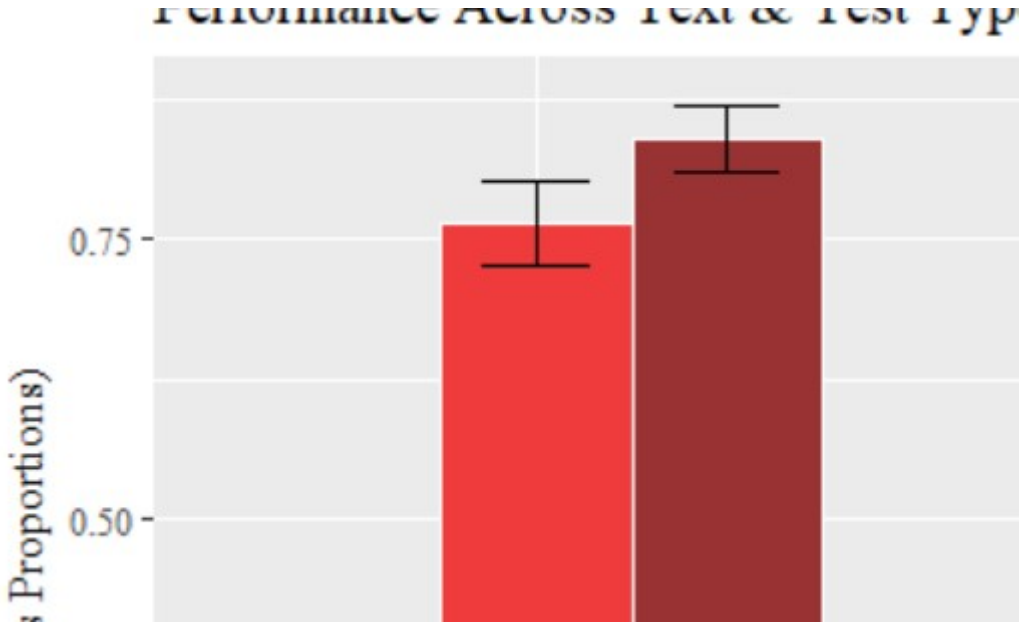

Figure 2.2. Text Type & Test Method Mean Responses with 95% Confidence Intervals

to examine specific text and test type comparisons. Results indicated that narrative and expository text types demonstrated statistically significant differences for sentence-completion (d = 0.91) and multiple-choice test types (d = 0.52), meaning that the scores on both sentence completion and multiple-choice sections were different, depending upon the type of text the learner was reading.

2. *With these students, what is the relationship between topic familiarity and L2 reading performance?*

Correlations between topic familiarity and performance are negligible and not statistically significant across all texts ($r < 0.2$; $p > 0.05$). Despite desires to further examine possible relationships with this construct, due to the lack of a significant correlation among topic familiarity and composite reading comprehension, other statistical tests were deemed no longer appropriate.

3. *How do Chinese EFL university students self-assess their L2 reading abilities? Are students able to accurately self-assess their L2 reading abilities as measured by an L2 reading task across varied text types and test methods?*

Much like reading performance, there is a wide, near-normal distribution in self-assessment scores (M = 58.45; SD = 6.05). Thus, some students assessed themselves as excellent readers while others assessed themselves as poor readers (Figure 2.3).

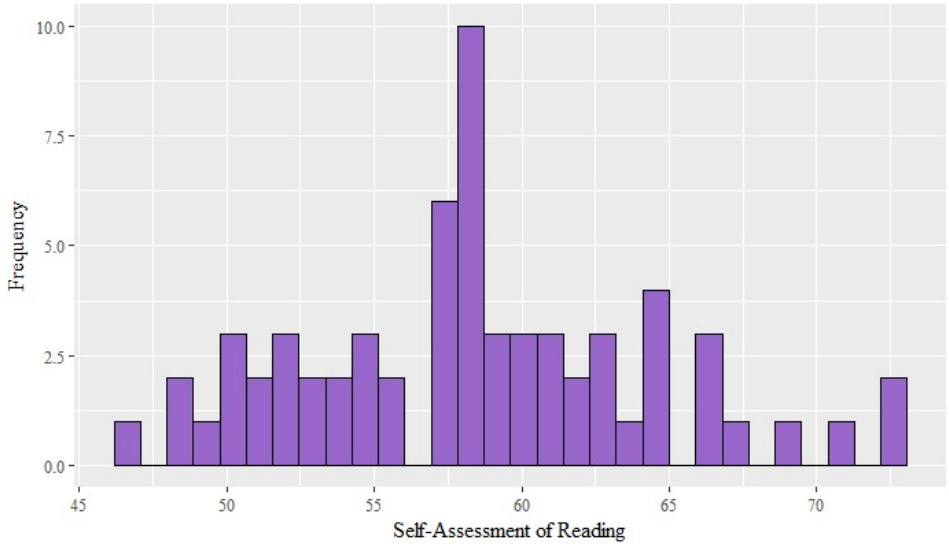

Figure 2.3. A Histogram of Overall Self-Assessment

As is illustrated in Figure 2.4, students' self-assessment ratings did correlate with performance ($r = 0.26$, $p < 0.05$). However, this overall effect may have been driven by specific relationships within the overall composite score. When considered by test method and text type (Tables 2.4 and 2.5), it is evident that self- assessment correlated with multiple choice tasks and narrative text types, reaching statistical significance with both of these sub-categories. However, self-

assessment did not significantly correlate with reading comprehension assessment scores for expository texts, free recall, or sentence completion. When further broken down by text and test

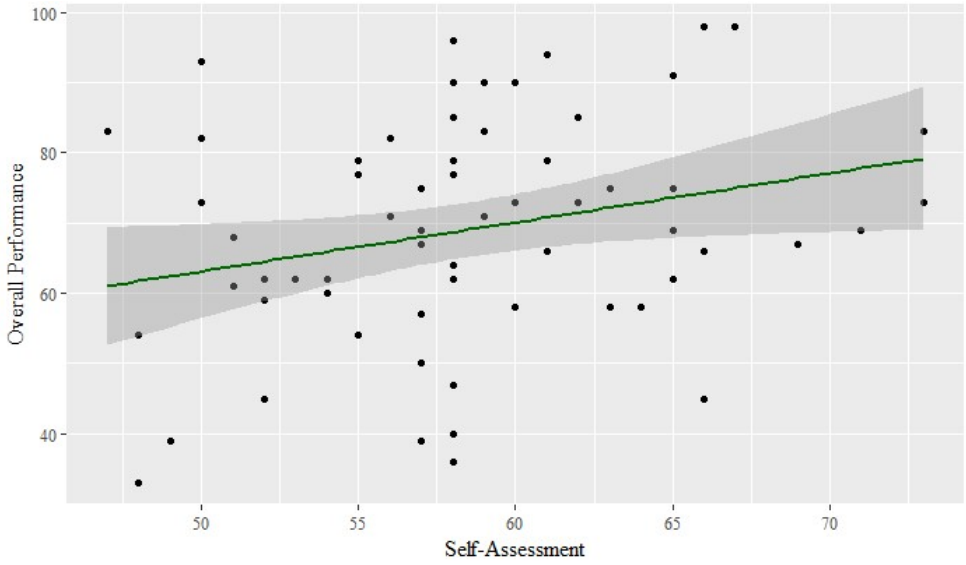

Figure 2.4. Scatterplot of Association Between Self-Assessment and Overall Performance with 95% Confidence Band

Table 2.4.*Correlation of Self-Assessment and Reading Performance Across Text Types*

	Self-Assessment	Narrative Text	Expository Text
Self-Assessment	1.00		
Narrative Text	0.28*	1.00	
Expository Text	0.21	0.76**	1.00

Note: * = p <0.05; ** = p < 0.01

Table 2.5. *Correlation of Self-Assessment and Reading Performance Across Test Types*

	Self-Assessment	Free Recall	Sentence Complet.	Mult. Choice
Self-Assessment	1.00			
Free Recall	0.19	1.00		
Sentence Completion	0.23	0.54**	1.00	
Multiple Choice	0.35*	0.45**	0.42**	1.00

Note: * = p <0.05; ** = p < 0.01

type in a single statistical test, only multiple-choice assessment of the narrative passages held a statistically significant relationship with criterion-referenced self-assessment ($r = 0.21$, $p < 0.05$). Therefore, although self-assessment appears to be significantly related to performance when these measures are combined and when, therefore, reliability is at its highest, there is evidence here that such findings may be driven by specific sub-categories of tests and texts.

2.5 Discussion

These findings, which indicate that test type or assessment task did have an impact on participants' reading comprehension performance, mirror prior research, with students scoring more correct answers when they were able to select rather than construct responses (Alderson, 2000; Wolf, 1993). While the increased cues present in the selection-based responses could explain some of this test method difference, as Transfer Appropriate Processing (TAP) might suggest (Morris et al., 1977), such an effect could also be born out of students' preparation and training for reading exams, which are largely multiple choice in the English language classroom in China (Cheng, 2008). These effects could be further heightened by the context of these Chinese university students' language learning experiences, which culminate in high-stakes

English language exams including their National College Entrance Exam (高考 /gāo kǎo) and two levels of the College English Test (Cheng, 2008). Thus, such exams may result in classroom instruction and student preparation to match these test formats, leading students to develop strong abilities with specific testing tasks. Further, the statistically significant small effects of text type could also be tied to learning experiences and fit within the paradigm of TAP, as these students are often asked to read from a variety of genres and answer questions in order to prepare them for possible short passages that exist on these high-stakes texts (Cheng, 2008; Morris et al., 1977). The importance and centrality of these assessments within Chinese culture and societal structure require that research continue to be explored within China to fully capture the realities of students' capabilities and expectations within the local context.

Results of this study also suggest that topic did not play a key role for these learners in comprehending these texts. These findings contradict much of existing scholarship on topic familiarity, which indicate that topic can have a central role in understanding a text (Barry & Lazarte, 1998; Brantmeier, 2005a; Uso-Juan, 2006; Horiba & Fukaya, 2015). However, these prior studies have often looked at the extremes of high-knowledge and low-knowledge readers, without giving weight to the variability of possible topic familiarity on a more typical reading task, such as a standardized assessment. Much like Bugel and Buunk (1996) found with a more neutral passage on a standardized examination, this study suggests that when topic familiarity is not at the extremes of high or low knowledge but rather lies across a continuous spectrum, it may impact reading comprehension less completely. Further, the importance of topic familiarity may be clouded by the proficiency of these students, which was at an advanced level (Bernhardt, 2011; Chen & Donin, 1997). At this level, it could be possible that students have developed their strategies and skills of reading to compensate for a lack of prior knowledge about a given topic

much like Chen and Donin (1997) found with graduate students studying in English (Bernhardt, 2011). Yet, with similarly advanced students, Brantmeier (2005a) found that topic knowledge was associated with comprehension outcomes. In addition, this relationship was similar for intermediate and advanced students (Brantmeier, 2005a). As findings of both this study and prior research do not coincide perfectly with our understanding, it is evident that this complexity of topic familiarity must be addressed further in future studies. While this study sought to understand topic familiarity with the context of the many variables that shape successful L2 reading (Bernhardt, 2011; Grabe, 2009), such as text type and test method, the resulting data prevented statistical tests which would allow for a multivariate examination of text, test, and topic together. These results harken to Nassaji (2002), who recommended that the relationship between topic familiarity and L2 reading be examined with caution, seeking out the nuance of this multifaceted intersection. As a result, it is vital that researchers continue to examine topic familiarity throughout the world, especially in nations such as China that rely heavily on L2 reading assessment as a pathway to success. In these high-stakes circumstances, test constructors must be certain that they are measuring reading capabilities rather than topic knowledge. In such future examinations across language learning contexts, researchers should study topic familiarity with more nuanced passages along a range of familiarity using instruments attuned to measuring topic familiarity alongside measures of the key variables of L2 reading comprehension so as to better understand this relationship among topic familiarity and L2 reading comprehension while honoring its compensatory nature (Bernhardt, 2011; Nassaji, 2002).

In addition, findings of the present study suggest that these students' self-ratings do relate to their reading comprehension performance. Such findings substantiate prior findings that support the use of contextualized items for successful SA (Brantmeier, 2005b; Brantmeier &

Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012; Butler & Lee, 2006; Author et al., 2016; LeBlanc & Painchaud, 1985; Ross, 1998). In addition, these findings may provide insight into learners' experiences that may have allowed them to successfully self-assess, as experience has been found to be a key factor in students' ability to self-assess (Suzuki, 2015). However, these findings differ somewhat from prior research on self-assessment in China (Schultz, 2017), which indicates that adult Chinese EFL students are unable to self-assess. In addition, these new findings indicate that there is a test method effect present in self-assessment as well as performance within reading comprehension. Such findings, when put in conversation with prior findings and the context of language learning, appear to indicate that self-assessment, like performance assessment, may be affected by TAP, wherein learners are better able to perform on tasks with which they already have extensive experience.

2.6 Conclusion

Despite limitations such as the lack of an L1 reading test and exact L2 proficiency measure, these results respond to a need to better understand Chinese EFL readers and L2 reading comprehension as they increasingly encounter high-stakes uses of their L2 reading skills. Findings of this study demonstrate that the role of topic in L2 reading could be more nuanced than prior research has indicated, particularly with advanced Chinese students of EFL. Further, self-assessment, text types, and assessment tasks need to be considered in concert, as findings here demonstrate that they are related and interacting variables that may provide an opportunity to unravel the underlying, multifaceted nature of L2 reading. In addition, these findings suggest that educators and researchers alike should carefully select text and test types for local contexts and the purposes of the assessment.

With consistent changes and developments to high-stakes testing of English both in China and around the world, it is vital that findings such as these are taken into account when constructing tests. The reality that preparation and assessment must align for students both to have a good understanding of their abilities and to succeed fully on the task are supported both by this study and prior research. These results might also indicate that instructors should prepare students to succeed by providing them with experiences that are similar to the tests their students will face. Such techniques may bolster both students' abilities and their awareness of their strengths and weaknesses in completing the L2 reading assessments.

In the future, researchers should undertake similar studies across varied contexts, examining these intersecting factors alongside variables that have been established in the field to impact second language reading such as L2 language proficiency, L1 reading abilities, and topic interest. Together, these factors should be examined to expand directly upon the Compensatory Model of Advanced L2 Reading and wider understandings of L2 reading. Furthermore, as suggested by Brantmeier, Vanderplank, and Strube (2012), future research should examine the value of incorporating SA as part of course requirements where students could begin to self-diagnose their reading abilities and may become better equipped to be lifelong L2 readers. Through such investigations, it may be possible to gain a more complete understanding of the complex and multifaceted experience of L2 reading comprehension.

Chapter 3: An Investigation of Interactive Online Self-Assessment Training and Advanced Second Language Reading of Spanish

Language educators and researchers increasingly support learner-centered approaches that foster greater learner autonomy within second language (L2) classrooms across levels of instruction and languages taught (Butler, 2018; Yamagata, 2018). Yet, how to empower learners as agents of their own learning (Bandura, 2001) remains a broad area of exploration. Within language learning frameworks, scholars affirm that accurate self-evaluation or self-assessment is crucial for learners to know which steps they ought to take to arrive at their ultimate language learning goals (Rivers, 2001; Ziegler, 2014). Further, the development of such accurate depictions of individual strengths and weaknesses gains further importance as more L2 learners advance through their university curriculums, reaching more advanced courses (Brantmeier, 2006b). When students reach these advanced levels of language learning, they are likely to encounter a literature-intensive curriculum where they will experience more reading of rich, complex texts that shift focus away from language learning and toward literary interpretation (Brantmeier, 2006b). Therefore, it is crucial for these students to be able to self-evaluate their own comprehension of more intricate texts so that they are able to prepare for and succeed both within their classes and beyond the classroom (Brantmeier, Vanderplank, & Strube, 2012). Yet, to date, research appears to have primarily focused on self-assessment training with productive modes, focusing on speaking and writing outcomes (Huang, Samuelson, & Chen, 2016; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Poehner, 2012). Where L2 reading has been examined with self-assessment training, research has concentrated on beginner and intermediate

readers of English, producing mixed results, with one study indicating that these learners make gains in both self-evaluative and reading comprehension abilities (Shahrakipour, 2012) while no reading comprehension or self-evaluative gains were found in another investigation (Jafarpur & Yamini, 1995). To date, despite a need to examine advanced learners' self-awareness of capabilities and online training programs which may support such an awareness, it appears that no study has directly examined online self-assessment training with advanced L2 reading. Therefore, with 136 university students studying Advanced Spanish, this quasi-experimental study seeks to determine the accuracy of self-assessment for advanced L2 readers with and without online training, evaluating whether this training yields accuracy in self-evaluations and enhances second language reading comprehension test performance.

3.1 Literature Review

3.1.1 Self-Assessment in L2 Learning

Within this study, self-assessment (SA) is defined as learners' evaluations of their own language performance or abilities, and accuracy in this self-assessment is understood to be when these self-ratings of individuals positively correspond with more traditional measures of linguistic competence such as teachers' evaluations or standardized test results (Butler, 2018; Sweet & Mack, 2017). Capabilities to self-assess with accuracy are vital because such capabilities could provide language learners with opportunities to set realistic, challenging goals which may motivate them as autonomous learners both within and outside of the language classroom (Sweet & Mack, 2017; Ziegler, 2014). In fact, since the first use of SA within the context of language learning, applied uses of SA tools were encouraged as means to develop learner autonomy and responsibility (Blanche & Merino, 1989).

Despite the intentions of Blanche and Merino (1989) many researchers first sought to examine the construct and instruments of self-assessment rather than investigating the benefits it may have in the classroom. In this way, a variety of instruments were developed. Consolidating prior work through a meta-analysis, Ross (1998) examined a variety of self-assessment studies within language learning, finding that self-assessment items which were more "concrete" in nature were more often associated with L2 performance measures. Little (2005) expanded this notion, supporting SA items that aligned with the Common European Framework (CEFR) and the DIALANG project. Expanding on these findings and her own complex SA results (Brantmeier, 2005a; 2006a), Brantmeier and Vanderplank (2008) utilized criterion-referenced SA items which were designed to situate the learner in a specific, concrete language use task. In multiple experiments with these instruments, findings indicate that such criterion-referenced self-assessment items demonstrate clear associations with measures of linguistic performance (Brantmeier & Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012). Through this research, it is evident that concrete, criterion-referenced items may provide more reliable self-assessments. Further, researchers also examined SA in terms of the learners who are completing the assessment instruments. In these examinations, results to date indicate that a learner's anxiety (MacIntyre, Noels, & Clément, 1997), linguistic proficiency (Heinlenman, 1990), experience using the language (Dolosic, Brantmeier, Strube, & Hogrebe, 2016; Suzuki, 2015), cultural background (Huang, Samuelson, & Chen, 2016), or even the language that they are studying (Ashton, 2014) may shape their ability to self-assess. In the end, the research examining SA provides an understanding of the complexity of this construct despite its straight-forward appearance. Awareness of these findings is vital to those who seek to harness the power of SA to bolster student learning.

3.1.2 Self-Assessment Training in L2 Learning at the University

When self-assessment training is used for student learning, it is embedded into courses and programs, often with the goal of providing benefits to students through close ties to course content and linguistic skill development. Several prior studies (outlined in Table 3.1) have examined the results of such training and developed foundational knowledge about the application of self-assessment. Most saliently, these studies indicate that self-assessment in the classroom allows learners to develop an awareness of their strengths and weaknesses, helping these students to accurately self-assess (de Saint-Leger, 2009; Duque Mican & Cuesta Medina, 2017; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Poehner, 2012). Further, self-assessment may allow university language learners to take greater ownership of their own

Table 3.1. *Selection of Investigations of Self-Assessment Training in Language Classrooms*

Author(s)	Participants	Training Paradigm	Results
de Saint-Leger (2009)	32 English-speaking university students studying Advanced French	Every four weeks for one semester, students completed a self-assessment and goal-setting activity with speaking	Students appeared to have an increased awareness of their abilities and responsibilities as language learners, although awareness did not always lead to concrete modifications in language learning behaviors.
Jafarpur &Yamini (1995)	60 Iranian university students in two intermediate English classes	Students in the experimental class were trained on rubrics to assess themselves and their peers on reading, listening, and speaking	Students were not able to self-assess accurately, even after training. The student who experienced self-assessment training did not score better than their "traditionally" educated peers.
Malzloomi & Khabiri (2016)	60 Iranian university students, studying English translation at the intermediate level	Students in the experimental class completed a dynamic self-assessment paradigm, self-assessing and meeting with a teacher to compare their self and teacher assessments of writing regularly for half the semester.	With these highly situated self-assessments, learners in the experimental group learned how to self-assess. Further, these learners' final papers were rated more highly than learners who had experienced more traditional instruction.

Naeini (2011)	150 Iranian university students, studying English translation at the intermediate level	Students in the experimental group used a checklist to self-assess writing throughout the entire semester.	The students who used this self-assessment checklist regularly throughout their class outscored their peers on both writing and speaking, although they were not trained on speaking. They also appeared to have higher motivation and self-esteem.
Nguyen & Gu (2013)	91 Vietnamese university students studying intermediate EFL	Students in the experimental group experience nine hours of specialized metacognitive training to help them in their writing.	Those who were trained outperformed their peers both with accuracy in their self-assessment and their actual writing performance.
Poehner (2012)	6 advanced adult learners of French in the US	Learners were encouraged to use self-assessments to reflect on the French speaking.	Learners were able to take responsibility for their learning through self-assessments.
Shahrakipour (2012)	120 Iranian university students, studying EFL at beginning and intermediate levels	Teachers led students in the experimental group through frequent self-assessments, with learners taking more responsibility for assessing themselves later in the semester.	Results indicated that both beginner and intermediate learners benefited from the self-assessment training, outscoring their peers who experienced "traditional" instruction for both reading and listening. Effect sizes differed across proficiency levels, indicating a possible interaction among these factors.
Sweet & Mack (2017)	367 American university students studying intermediate Spanish	Students completed two self-assessment tasks where they reflected on oral performance with a partner.	Learners were able to gain a better awareness of their capabilities, the process of learning language, and monitor their own progress.
Sweet, Mack, & Olivero-Agney (2017)	340 American university students of varied L2 at beginning and intermediate stages	Students completed two reflective self-assessment tasks, focusing on oral performance. Some completed the second using an online format.	All learners were more able to self-assess following training. Face-to-face formats of self-assessment training demonstrated a slight advantage over the online format.

learning (Poehner, 2012; Sweet & Mack, 2017), although it does not always lead to concrete behavior changes (de Saint-Leger, 2009). In one such study, however, Sweet and Mack (2017) outlined the use of what they refer to as the "proximal performance event" wherein learners are led through tasks which have them self-assess and demonstrate their oral performance capabilities, reflecting on the accuracy of their own self-assessments independently and with their class to establish goals for improving their language capabilities. Their results indicated that learners gained an awareness of the process of language learning and their own capabilities in addition to being more able to monitor their own progress (Sweet & Mack, 2017). Sweet and colleagues (2017) also sought to expand their project, using online formats to develop awareness of speaking capacities without requiring the proximal performance event or collaborative peer conversations. However, while learners did gain some of the same benefits through this online experience, in-class training with proximal performance and peer collaboration had a slight advantage over the online format (Sweet, Mack, & Olivero-Agney, 2017). Together these findings demonstrate the value of proximal performance as well as the possibilities that online environments could hold.

Beyond an awareness of their own skills and expanded autonomy in the language learning process, self-assessment training has demonstrated an association with higher performance for some L2 tasks that are being studied within each investigation. Specifically, learners who are trained in self-assessment outperform the students who are experiencing a more "traditional" classroom experiences on final assessments of varied linguistic skills (Malzloomi & Khabiri, 2016; Naeini, 2011; Nguyen & Gu, 2013; Shahrakipour, 2012). Despite the variation in duration, level, and time of the semester that testing was conducted throughout these studies, together, these findings clearly demonstrate that self-assessment can be a powerful tool.

For studies focused on writing, results appear to be especially strong even when varied types of self-assessment and metacognitive training are used across differing language learning contexts. For example, with Iranian university students during one semester, Naeini (2011) trained one class to use a check-list to self-assess their writing in English while the other class was taught with more traditional language learning activities. Findings indicated that students who used this self-assessment checklist regularly throughout their semester-long class outperformed peers who had not been exposed to the checklist. Further, these students who received the checklist training also appeared to outscore their peers on writing. In another study of self-assessment training and writing, Malzloomi and Khabiri (2016) used a more dynamic training paradigm, having students self-assess and discuss their self-assessments with an instructor regularly throughout the course while their peers in another class were the control group, receiving no intervention. In the end, those who had experienced the guided self-assessments and regular meetings over eight weeks of class were better able to self-assess and produced better writing samples (Malzloomi & Khabiri, 2016). In another setting, Nguyen and Gu (2013) implemented a single metacognitive training session prior to the start of the semester to understand how such self-evaluative training might support learners' development across the semester. Findings indicated that learners who were exposed to this training, as compared to peers who were not, outperformed their classmates on both self-assessment accuracy and the quality of their writing samples. Therefore, studies of self-assessment training with writing appear to promise benefits to learners.

Despite these findings demonstrating benefits for learners who receive SA training, other studies reveal complexities in harnessing SA for supporting student learning. For example, Jafarpur and Yamini (1995) conducted a study training intermediate EFL learners to self-assess

and peer-assess using specific rubrics across a semester for reading, listening, and speaking. However, these learners were not able to outperform their peers in measures of self-assessment accuracy or performance (Jafarpur & Yamini, 1995). Jafarpur and Yamini (1995) explained these results through the lack of direct experiences that their learners had with using English and this explanation aligns with growing understandings of the complexity of SA as a construct and the difficulty in implementing SA as a pedagogical tool. While results were more positive for Shahrakipour (2012) whose study demonstrated that beginner and intermediate learners who completed reading and listening SA throughout their course outscored their peers, their findings also demonstrate complexity. Specifically, the effect sizes of SA training for intermediate and beginning learners differed, signifying an interaction in the relationship between training and performance. Shahrakipour (2012) argues that this difference is due to learners' proficiency, indicating that training may have different effects on students of differing initial proficiency. Such findings align with investigations of self-assessment which demonstrate that self-assessment capacities are different across stages of acquisition (Brantmeier, et al., 2012; Heinlenman, 1990). In addition, de Saint-Leger (2009) found that learners who were encouraged to self-assess and set goals once per month throughout a semester-long French course gained abilities in defining their own strengths or weaknesses, but these students did not necessarily use this knowledge to improve their oral performance abilities. Together, therefore, these studies demonstrate that learner factors may also shape both accomplishments and disappointments in pedagogical implementation of SA training. As a result, the great promise that self-assessment offers to learners must be explored more closely, particularly across linguistic skills, varied proficiency levels, and diverse language learning contexts. Therefore, this study specifically

examines L2 reading comprehension self-assessment training with advanced students studying university-level Spanish in order to expand our current understandings.

3.1.3 L2 Reading Comprehension

The processes and products of L2 reading comprehension are also complex and are frequently being shaped by the reader, the text, and the task in a comprehensive system of factors that contribute to ultimate comprehension. Reading, even in a first language, is an interactive process which occurs through both "bottom-up" and "top-down" processes where learners are uniting information from individual words as well as using their knowledge of discourse structure to understand the text they are reading (Grabe, 2009; Koda, 2005). Specifically, features of the text, including its complexity, length, and genre can have an interactive effect with learner approaches to and understanding of a text (Bernhardt, 2011; Grabe, 2009; Koda, 2005). Further, learners' prior experiences and knowledge can contribute to their comprehension of a text, with an individuals' first language (L1) reading abilities and L2 language knowledge (of grammar and vocabulary) explaining nearly 50% of his or her abilities to comprehend L2 passages (Bernhardt, 2011).

Adding complexity to L2 reading, other correlates such as topic familiarity, working memory capacity, and topic interest can have meaningful, sometimes interactive, impacts on L2 reading comprehension performance (Jeon & Yamashita, 2014; Juffs & Harrington, 2011; Puldio, 2007). In addition to these correlates, when assessing L2 reading, the methodologies used to capture reading comprehension performance can also shape the demonstrated comprehension of similar passages (Alderson, Bachman, Perkins, & Cohen, 1991; Alderson, Clapham, & Wall, 1995; Brantmeier, 2005a; Carrell, 1985; Riley & Lee, 1996; Shohamy, 1984). For example, when students complete a free recall or answer multiple-choice questions, they will

demonstrate different facets of comprehension for the same text. In addition, learners may be more familiar or more readily equipped to handle specific text types, which can shape not only performance but also abilities to evaluate one's own performance (Dolosic, 2018). Specifically, the second chapter of this dissertation found that self-evaluations of performance may more closely align with task types that learners are comfortable and familiar performing, such as multiple-choice (Dolosic, 2018). Together, these features of the text, the reader, and the task represent the complexity that shapes L2 reading as they all contribute to our understanding of an individual's ability to read. While researchers have sought to examine reading both within and outside of L2 contexts for decades, much remains to be understood about these processes and the assessment of L2 reading, including opportunities which may support greater L2 reading comprehension achievement.

Beyond the challenges of L2 reading, advanced L2 reading is distinct from beginning and intermediate L2 reading. First, as is evident in the ACTFL and CEFR frameworks, becoming an advanced learner of language requires a large commitment to improving the precision of language capabilities, well beyond that of an intermediate learner. Moving further toward language learning goals, developing skills within the "Advanced" language learning level, perhaps seeking superior or near-native status, is likewise challenging, and at times, only small gains may be visible within any given semester (Byrnes, 2006). In addition, once learners have reached advanced stages of courses and curricula, they are likely to be reading texts for much more than linguistic skill development, learning about examining the many layers within a text to gain access to more complete cultural, historical, and political meanings. More precisely, as Brantmeier (2006b) reminds us, within advanced language classrooms, "the instructional practices shift from a focus on language skills to an emphasis on text analysis and interpretation"

(p. 3). Therefore, there is an on-going need to further understand students at these levels, particularly so that they may be able to develop an awareness of their capabilities to understand a given L2 text in order to read it more deeply, within its larger context.

3.1.4 Online Training for L2 Learning

Prior research indicates that self-assessment training has not traditionally relied on online resources. Yet, online training has proven effective for training students to use other technologies to bolster language learning; for example, online tools have been used to introduce learners to computer assisted language learning (CALL) environments, supporting knowledge of the computer technologies and developing linguistic capabilities (Lai & Gu, 2011). Implementing Hubbard's (2004) framework, researchers created a series of modules which first provided pedagogical support, explaining the usefulness of the training. Then, researchers moved into tasks that allowed the learner to implement what they had learned about, and finally reflect on their experiences (Lai & Gu, 2011). As Smith and Craig (2013) discovered, such training can benefit learners when it comes to using CALL environments, providing experiences to support their successful use of CALL to gain linguistic competencies. Using technology-driven applications to build learners' self-awareness could benefit students in their abilities to self-assess and take ownership for their own learning.

The tool used in the present study is an online digital environment built to mirror prior criterion-referenced and reflective approaches to self-assessment (Brantmeier, et al. 2012; Little, 2009; Sweet & Mack, 2017; Sweet, et al., 2017). It is designed so that learners were given an opportunity to complete a reading comprehension exercise and evaluate their own performance, comparing it to their own ratings. Such a program could provide learners with opportunities to

improve their self-assessment accuracy, goal setting practices, and L2 reading comprehension. Yet, to date, no study of this kind has directly examined such a tool with Advanced L2 readers. Implementing Hubbard's (2004) framework to provide principled self-assessment training, the following quasi-experimental study examines the benefits of such an online training environment, guided by these questions:

3.2 Research Questions

1. With university students enrolled in an Advanced Spanish course for one semester, what are differences in the L2 reading comprehension self-ratings with and without self-assessment training?
2. How do these students' self-ratings align with performance? Are these learners able to accurately self-assess their L2 reading comprehension across three different testing tasks? How does the relationship between self-assessment and performance differ among learners with and without the self-assessment training?
3. With these learners, are there significant differences in reading comprehension performance between groups who did or did not complete self-assessment training?
4. For students who completed self-assessment training, what did they report as their own strengths and weaknesses within the training? How did they set goals to overcome their weaknesses in Spanish reading comprehension?

3.3 Methods

3.3.1 Participants

Participants for this study were drawn from university students (aged 18-21) who placed into and subsequently enrolled in an Advanced Spanish Course at a mid-sized, highly-selective private university in the United States. The course, entitled "Grammar and Composition I,"

focused on refining linguistic skills in listening, speaking, reading, and writing while exploring the culture of the Spanish-speaking world. Participants completed this course during the Spring semester of 2017 or 2018. The students completing this course during the first semester of this study (in 2017) were not provided with any self-assessment training, but the students who completed this course in the second Spring (in 2018) completed at least three of four possible self-assessment modules to be included in the study. All students earned participation credit in their courses for completing paper reading tests that were used as measures in this study. In the end, the 136 individuals who completed their assigned training and measures were included in this analysis. Within this sample, 74 (21 self-identified men; 53 self-identified women) completed no training and 62 (16 self-identified men; 44 self-identified women) completed training.

3.3.2 Procedure

In this quasi-experimental design, students taking the same course but in different semesters were considered comparison groups. For students who completed the Advanced Spanish Course in first Spring (2017), no interventions were put into place. During Week Thirteen of the semester at a regularly scheduled class meeting, students completed a written test measuring their self-assessment and reading comprehension. Students were given one hour to complete the test. All students progressed linearly, completing one page before moving onto the next and never returning to the pages that had come before. The researcher or an assistant was present at all data collection times which took place in the students' normal classrooms.

The students who completed the Advanced Spanish Course in second Spring (2018) completed at least three homework assignments where they were asked to critically evaluate their abilities in reading comprehension and complete a reading comprehension task. There were

four possible opportunities to complete self-assessment training, and students were considered to have completed their training when they completed any three of these four. These assignments were spaced across the first twelve weeks of the course, with students completing about one training each month through April. Then, in Week Thirteen of their course, these students were given a test measuring their self-assessment and reading comprehension during a normal class session. This test was identical to the test that had been distributed to the students the year before. These students were also given one hour to complete their written test. Again, students progressed one page at a time, completing each page before moving onto the next and never returning to prior pages. All test packets were collected for analysis by the researcher or a trained assistant.

3.3.3 Materials

3.3.3.1 Online Self-Assessment Training Modules. Four unique Spanish reading self-assessment activities were constructed for the Advanced Spanish course. Each self-assessment training activity consisted of information about self-assessment, a self-assessment questionnaire, a short reading passage and comprehension task, and a guided reflection on performance and self-assessment. The texts and comprehension tasks were varied across open and closed response types, providing a variety of reading comprehension experiences, including summaries, free recalls, multiple-choice questions, and sentence completion tasks for both narrative and expository text types. Samples of the self-assessment items are included in Appendix A. The sequence of these trainings was built to mirror successful interventions in L1 reading (McDaniel, Howard, & Einstein, 2009) and L2 speaking (Sweet & Mack, 2017) where learners were provided with an opportunity to complete an activity and then evaluate their own performance, comparing their actual performance with expectations. The technological design of the web-

based platform was built to capitalize on the affordances of technology while minimizing the risks of integrating new online programs into courses. (See Chun, Kern, and Smith (2016) for full discussion.) Through the use of this digital space, along the guidelines of Hubbard's (2004) framework, the activity is carefully constructed such that learners (1) are introduced to the activity and its pedagogical purpose, (2) complete an activity using self-assessment and L2 reading skills, and (3) compare the passage to their responses by answering reflective prompts.

3.3.3.2 Criterion-Referenced Self-Assessment. For this self-assessment instrument, learners were situated into a specific context of reading comprehension with a short prompt. Then, the learner was asked to rate their ability to complete this task within the prompt. Learners could reply with a range of responses on a five-point Likert scale from five ("Strongly Agree") to one ("Strongly Disagree"). For example, an item might read, "When reading a story, I understand action sequences, knowing who is doing what action in the story." A student was able to rate their own abilities by agree or disagreeing with the statement across the five-point scale of responses. The instrument used in this study was tailored from prior research which brought the DIALANG framework into self-assessment, utilizing criterion-referenced self-assessment (Brantmeier, et al., 2012; https://dialangweb.lancaster.ac.uk/). Examples of these items can be found in Appendix A.

3.3.3.3 L2 Reading Comprehension Assessment. L2 reading comprehension was measured through a carefully constructed set of tasks, developed using Wolf (1993) and Alderson's (2000) guidelines. Two texts were selected, representing narrative and expository text types. These texts were equated and scored by a native speaker of Spanish for expected proficiency level of the reader, length, and syntactic complexity. Length was measured through number of words, and syntactic complexity was measured through embedded clauses. (See Table

3.2 for further details.) Free recall, sentence completion, and multiple-choice were selected as reading comprehension tasks to provide a comprehensive picture of learners' L2 reading. Free recall was conducted in English and scored in terms of pausal units (Brantmeier, Strube, & Yu, 2014). These pausal units are said to be where a native speaker would pause as reading the text aloud (Brantmeier, et al., 2014). Units scored were verified by a second trained rater. Sentence completion items were scored as correct or incorrect based on pre-determined answers. Multiple-choice items were developed using Wolf's (1993) guidelines and validated by a native speaker. The results of this instrument are examined both as a composite and as individual tasks.

Table 3.2. *Summary of Reading Passages & Comprehension Tasks*

Passage	Language Assessed	Word Count	Embedded Clauses	Assessment Items
La Antártida: Historia y Leyendas (Antarctica: History and Legends)	Spanish	656	17	Written Free Recall 10 Sentence Completion 10 Multiple Choice Items
La Tortuga Gigante (The Giant Turtle)	Spanish	656	20	Written Free Recall 10 Sentence Completion 10 Multiple Choice Items

3.3.4 Analyses

Statistical analyses were conducted to answer the research questions posed. The statistical analyses presented here were conducted with only complete observations of all variables without removing any extreme cases because all extreme cases were determined to be part of the true population of learners. Analyses were completed in the open source statistical program, R version 3.3.3 (R Core Team, 2017), using additional packages including psych (Revelle, 2017), psychometric (Fletcher, 2010), and car (Fox & Weisberg, 2011). Figures were generated using GGPlot2 (Wickham, 2009).

3.4 Results

RQ1: *With university students enrolled in an Advanced Spanish course, what are the differences in the L2 reading comprehension self-ratings with and without self-assessment training?*

On the Criterion-Referenced Self-Assessment questionnaire (Cronbach α = 0.85), the actual self-ratings of individuals who received training were not different from those who had not received training, with similar means ($t = 0.1$; $p = 0.9$) and variances ($F = 0.04$; $p = 0.85$). Table 3.3 and Figure 3.1 present these data and demonstrate that the ratings provided across students did not substantially change on average, approximating a nearly normal distribution in both groups.

Table 3.3. *Descriptive Statistics of Criterion-Referenced Self-Assessments: Treatment & Control*

	Mean	Standard Deviation	Min.	Max.
Control	34.39	4.87	17	44
Treatment	34.50	5.26	20	47

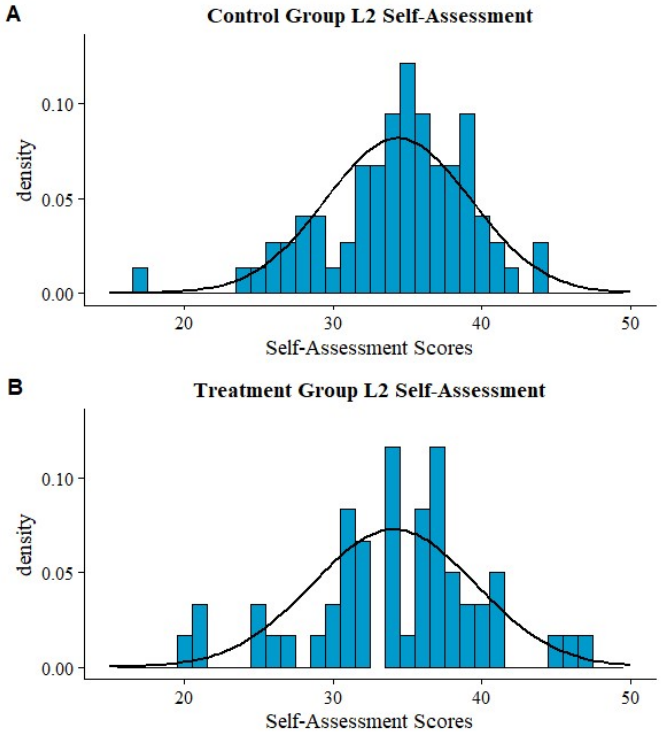

Figure 3.1. *Histograms of Criterion-Referenced Self-Assessments for Treatment & Control*

RQ2: *How do these students' self-ratings align with performance? Are these learners able to accurately self-assess their L2 reading comprehension across three testing tasks (Free Recall, Sentence Completion, and Multiple-Choice)? How does the relationship between self-assessment and performance differ among learners with and without the self-assessment training?*

Overall, the self-ratings of those with training matched their L2 reading comprehension performance more closely ($r = .47$, $p < 0.05$) than those who had not had training ($r = 0.16$, $p > 0.05$). Using a Fisher's Z test, the difference between these two Pearson product moment correlations was determined to be statistically significant ($p < 0.05$) (Meng, Rosenthal, & Rubin, 1992). Together these statistical results indicate that the learners who had completed training were able to produce self-ratings that aligned more closely with their actual L2 reading comprehension performance, as is evident in Figure 3.2.

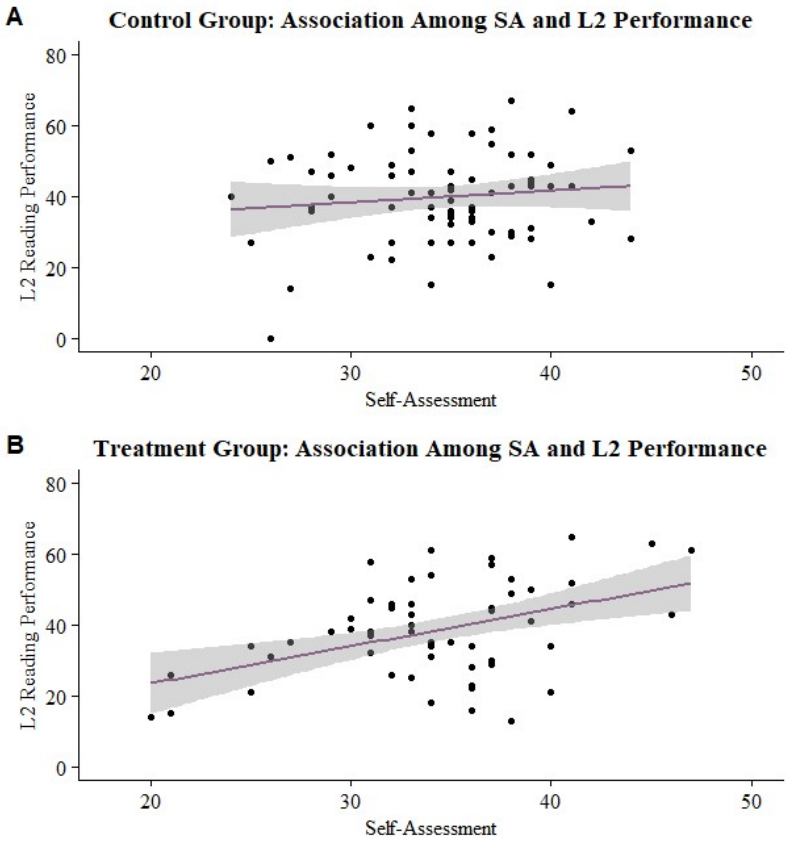

Figure 3.2: *Associations Between Self-Assessment and L2 Reading Performance*

When examining these correlations for specific test types, the difference between those who experienced training and those who did not becomes more apparent for specific types. (See Table 3.4.) More precisely, with Sentence Completion and Multiple-Choice testing types, correlations are both significantly different, demonstrating statistical significance only with training. For Sentence Completion, the correlation with training is 0.20 higher and for Multiple choice this difference is 0.40, indicating a substantial difference that was verified through another Fisher's Z test of significant differences ($p < 0.05$). However, students' ability to accurately self-assess did not demonstrate statistically significant differences for Free Recall measures of reading comprehension ($p > 0.05$). In the end, for those who received training the correlations are highest among self-assessment and more typical measures of reading comprehension such as sentence completion and multiple-choice.

Table 3.4. *Correlations Among Self-Assessment and L2 Reading Comprehension Performance*

	Control	Treatment	Z-score (p-value)
Composite Scores	0.16	0.47	1.98 ($p = 0.05$)
Free Recall	0.17	0.37	1.23 ($p > 0.05$)
Sentence Completion	0.08	0.48	2.51 ($p < 0.05$)
Multiple Choice	0.15	0.48	2.11 ($p < 0.05$)

RQ3: *With these learners, are there significant differences in reading comprehension performance between the groups who did or did not complete the self-assessment training?*

Reading comprehension performance, much like self-assessment ratings, are nearly identical across the two groups. Therefore, differences between the composite scores of students who have experienced self-assessment training and those who have not are not statistically significant when examined through simple ANOVA comparisons. (Results presented in Table 3.5.)

Table 3.5. *Comparisons of Mean L2 Reading Comprehension Performance*

	Means	F-Statistic	Significance
Composite Scores			
Control	39.6	0.30	$p = 0.587$
Treatment	39.5		
Free Recall			
Control	14.4	0.66	$p = 0.418$
Treatment	16.3		
Sentence Completion			
Control	10.9	8.53	$p = 0.004$
Treatment	9.3		
Multiple Choice			
Control	14.2	0.65	$p = 0.422$
Treatment	14.0		

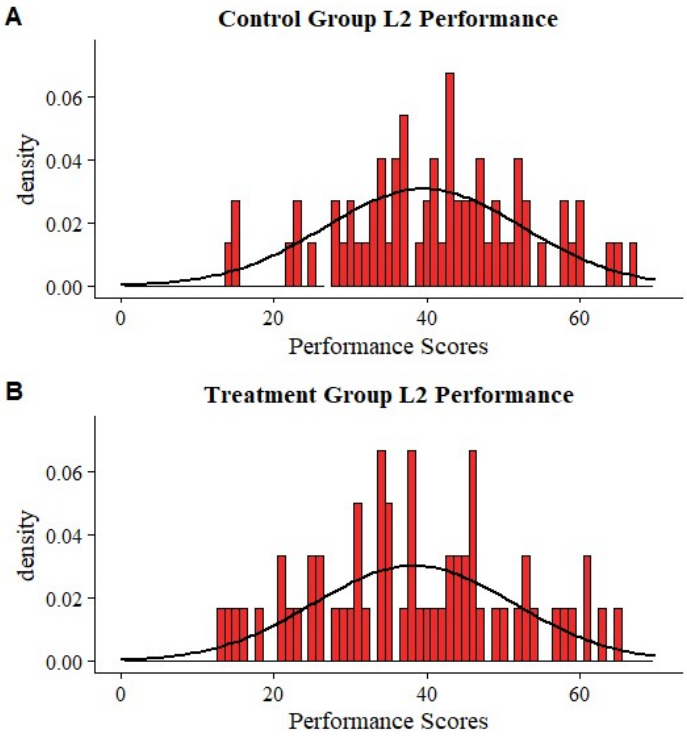

Figure 3.3. *Histograms of Composite L2 Reading Comprehension for Treatment & Control*

However, performance on sentence completion tasks demonstrated statistically significant differences, with the students who had not experienced self-assessment outperforming students who had experienced self-assessment, although this difference held a small effect size ($\eta^2 = 0.05$).

RQ4: *For students who completed self-assessment training, what did they report as their own strengths and weaknesses within the modules? How did they set goals to overcome their weaknesses in Spanish reading comprehension?*

When students completed their self-assessment training, they practiced self-assessing at least three times throughout the semester, providing a score by answering a series of items asking about their abilities to read in Spanish on a one to five Likert scale. The average score across all the items for each individual training module is reported in Table 3.6. These scores demonstrate that the learners rated themselves somewhat consistently across the semester with variations being limited to less than one point on average.

Following their self-assessments and reading comprehension task, students were asked to reflect upon their abilities and performance, stating what they found to be their strengths and weaknesses. In a quantitative analysis of these open-ended responses within two select modules, students' responses were read, categorized, and counted based on the themes that emerged from student answers. Specifically, when answering the question: "What surprised you when you compared the text to your response?" for their first training, while 32% of students responded that nothing at all had surprised them, 22% stated that they had done better than they

Table 3.6. *Self-Assessment Ratings from within Online Training Sessions*

	Training 1	Training 2	Training 3	Training 4
Mean (SD)	3.45 (0.50)	3.76 (0.45)	3.51 (0.50)	3.48 (0.43)

had expected in some way and 15% reported that they had done worse than expected. Within this group, some (16%) specifically stated that they were surprised they had left so much detail out of their answers to the short questions in training one.

Students also reported their exact strengths and weaknesses when prompted to provide examples that they may have identified while completing these activities. For one such training, 69% of students stated that they had strengths in identifying the key plot while 30% indicated that their weaknesses were in vocabulary knowledge and 40% reported that they had a weakness when it came to parsing and recalling key details in the reading passage. For the final question of each training task, students were asked what actions they might take to improve their reading after completing these tasks. Although 6% of students did report that they would take no concrete steps outside of class, 37% indicated that they would read more, 22% suggested that they would work on their knowledge of Spanish grammar and vocabulary, and 35% stated that they would adjust their approach to reading in Spanish, by slowing down or reading certain sections with greater care.

3.5 Discussion

Within this study, students who experienced self-assessment training provided self-ratings that correlated more directly with their actual reading comprehension performance with the largest correlation differences present among self-assessment scores and performance on sentence completion and multiple-choice test items. These results align with prior results of self-assessment training across skills which demonstrate that learners are better able to diagnose their own strengths and weaknesses through training (de Saint-Leger, 2009; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Poehner, 2012). In addition, these results align specifically with the findings of Sweet and Mack (2017) who indicated that for speaking, self-assessment tasks across

one semester that involved a "proximal performance event" produced an awareness of capabilities. Much like Sweet and Mack's (2017) paradigm, the procedure and materials used in this study allowed learners to self-assess and complete a linguistic task, comparing their performance to their self-assessment and setting goals for their future experiences with the language. As a result, it is evident that "proximal performance events" may be likewise valuable in semester-long courses to developing self-assessment and goals setting capacities for reading comprehension abilities with advanced students throughout the semester. In addition, findings of this study are promising for the use of online formats to facilitate the benefits of proximal performance in an online setting as this study provides an implementation of an online format of such self-assessments which include a required proximal performance event. As a result, this study indicates that programs like this one may be able to encourage learners to self-assess with accuracy even in the home environment, expanding successful applications of such online self-assessment training programs to L2 reading comprehension.

Examining the self-assessment results more closely, they also clearly indicate that differences in accuracy of self-assessment are greatest for sentence completion and multiple-choice measures, with those who were trained outscoring those who were not trained in terms of accuracy. Despite efforts to provide learners with diverse practice conditions in the online training such as summarizing, completing question items, and recalling the content of the passage, the correlations for sentence completion and multiple-choice have the most evident differences. While these findings highlight the complexity of the construct of self-assessment (de Saint-Leger, 2009; Jafarpur & Yamini, 1995), they also align with previous understandings of the interaction among test method and self-assessment in L2 reading comprehension (Brantmeier, 2005a; 2006a; Dolosic, 2018). More precisely, studies have previously

demonstrated that learners' self-evaluations aligned more directly with specific testing conditions, such as multiple-choice or sentence completion (Brantmeier, 2005a; 2006b; Dolosic, 2018). In one study, Brantmeier (2005a) even found that learners were able to accurately self-assess with descriptive items only on measures of free recall. Yet, the findings of the first study presented in this dissertation suggested that learners may be more able to accurately self-assess when familiar with the testing task. Such an explanation would align with learners' experiences within this context because these two types of reading comprehension assessment, sentence completion and multiple choice, are quite common in the context of language classrooms at the university, existing in textbook exercises, in-class practice, placement tests, and even final exams. Therefore, these results highlight the complexity associated with self-assessment and self-assessment training of reading comprehension across test types. As a result, such findings ultimately indicate that for advanced L2 readers self-assessment training with a reflective task could help learners to develop abilities to self-assess with accuracy on typical forms of assessment, expanding previous understandings of who might benefit from such training.

Findings of this study also indicate that these L2 learners were able to develop and set individualized goals for themselves through this reflective self-assessment training, focusing on their approaches to texts, linguistic development, and a variety of other possible steps they could take to succeed in future L2 reading tasks. Yet, prior research highlights complexity of this goal setting and implementation within self-assessment training has offered some mixed results. For example, while learners often gain capacities as autonomous learners through training with self-assessment and metacognitive foci (de Saint-Leger, 2009; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Poehner, 2012), de Saint-Leger (2009) discovered that, even when students were able to state their own capabilities with greater accuracy following the training, they did not

seem to take on more responsibility or follow-through with the goals they had set. Therefore, findings of this study align with prior research in that learners were able to set individualized goals and develop an awareness of their own capabilities; however, it remains unknown as to whether or not these students implemented these individualized goals. For these advanced L2 readers, their responses indicate that they are taking initiative to develop their own goals that correspond with their individual needs which could be especially fruitful for achieving personal and professional goals of language learning across their lifetimes.

While findings that indicate that these L2 learners who experienced self-assessment training developed capabilities to self-assess with accuracy and set individual goals, these students' L2 reading comprehension performance was not significantly different from their untrained counterparts, following three online trainings in twelve weeks of one semester. Such findings align with broader understandings of L2 reading. Specifically, Brantmeier (2006a), Bernhardt (2011), and Byrnes (2006) have described the complexity of advanced L2 learning, identifying it as distinct and separate from earlier stages of language learning. For example, Byrnes (2006) and Swain (2006) indicate that such advanced learners are able to use their advanced competencies to overcome the challenges that they encounter in the language, unlike learners at other stages of acquisition. Bernhardt's Compensatory Model (2011) likewise highlights the adaptability of these advanced language learners specifically when reading in a L2.

In addition, within the ACTFL and CEFR guidelines attaining "Advanced" or higher reading comprehension capabilities requires a large commitment to language learning, and as such, discrete skill improvement may not be evident for every skill each semester for all learners. In addition, within this study, ultimate comprehension was also the measure, meaning that "ease" and "effort" could not be adequately measured, although the ease of reading is considered in the

ACTFL guidelines determining "Advanced" reading capabilities (Swender, Conrad, & Vicars, 2012). Therefore, the discrete improvements that self-assessment intervention might have made with these learners could also have been more challenging to discover at this advanced level with this reading test. As a result, while Shahrakipour (2012) found that beginning and intermediate learners gained self-assessment and reading comprehension abilities that outscored untrained counterparts, current findings match these larger understandings of 'advancedness' in L2 reading, suggesting that advanced learners differ from patterns exhibited by beginning and intermediate learners. Further, in the case of Shahrakipour, (2012) tests were conducted at the very end of the semester with a full course being used for training. This length of training may likewise have affected students' results. Therefore, future studies and meta-analyses should examine varied lengths of self-assessment intervention in detail to better understand possible impacts. With these advanced L2 learners, therefore, further examination may be needed to discover exactly how self-assessment training can be implemented to best benefit these unique advanced L2 readers by increasing overall quantity, number of semesters, frequency of use, or other such factors in self-assessment training.

3.6 Conclusion

Together, findings of the present study demonstrate that university students studying Spanish were able to self-assess more accurately if they had received online self-assessment training, reporting specific individualized plans to overcome their self-identified weaknesses. Further, while advanced L2 learners who had experienced self-assessment training did not outperform as compared their counterparts who had not been trained, such results aligned with prior understandings of the complex, compensatory nature of L2 reading at advanced stages. These findings expand previous understandings of pedagogical implementations of self-

assessment, discovering nuance at this advanced level of language learning. In applying these findings to classrooms, instructors and administrators should consider implementing these initiatives to support the development of learner autonomy, tailoring the length and timing of the training to meet their curricular goals. In addition, future studies should build upon this work, examining the paradigm of online self-assessment training for L2 reading comprehension at varied stages of language learning with varied levels of frequency and duration in order to establish a more complete understanding of its possible benefits. Further, such examinations should include measures of motivation, self-confidence, and other key correlates for developing learner autonomy in modern learner-centered classrooms.

Chapter 4: An Individualized Approach to Self-Assessment with Readers in the French Village

In recent years, the number of American college students studying abroad has been increasing, with the number of students studying abroad doubling from 2000 to 2015 (Institute of International Education, 2015). Furthermore, the focus of this growth is on short-term, intensive programs, with about 95% of students staying abroad for one semester or less (Mitchell, Tracy-Ventura, & McManus, 2017). Much like this growing interest in intensive short-term abroad programs, students' interests in intensive experiences in language learning in their home country is also growing (Isabelli-Garcia & Lacorte, 2016). Such programs have been examined for their abilities to foster oral production as well as their capacities to support learners' awareness of their own strengths and weaknesses (Dolosic, Brantmeier, Strube, & Hogrebe, 2016). While enhanced oral production was found to be a strength these programs, focuses of such immersion programs are shifting as directors must respond to the 21st century demands that learners encounter where skills to interact in written formats in the L2 is fundamental (Grabe, 2009).

However, despite these growing demands, it appears that few studies to date have examined second language (L2) reading outcomes with short-term intensive immersion programs. Rather, research has often focused on (1) first and second language outcomes from long-term schooling programs (such as Fortune & Tedick, 2008; Genesee, 2008; Lindholm-Leary, 2011; Lyster & Mori, 2006; Montanari, 2014), (2) cultural and communicative outcomes from foreign exchange programs (Llanes & Muños, 2009; Llanes Tragant, & Serrano, 2012; Savage & Hughes, 2014) or (3) specific grammatical or vocabulary development in short-term settings (Briggs, 2015; Isabelli-Garci & Lacrote, 2016). Therefore, grounded in broader

understandings of immersion for language acquisition, the following study examines high school students' French language knowledge and L2 reading development through a short-term domestic immersion experience, answering (1) how learners' knowledge of the language, in terms of grammar and vocabulary, develops; (2) how learners' discourse-level French reading comprehension develops; and (3) how accurate individual students are in their own assessments of their capabilities before and after this program.

4.1 Literature Review

4.1.1 Reading in a Second Language

Reading comprehension is a multifaceted, complex process that lies at the center of three key components: the reader, the text, and the task (Bernhardt, 2011; Grabe, 2009; Koda, 2005; Urquhart & Weir, 1998). Distilling this process and its products into a single sentence, Koda (2005) indicates that reading is (1) "extracting information from the text" and (2) incorporating it "with what is already known" (p. 4). Urquhart and Weir (1998) refer to this incorporation as an interpretation in their definition which describes reading as "receiving and interpreting information" that is presented in print, going beyond decoding the words to fully understand the message of the text (p. 22). Within this overarching definition, Grabe (2009) further contends that the linguistic process of reading is mutable to the forces of purpose and strategy that a learner may use to comprehend and learn from a text. In their work, Koda (2005; 2007) and Bernhardt (2011) highlight the complexities and intricacies of the multiple language systems that are used to understand texts in the readers non-native language, stating that this process is unique in that it relies on both the first and second language capacities to read. Therefore, learning to read in a second language can be a complicated process where both L1 and L2 resources are

brought forth in order to make sense of the text (Bernhardt, 2011; Grabe, 2009; Koda, 2005; Wu, 2017).

Within this complex process, many factors have been found to shape learners' successes in L2 reading comprehension including L2 grammar knowledge, L2 vocabulary knowledge, L1 reading comprehension, working memory capacity (WMC), topic interest (TI), topic familiarity (TF), text structure, and methods of assessment (Alderson, 2005; Bernhardt, 2011; Brantmeier, 2006a; Brantmeier, 2006b; Grabe, 2009; Jeon & Yamashita, 2014; Koda, 2005; Lee & Pulido, 2017; Pulido 2007). The Compensatory Model of L2 Reading (Bernhardt, 2011) specifically highlights the roles of L2 language knowledge and first language (L1) reading comprehension skills. This model, based on empirical studies with adolescents and adults, states that these two factors explain 50% of the variability in their L2 reading comprehension performance, with L1 reading comprehension explaining 20% while L2 language knowledge is able to explain 30% (Bernhardt, 2011). Language knowledge is conceptualized as vocabulary and grammar knowledge of the L2 (Bernhardt, 2011). Extending this framework, while indicating clear L1 and L2 reading comprehension connections, Grabe (2009) also highlights the central value of grammar knowledge and vocabulary in knowledge in L2 reading across stages of acquisition, signifying that word recognition and syntactical processing are crucial to learners' abilities to read. Further, empirical meta-analysis indicates that L2 vocabulary and grammar are two of the greatest correlates with L2 reading comprehension performance (Jeon & Yamashita, 2014). As a result, the present study examines both the development of discrete linguistic knowledge in terms of grammar and vocabulary while also examining the complex, discourse-level reading comprehension capacities of these learners in their L2: French.

4.1.2 Reading in Immersion Settings

Part of a larger tradition of bilingual education, immersion education is a specific schooling model wherein learners gain access to common curricular subjects such as math, science, and history through the use of a non-native language, for at least 50% of their instructional time, in order to promote additive bilingualism (Genesee, 2008; Tedick, Christian, & Fortune, 2011). Instructors within these contexts use "immersion techniques" which support language acquisition through content learning, scaffolding vocabulary and grammatical structures that are necessary for learning (Fortune, 2008). In synthesizing immersion teachers' experiences, Met (2008) described teachers' constant focus on developing students' vocabulary. In this study, teachers provided precise words to describe situations to improve student understanding and to support students' literacy development (Met, 2008). In these traditional immersion classrooms, researchers have closely examined literacy outcomes, particularly with children.

Primarily, these studies have examined young learners' abilities to achieve educational benchmarks in comparison with their peers (Genesee, 2008). To date, research that investigates outcomes has consistently found that elementary school learners who experience immersion education have demonstrated advantages, scoring equally well or better than their monolingually schooled peers on varied metrics of reading, writing, and other curricular subjects (Genesee, 2008; Lindholm-Leary, 2011; Montanari, 2014; Padilla, Fan, Xu, & Silva, 2013). Further, recent research also demonstrates that students in immersion education programs are successful in attaining L2 literacy over their elementary years in both Italian and Mandarin Chinese (Montanari, 2014; Padilla et al., 2013).

Specifically, with K-3 students in an Italian immersion setting, Montanari (2014) examined students' literacy in both their first and second language simultaneously. Findings

indicate that literacy development continued over time, with all students becoming more accurate in both Italian and English reading as time passed (Montanari, 2014). Similarly, in a Mandarin Chinese immersion school setting, Padilla and colleagues (2013) examined learners' abilities to read in both the Mandarin Chinese and English, the dominant language in larger society, following five years in the immersion program. Individuals were able to read in both languages, gaining literacy in both the L2 and the L1 (Padilla et al., 2013). Together these studies indicate that gaining L2 literacy is possible for young learners in immersion settings across multiple years.

Outside of these immersion school paradigms, immersion education experiences are also part of many study abroad contexts (Isabelli-Garcia, Bown, Plews, & Dewey, 2018). When learners seek similarly immersive experiences at the university through study abroad, they frequently accomplish daily tasks and courses in their L2, learning language through immersion. Yet for American students abroad, these experiences are often short and intensive, lasting only a few weeks (Institute of International Education, 2015). In these study abroad contexts, few studies have been conducted that examine L2 reading. Yet, the findings of one study suggest that learners are able to improve reading outcomes in such intensive, short-term immersion settings (Savage & Hughes, 2014). In this study of L2 reading development in a short-term study abroad program, 140 native English-speaking United States Air Force Cadets learning Mandarin Chinese studied abroad in China for a summer term (Savage & Hughes, 2014). Analyzing performance before and after their summer in China, Savage and Hughes (2014) found that students made notable improvement on reading measures, in addition to gaining confidence in their abilities to use Mandarin Chinese. Briggs (2015) examined vocabulary development with learners studying English abroad in Britain. With these learners who resided in the context for

variable lengths of time, Briggs (2015) determined that informal vocabulary contact used most often by language learners did not facilitate meaningful vocabulary gains. Together, these results demonstrate that it may be possible to gain reading competencies in a short-term immersion experience, yet how specific contexts facilitate language learning may impact overall outcomes. Therefore, there is a need to expand upon current understandings, examining such phenomena across languages or in other immersive contexts.

Domestic immersion programs have provided another popular educational space for learners to gain access to an immersive language learning as these are short-term experiences offering a similar experience to the all-encompassing language environment of immersion education and study abroad while the learner remains in their home country. It appears, at present, that no study has examined L2 reading outcomes of a USA domestic immersion program. However, with American high school students studying French in a 4-week intensive immersion program, oral production has been shown to improve, and learners also appeared to gain a sense of awareness about their own strengths and weaknesses in the language (Dolosic, et al., 2016). Such results indicate that learners do gain language skills through these domestic immersion summer programs.

Further, in another study examining domestic immersion settings, students developed key facets of L2 reading, learning grammatical structures through a domestic immersion university program (Isabelli-Garcia & Lacrote, 2016). In this domestic immersion program with native speakers of English, studying Spanish, researchers examined these university students' learning of specific grammatical structures throughout the 7-week program, comparing accuracy on a written pre-test and a post-test. The tested structure was not included in any explicit instruction. Yet, findings indicated that learners did increase their accuracy as compared to the pre-test

(Isabelli-Garcia & Lacrote, 2016). Taken together, these results from domestic immersion settings are encouraging. However, further examination is needed to understand the benefits and challenges of L2 reading within short-term domestic immersion programs.

4.1.3 Self-Assessment

Based on prior definitions of self-assessment (SA) in L2 contexts (Blanche & Merino, 1989; Brantmeier, 2005a; 2006a; LeBlanc & Painchaud, 1985; Ross, 1998), within this investigation, self-assessment is understood to be learners' evaluations of their own capabilities. Accuracy in such self-assessment is defined as a close association among learners' self-ratings and their performance on a test demonstrating their linguistic capabilities (Butler, 2018; Sweet & Mack, 2017). Such SA has been examined as both an instrument of assessment and a pedagogical tool since the first investigations of this construct. When SA has been discussed in the process of acquiring language, accuracy in SA has often been associated with learners' linguistic gains, focusing on how learners who are successful in self-diagnosing their strengths and weaknesses may have greater capacities for goal-setting and autonomous learning (Sweet & Mack, 2017; Ziegler, 2014).

Prior evaluations and investigations of learners' self-ratings, self-assessment instruments, and self-assessment accuracy have indicated that, much like L2 reading, L2 SA is shaped by a number of variables. Specifically, learners and the instruments that they use to self-assess can impact the accuracy of these self-ratings. For example, L2 SA instruments are more reliable depictions of learners' abilities when the items in the assessment are associated with specific, concrete language tasks and knowledge that the learner can use to situate their response (Brantmeier & Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012; Ross, 1998). Therefore, SA instruments comprised of items that are more descriptive and do not situate the

learner in a specific context of language use do not always reliably correlate with measures of performance (Brantmeier, 2005a; 2006b).

Further, learners' prior experiences may shape their ability to respond to question items, even when the question items offer criteria that situate the learner in a specific language use context (Dolosic, Brantmeier, Strube, & Hogrebe, 2016). In fact, much like other facets of L2 learning, self-assessment has been connected to individual learner differences (IDVs) which are particularly pronounced in language learning in terms of both learners' ultimate successes and means of arriving at those successes (Dornyei & Ryan, 2015). Within self-assessment research, prior findings have indicated that factors such as learners' linguistic proficiency (Heinlenman, 1990), level of anxiety (MacIntyre, Noels, & Clément, 1997), experience using the language in authentic interactions (Dolosic, et al., 2016; Suzuki, 2015), and cultural context of education (Huang, Samuelson, & Chen, 2016) may impact a learners' skills to accurately represent their abilities through self-ratings. In addition, prior studies of SA have almost exclusively relied upon Pearson correlations between learners' SA ratings and performance scores and regression modeling to indicate an overarching relationship for a group of students (Ross, 1998; Brantmeier, et al., 2012). As this same research indicates individual variability in accuracy of self-assessment, these findings about L2 SA underscore a need for individualized approaches. Consequently, the present study examines L2 SA with methodologies that allow for an "individual" outcome of accuracy by incorporating an individualized analytical approach.

4.2 Research Questions

1. To what extent do high school students studying French demonstrate gains in second language linguistic skills and reading abilities following a short-term domestic immersion experience?

2. How do these learners self-assess their second language linguistic skills and reading abilities before and after their domestic immersion experience? Are they able to accurately represent their abilities?

3. When examining the relationship between students' SA and performance, do individual differences emerge? If variability is present, do student characteristics such as proficiency level explain this variability?

4.3 Methods

4.3.1 Participants

Learners in this context are high schoolers aged thirteen to seventeen who attended a four-week immersion summer program. Forty-nine learners and their parents consented to allow the use of their test scores for research purposes, yet forty-seven individuals were included in this analysis as two learners who had consented were unable to complete the program for personal reasons. Learners identified as male or female with 14 self-reporting male gender while 33 reported female gender. Learners indicated their prior study experiences on the entry questionnaire, reporting zero to ten years of French study with an average of 2.5 years of prior study of French (with a standard deviation of 1.8 years). Therefore, learners varied in their prior knowledge and experiences with the French language.

4.3.2 Context

The unique language learning context for this study is referred to as "domestic immersion" where learners come to an educational space wherein the L2 is the medium of

communication without having left their home country. Specifically, this program fosters a supportive educational environment where learners are engaged in active learning as community and in nature, all in the L2. This total immersion includes meal times, independent study, song sessions, and varied activities where learners put their French language to use both with learners at the same stage of acquisition and in mixed groups all under the leadership of native and native-like speakers of French (Hamilton & Cohen, 2004). Upon arrival, learners are split into groups based on their individual stages of acquisition. These levels are determined by their performance on a paper and spoken placement test that includes measures of listening, speaking, reading, writing, and grammar and vocabulary knowledge in French. Learning groups consist of less than fifteen students each and are seen as "classes." Each section has a well-qualified French instructor who meets with the class regularly for 2.5 hours of class daily. In addition, students experience total-immersion courses about the culture of the Francophone world that are taught by another instructor for one hour each day. Beyond class sessions, throughout the day each individual student has many opportunities to engage with French and learn from the context without the distraction of cell phones, television, or online media, which are forbidden in this context (Hamilton & Cohen, 2004). For example, learners are able to dance to French music with friends, buy items at the village store, or play outdoor games led by a counselor, using their French language skills to communicate with those around them. At the end of each day, learners have about thirty minutes of English-led discussion to ask questions of their counselors and talk about the best and most challenging parts of their day. This moment in the day is meant to allow learners at all levels of French to have an opportunity to connect with peers and counselors as language learners striving to learn more French.

4.3.3 Instruments

The assessments used in this study include a demographic questionnaire, self-assessment instrument, and vocabulary, grammar, and reading performance tasks. These represent a portion of the assessments completed for placement and assessment purposes. (Examples of items can be found in Appendix B.)

4.3.3.1 Demographic Questionnaire & Self-Assessment. The demographic questionnaire was used to collect basic information about each participant, consisting of questions about the length of time and typical learning context for the students' French classes prior to arrival at the camp. For the self-assessment, students completed an instrument that is a criterion-referenced self-assessment developed based on items and underlying principles of prior research in self-assessment (Brantmeier & Vanderplank, 2008; Brantmeier et al, 2012; DIALANG Project, https://dialangweb.lancaster.ac.uk/; Little, 2005; Zhang & Thompson, 2004). In total, this instrument consisted of forty-eight items situating the learner in a specific context where the learner then rated their agreement with the statement. High agreement, with a rating of "5" on a Likert scale, indicated that they believed that they were able to successfully complete the linguistic skill while disagreement, with a rating of "1" on a Likert scale, indicated that they thought they would not be able to complete the skill stated.

4.3.3.2 Vocabulary & Grammar. The vocabulary instrument was developed based on a Levels Test (Nation, 2000). This test was constructed of matching items. Students were asked to match more common vocabulary to less common vocabulary, demonstrating their comfort with more or less frequent words in French. Then, students were asked to produce the equivalent word in English to further demonstrate their understanding. Words for this task were selected from a frequency list and verified by a second reader. The two-part section focused on grammar was comprised of a production and a recognition task. The production task was a fill-in-the-blank

task, where the correct grammatical form was requested to complete the sentence. The second portion was a multiple-choice format where the correct form must be selected from four options. As this test was used for placement, these grammar questions were designed to elicit forms across typical instructional sequences, soliciting basic forms and advanced forms to discriminate which of these learners had encountered previously.

4.3.3.3 Reading. The reading section included two passages that learners read and used to complete tasks. Each passage was followed by three tasks: Free Recall, Sentence Completion, and Multiple Choice. Moving from less to more cued response, these tasks elicit information about the learners' understanding of the passages read. All passages were equated for length in number of words, pausal units, and grammatical complexity through an examination of embedded clauses. Further, these passages are all considered "Advanced" by the American Council of Teachers of Foreign Languages (ACTFL) Guidelines. All tasks testing the comprehension of the passage were conducted in English so as to be true tests of reading skill rather than writing skill, based on the guidelines outlined by Bernhardt (1991) and Wolf (1993) and furthered by Brantmeier (2006b).

4.3.4 Procedures

4.3.4.1 Test Taking. Learners completed equivalent versions of these measures both at the beginning and at the end of their experiences in the immersion setting. They completed the pre-test in sections, filling out the demographic questionnaire and self-assessment during their "check-in" to the program. Then, on their first full day in the setting, they completed a researcher-designed placement test, including these measures of French vocabulary, grammar, and reading comprehension. They progressed through all tasks in a linear fashion, completing a page and then turning to the next without the opportunity to turn back. After four weeks in the

language learning setting, learners completed the post-program test which included the measures of self-assessment and French vocabulary, grammar, and reading comprehension. Again, they moved through the packet without being able to return to previous pages.

4.3.4.2 Scoring. When examining the data, two raters independently scored all sections. Then, raters met together to discuss disagreements until an agreement was reached for each response on each learner's packet. Ultimate rater agreement was determined to be greater than 90%. Vocabulary items were scored correctly when (1) matched correctly and (2) defined correctly, as determined by the raters. Grammar fill-in-the-blank items were said to be correct when they were perfectly correct, including accents and spelling. Free recall was scored using pausal units (Brantmeier , Strube, & Yu, 2014). Such pausal units are defined as the natural pauses that a native speaker takes when reading a given passage aloud (Brantmeier, et al., 2014). Sentence completion was scored based on acceptable answers that responded to the questions with knowledge from the passage. These acceptable answers were agreed upon by both raters. Multiple-choice items for both grammar and reading comprehension were correct when the correct answer, as determined by a native reader, was selected. Both multiple-choice items and sentence-completion items were based on the main idea and key details presented within the passage.

4.3.5 Analyses

Analyses for this examination of data included typical univariate statistics for the field of applied linguistics (Plonsky, 2015), bootstrapping, and hierarchical approaches to seek direct answers to the research questions presented. No data were excluded due to missingness because data were found to be missing at random. Further, no outliers were excluded as no individuals had scores that were greater than 2.5 standard deviations from the mean. All analyses were

conducted with the open-source statistical program, R version 3.3.3 (R Core Team, 2017), using additional packages including psych (Revelle, 2017), psychometric (Fletcher, 2010), lme4 (Bates, Machler, Bolker, & Walker, 2015), and car (Fox & Weisberg, 2011). Figures were created using GGPlot2 (Wickham, 2009).

4.4 Results

All data were first visually examined and tested to understand the means and variability in the data. Descriptive statistics demonstrating these facets of the data are presented in Table 4.1.

Table 4.1. *Descriptive Statistics of Key Variables*

		Minimum	Maximum	Mean	Standard Deviation
Reading SA	*Pre-Test*	1.00	5.00	3.59	1.08
	Post-Test	3.43	5.00	4.33	0.45
Language Knowledge SA	*Pre-Test*	1.00	4.57	2.94	0.89
	Post-Test	2.43	4.93	3.64	0.66
Reading Score	*Pre-Test*	0.07	0.80	0.42	0.17
	Post-Test	0.17	0.83	0.48	0.15
Language Knowledge Score	*Pre-Test*	0.02	0.72	0.29	0.22
	Post-Test	0.12	0.92	0.54	0.22

Note: Reading and Language Knowledge Scores are percentages of possible correct answers. Self-Assessment is on a scale of one to five.

RQ1: *To what extent do high school students studying French demonstrate gains in second language linguistic skills and reading abilities following a short-term domestic immersion experience?*

To answer this research question, paired t-tests were indicated. In this comparison, students' pre-test scores were compared to their own post-test scores. Findings indicated that for

both language knowledge and reading comprehension post-test scores were significantly different from pre-test scores ($p < 0.05$). This difference was positive. Examining these results further, effect sizes indicated that learners demonstrated larger differences between pre-test and post-test for language knowledge scores than between pre-test and post-test for reading comprehension scores (with Cohen's d equaling 0.594 and 0.257 respectively). Results are summarized in Table 4.2 and presented visually in Figure 4.1 and Figure 4.2. Figures 4.1 and 4.2 are boxplots that show the entire distribution of scores, with the median score highlighted at the center and standard deviations represented by the box and the line segments, commonly referred to as "whiskers." Due to the non-normality of the distribution of language knowledge scores, these findings were also verified using bootstrapping methods. These secondary tests indicated that the t-test results were correct, with significant differences remaining evident ($p < 0.01$).

Table 4.2. *Results of Paired t-tests of Reading Comprehension and Language Knowledge*

	t	Degrees of Freedom	p-value	Cohen's d
Language Knowledge Scores	16.93	46	< 0.0001	0.594
Reading Comprehension Scores	3.3641	46	0.0016	0.257

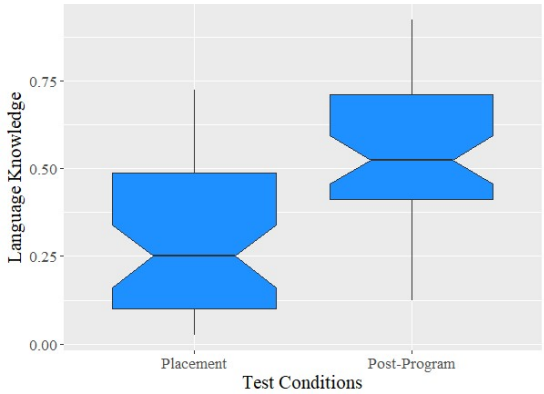

Figure 4.1. *Comparison of Language Knowledge Scores*

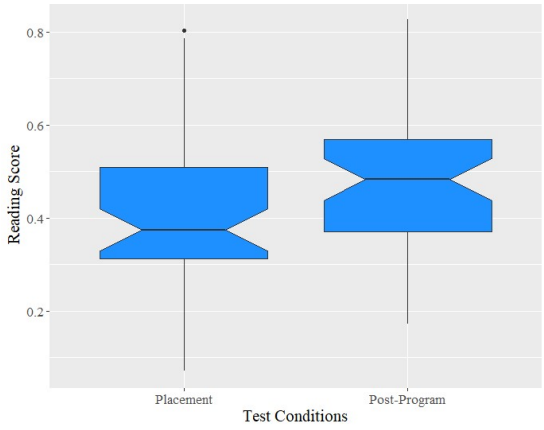

Figure 4.2. *Comparison of Reading Comprehension Scores*

RQ2: *How do these learners self-assess their second language linguistic skills and reading abilities before and after their domestic immersion experience? Are they able to accurately represent their abilities?*

Self-assessment score means are presented in Table 4.3. As these means indicate, students' self-assessments fall near the center of the scale, particularly on the pre-test where students often rank themselves near 3 or "Neutral." Yet, pre-test self-assessment is lower on average than post-test self-assessment with statistically significant differences between how learners rated themselves before and after the program ($p < 0.05$). Results of this test, including effect sizes, are presented in Table 4.4. In addition to these self-assessment trends, learners also rated themselves more highly on the applied skill of reading than on their discreet knowledge of linguistic features both on the pre-test and the post-test.

In examining the relationship between learners' self-assessment and their performance, traditional approaches were applied first. Pearson correlations revealed that learners were able to accurately self-assess both before and after the program on dimensions of both language knowledge and reading comprehension ($p < 0.05$). Table 4.5 presents these results. Figures 4.3 and 4.4 represent these findings.

Table 4.3. *Self-Assessment Score Means by Skill and Test Condition*

	Reading SA	Language Knowledge SA
Pre-Test Mean (SD)	3.59 (1.08)	2.94 (0.89)
Post-Test Mean (SD)	4.33 (0.45)	3.64 (0.66)

Table 4.4 *Results of Paired t-tests of Reading Comprehension and Language Knowledge Self-Assessment*

	t	Degrees of Freedom	p-value	Cohen's d
Language Knowledge SA	8.37	45	< 0.0001	0.891
Reading Comprehension SA	5.52	45	< 0.0001	0.897

Table 4.5. *Correlations among Self-Assessed Values and Performance Values*

	r	p-value
Language Knowledge Pre-Test	0.63	< 0.01
Language Knowledge Post-Test	0.73	< 0.01
Reading Comprehension Pre-Test	0.48	0.01
Reading Comprehension Post-Test	0.45	0.01

Further tests were conducted to examine the differences in these correlations among self-assessment and both language knowledge and reading comprehension because language knowledge correlations are larger than reading comprehension correlations both before and after the immersion program. Using a Fisher's Z statistical test, statistically significant differences were found among only the post-test correlations, with learners having a stronger association

among self-assessed values of language knowledge and language knowledge performance ($p < 0.05$) (Meng, Rosenthal, & Rubin, 1992).

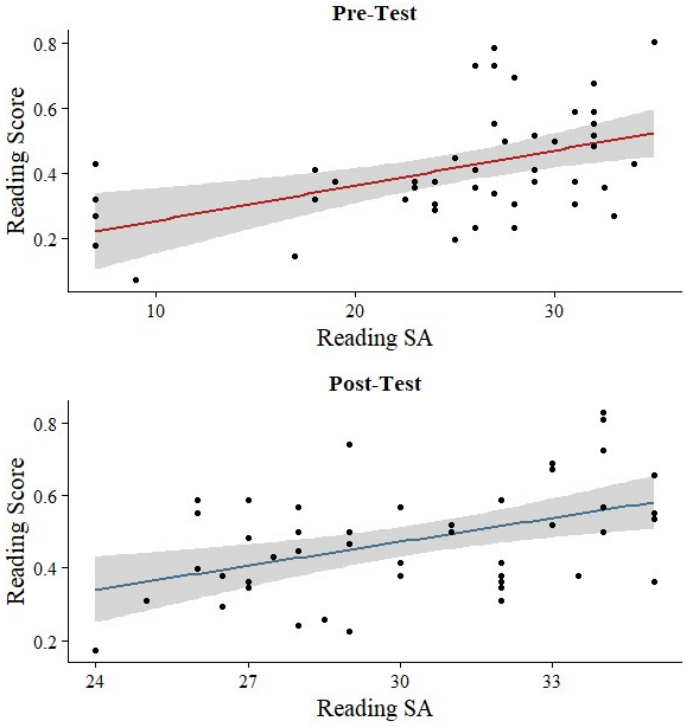

Figure 4.3. *Relationship Among Self-Assessed Reading and Performance Scores*

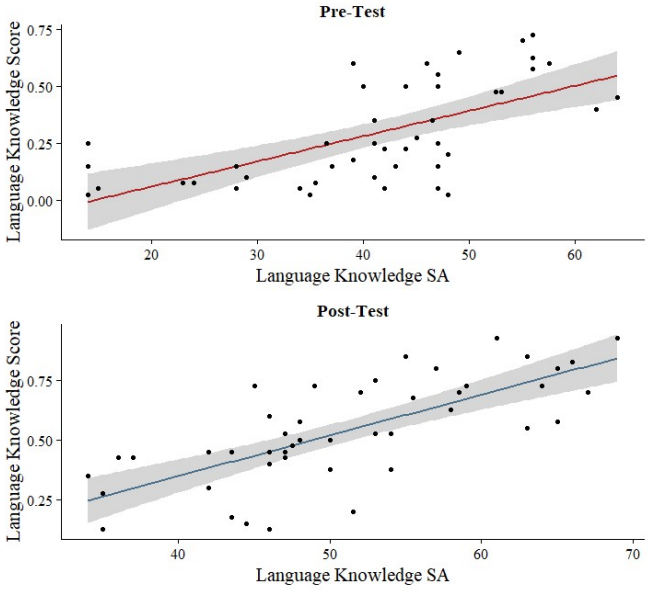

Figure 4.4 *Relationship Among Self-Assessed Language Knowledge and Performance Scores*

RQ3: *When examining the relationship between students' SA and performance, do individual differences emerge? If variability is present, do student characteristics such as proficiency level explain this variability?*

In order to examine this final research question, the present investigation employed hierarchical linear modeling (HLM) to elucidate individualized measures of self-assessment in order to seek a better understanding of individuals' capabilities to self-assess with accuracy. This flexible type of analysis provides opportunities to analyze data at multiple levels, separating the within-group and between-group variability (Kashy, Campbell, & Harris, 2006; Snijders & Bosker, 2012). In this analysis, each "group" was an individual learner who completed multiple measures of self-assessment and reading comprehension. These measures were said to be "nested" within the individual. Using a dummy coded no-intercept model, slopes and intercepts were able to be calculated for each individual in the study. Results of this model are presented in Table 4.6. Comparing this model with a null model, the No-Intercept Dummy-Coded Model of Self-Assessment and Performance was preferred ($\chi^2 = 17.91$, $p < 0.05$). Much like the slopes presented in Figures 4.3 and 4.4, the random slopes in this model are said to represent the association among self-assessment ratings and performance (Snijders & Bosker, 2012). However, these slopes and intercepts vary for each individual within the sample. Further, within the model, one set of slopes and intercepts represents the pre-test while another represents the post test. (For a more detailed description of the methods and formula used to calculate this analysis, see Appendix C.)

The results of this model further substantiate the average relationship for self-assessment with significant slopes for both Pre-Test and Post-Test self-assessment. Using this HLM analysis, comparisons between pre-test and post-test self-assessment accuracy slopes and

Table 4.6 *Results of No-Intercept, Dummy Coded Model: Self-Assessment and Performance Measures*

	Fixed Effects	
	β (SD)	*t*
Pre-Test	0.02 (0.11)	0.18
Post-Test	-0.01 (0.11)	-0.08
Pre-Test Self-Assessment	0.25 (0.09)	2.87
Post-Test Self-Assessment	0.19 (0.06)	2.98
	Random Effects	
	Variance (SD)	
Residual	0.39 (0.63)	
Pre-Test	0.43 (0.66)	
Post-Test	0.43 (0.66)	
Pre-Test Self-Assessment	0.06 (0.14)	
Post-Test Self-Assessment	0.02 (0.63)	

intercepts were also examined. These examinations were undertaken to discover possible changes in self-assessment accuracy while learners engaged with this unique context of learning. The difference in variance among the Pre-Test Self-Assessment and the Post-Test self-assessment does suggest some effect of the immersion program. However, with these self-assessment and reading comprehension measures, no significant differences were indicated in these comparisons of pre-test and post-test relationships ($p > 0.05$), meaning that there were no significant differences found between pre-test and post-test slopes and intercepts representing the relationship among self-assessment and performance measures.

While other models often assume that remaining variance is "error," HLM analysis also allows for the introduction of predictors to explain systematic variance in the slopes and

Table 4.7 *Results of No-Intercept Dummy-Coded Model of Self-Assessment, Level, and Performance*

	Fixed Effects	
	β (SD)	*t*
Pre-Test	-0.01 (0.06)	-0.17
Post-Test	-0.04 (0.73)	-0.59
Pre-Test Self-Assessment	0.09 (0.07)	1.47
Post-Test Self-Assessment	0.09 (0.06)	1.61
Pre-Test: Level	0.69 (0.06)	10.70
Post-Test: Level	0.60 (0.07)	8.11
Pre-Test: SA: Level	0.02 (0.06)	0.41
Post-Test: SA: Level	0.08 (0.06)	1.38
	Random Effects	
	Variance (SD)	
Residual	0.39 (0.63)	
Pre-Test	0.06 (0.25)	
Post-Test	0.13 (0.36)	
Pre-Test Self-Assessment	0.07 (0.10)	
Post-Test Self-Assessment	0.00 (0.00)	

intercepts (Snijders & Bosker, 2012). Therefore, as this study examines learners across proficiency levels, the predictor of "Level" was added to the model. This predictor was the ACTFL proficiency of students, as rated by the team of instructors in order to confer grades for the high school immersion program. Results of this model with Level as predictor are presented in Table 4.7. When added to the model as a predictor, this model was considered to be an improvement over the previous models (as is evident in Table 4.8). Following these initial examinations, comparisons among pre-test and post-test coefficients were reexamined within the

new model. Through this individual analysis, however, no significant differences were evident, indicating no change in self-assessment accuracy. However, the contribution of Level accounted for individual variability and allowed individual differences to emerge.

Table 4.8 *Comparison of Models*

	AIC	*Deviance*	X^2	*P*
Null	1278.2	1268.2		
Self-Assessment Only	1268.3	1250.3	17.91	0.001
Self-Assessment & Level	1162.5	1136.5	113.74	<0.001

4.5 Discussion

Results of this study clearly demonstrate that these learners gained capacities in their French vocabulary and grammar knowledge and French reading comprehension. While no prior studies have examined these outcomes with adolescents in a domestic immersion setting, the growth in reading comprehension abilities following a short-term immersion experience aligns with the findings of Savage and Hughes (2014). These researchers found that university-aged learners of Chinese were able to develop their reading capacities during a short stay abroad in China, as compared to their abilities prior to the program. In addition, the present study's findings of grammar and vocabulary knowledge growth are also in line with outcomes of Isabelli-Garcia and Lacorte (2016) that demonstrated that university-level learners of Spanish were able to gain language knowledge even beyond their language classes in domestic immersion contexts. Further, as Briggs (2015) called for in her study, more structured vocabulary learning in this context does appear to align with evident vocabulary gains seen on these language knowledge tasks. As both the language knowledge and the reading comprehension

scores demonstrate growth, these findings connect to the findings and discussion of Brantmeier and colleagues (2012) where they assert that language skills are interconnected "facets of a single higher order construct" (p. 152) and highlight the co-dependent nature of varied skills that make up second language acquisition.

However, findings of the present study are also complex in that gains in language knowledge were more evident that gains in reading comprehension abilities. These results align with current understandings of the numerous variables that shape a reader's success in their L2 reading (Bernhardt, 2011; Grabe, 2009; Koda, 2005). Specifically, this study highlights that while language knowledge and reading comprehension are correlated, many more factors shape successful reading of discourse-level texts. For example, within this study, L2 language knowledge was demonstrated through tasks involving one or two words in French while reading comprehension was measured with texts of more than 200 words. Within these larger texts, various factors both within and beyond the explained variance of the Compensatory Model (Bernhardt, 2011) may have impacted reading comprehension scores, shifting our understanding of "gains" made within the learning experience. Therefore, the complexity of discourse-level reading comprehension may have obscured specific strategic and linguistic gains made throughout this short-term immersion program. Taken together, the results of the present study are a nuanced depiction of the development of linguistic proficiency across skills, but they also contribute to our understanding of this domestic immersion context through clear gains in student learning. Situating these results within the Compensatory Model, it is evident that these relationships further support the current understandings of "explained variance," as linguistic knowledge and L2 reading comprehension appear to be closely tied together (Bernhardt, 2011).

Yet, further work remains to fully understand the variables that shape the remaining "unexplained" variance.

Results of this study also indicated that students demonstrated accuracy in their self-assessments of language knowledge and reading comprehension both before and after their experiences in the four-week domestic immersion program. These findings align with prior research that indicates greater accuracy in self-assessment when instruments are comprised of criterion-referenced items that situate the learner in a specific context of language use (Brantmeier & Vanderplank, 2008; Brantmeier, et al., 2012; Ross, 1998). However, prior research on oral production has also indicated that adolescent learners' self-assessment was not always accurate in these domestic immersion settings (Dolosic, et al., 2016). More specifically, accuracy for oral production self-assessment was said to improve following increased language use experiences during domestic immersion (Dolosic, et al., 2016). However, the results of the present study do not demonstrate this pattern. Rather, learners are able to self-assess both before and after the immersive experiences of this program. This divergence from prior findings may rely on the linguistic skills being self-assessed before and after this program. More precisely, when high school learners study languages in their schools and homes, they are able to interact with authentic texts and interpret various written media far more readily than they are able to interact with fluent, interactive speakers and media (Isabelli-Garcia, Bown, Plews, & Dewey, 2018). Therefore, it is possible that participants in the present study were able to more accurately self-assess their L2 language knowledge and reading comprehension from the beginning of the program due to the many experiences they already possessed with text-based mediums in the L2 from their home and school learning environments. In this way, it could be postulated that these

learners were able to apply their experiences when self-assessing both before and after this immersive program.

When examining self-assessment for individual differences, relationships between students' self-ratings and performance indicated a variability with some learners being more accurate in their depictions of themselves than others. Such findings build upon reviews and analysis of self-assessment and its association with performance (Ross, 1998; Brantmeier, et al., 2012; Dolosic, 2018). Within these SA studies, it is evident that learners do not all share identical capacities to self-assess with some being represented by data points that deviate greatly from trend lines in correlational tests. Yet, students are treated as a group in most analyses, making individualized understandings of accuracy in self-assessment difficult to determine empirically. With this hierarchical analysis representing the accuracy of individuals' self-assessments, it is possible to measure this variability and determine how accuracy in self-assessment may be related to other facets of L2 language learning, such as IDVs (Dornyei & Ryan, 2015), and the complex nature of L2 reading comprehension (Bernhardt, 2011). While prior research has laid the foundation for these findings, it is for future research to determine our broad understandings of self-assessment accuracy with idiographic and longitudinal methodologies.

In implementing this new approach within the study, results indicate that self-assessment accuracy does have a relationship with the proficiency level of the individual learner. Specifically, when accounting for the level of learners, results showed that there was improvement in the model accounting for the variance in individuals' accuracy in their self-assessments. Such findings align with Heinlenman's (1990) study of accuracy in L2 self-assessment wherein learners proficiency level impacted the accuracy of their self-assessments.

Brantmeier and colleagues (2012) also further this understanding, indicating that more these more consistent, advanced learners were more accurate in their use of the criterion-referenced self-assessment. The present study further confirms these findings for adolescents learning French in an immersive context. For this study, HLM analysis ended with this model. In the future, HLM analyses should be conducted with larger samples and in longitudinal designs. In this way, researchers will be able to use this flexible framework to begin to unravel the underlying complexities of self-assessment and skills development, particularly in unique settings such as domestic immersion.

4.6 Conclusion

This study indicates that learners gained French language capabilities in terms of vocabulary and grammar knowledge as well as reading comprehension within this unique language learning context. Further, these students demonstrated capabilities to self-assess with accuracy both before and after their immersion experiences, indicating nuances to students' abilities to accurately self-assess across skills. In addition, examining accuracy in self-assessment led to findings that suggest that such self-assessment accuracy could be associated with the proficiency of the learner. Future research should expand on these findings, seeking to understand and underscore the benefits of domestic immersion for language learning. In the future, studies should also further explore these intriguing findings of self-assessment accuracy, examining how self-assessment precision may be a key component toward the development of lifelong language learners.

Chapter 5: Conclusion

The studies within this dissertation (1) examine self-assessment through three distinct methodological approaches across three languages and (2) analyze the many intersecting variables that impact both L2 self-assessment and L2 reading comprehension. Therefore, this research has attempted to broaden understanding of how, when, and where learners are able to accurately represent their own strengths and weaknesses, particularly with L2 reading comprehension. These studies respond particularly to the growing importance of (1) skilled L2 and FL reading across a variety of text types in the 21st century (Bernhardt, 2011; Brantmeier, 2009; Grabe, 2009; Koda, 2005) and (2) learner-centered teaching and learning of L2 and FL (Butler, 2018; Rivers, 2001; Yamagata, 2018; Ziegler, 2014).

In 2007, the Modern Language Association released a report calling for increased language teaching in "applied" capacities, encouraging language programs to provide learners with opportunities to gain skills valuable to their personal and professional ambitions (MLA Report, 2007). Within this dissertation, such applied skill learning has been examined in varied contexts of L2 reading comprehension, responding to a need to better understand this skill which has only grown in importance as learners engage with online and print media in their L2s (Grabe, 2009). In order to better prepare lifelong L2 and FL readers, the present studies examined learners' capabilities to read across text types using multiple methods of assessment, expanding our understandings of reading comprehension with adult and adolescent readers of a second language. Using the Compensatory Model (Bernhardt, 2011) as a theoretical underpinning for research design and interpretation of results, this dissertation examined facets of L2 reading processes and products that lie within both the explained and unexplained variance of this model. For example, in *An Individualized Approach to Self-Assessment with Readers in the French*

Village, L2 language knowledge, a construct considered to be part of the "explained" variance of L2 reading was examined, and in *An Examination of Self-Assessment and Interconnected Facets of Second Language Reading*, the "unexplained" variance associated with background knowledge, text type, and test method was investigated more closely. Results of these studies have therefore contributed to our understanding of this Compensatory Model of L2 reading (Bernhardt, 2005; 2011).

Further, as researchers and educators alike seek to establish learner-centered classrooms where autonomous learners use their agency to individualize their learning and meet their own language learning goals (Butler, 2018; Rivers, 2001; Yamagata, 2018; Ziegler, 2014), much remains to be understood about the possible power of self-assessment training in supporting learners as they develop these capacities. While prior research in self-assessment training in classroom environments has demonstrated that learners who are trained in self-assessment are able to self-assess with accuracy and often perform better than their untrained peers, these studies have focused almost exclusively on productive abilities (Huang, Samuelson, & Chen, 2016; Malzloomi & Khabiri, 2016; Nguyen & Gu, 2013; Poehner, 2012). Therefore, there is a clear need to further examine self-assessment training, seeking to understand its benefits within the larger understanding of the interconnected "facets of [the] single higher order construct" of second language acquisition, including L2 reading comprehension (p. 152, Brantmeier et al., 2012).

As some researchers have examined self-assessment in classroom settings, others have investigated self-assessment as an assessment tool in order to find the relationships that predict quality self-assessment instrumentation and analyses. Seeking means to place "borderline" learners, Brantmeier (2006a) and colleagues (2008; 2012) examined the self-assessment

construct. Results of their studies aligned with prior findings, such as the comprehensive meta-analysis of Ross (1998) and forward-thinking work of Little (2005), which indicated that self-assessment items should situate the learner in concrete and specific language use situations. It is evident that much of the prior self-assessment work relies upon Pearson correlations among a self-assessment instrument and a more traditional assessment of language performance or proficiency, with few using new methodologies. Yet, as the studies within this dissertation highlight where individual variability is also examined, explanations about "why" learners vary so greatly in their abilities to self-assess emerge. As a result, this dissertation has examined self-assessment through three unique studies in order to expand upon the pioneering work that has been completed in L2 self-assessment.

5.1 Brief Review of Findings

Through quantitative designs and analyses, this dissertation has examined L2 reading comprehension and self-assessment, contributing to the growing understanding of these two complex systems. Within each study, different findings led to advances in our understanding of these constructs in different ways. More precisely, the first study of this dissertation, focusing on Chinese university students studying English, findings suggest that (1) learners performed differently on L2 reading comprehension depending on both the measure used and the given text's structure, and (2) learners were able to accurately depict their strengths and weaknesses on a criterion-reference instrument of self-assessment, particularly with narrative text types and multiple-choice formats. Within this study, these findings were taken as evidence for learners' familiarity with specific texts and tasks as beneficial both for performance and self-assessment with their L2, English.

On another front, results from the second study demonstrate that university students studying Spanish in the US at the advanced level are able to accurately depict their capabilities with advanced Spanish reading after complete three sessions of online training. Yet, these same learners did not outscore their peers on L2 reading comprehension, demonstrating a significant difference only on Sentence Completion tasks with better performance in the control group than the trained group. This study also found that these learners used these online training sessions as an opportunity to evaluate their approaches to Spanish reading and their broader study of Spanish. For this study, the findings were taken to indicate that there are benefits such as developing abilities to self-evaluate and self-monitor, from this online intervention. However, results also indicate that this short-term intervention may not have been enough to enhance L2 reading comprehension and lead to substantive changes in learners' capabilities to read Spanish at the advanced level.

Finally, with adolescents in the language immersion village, findings indicated that these learners (1) gained capacities in both language knowledge and reading comprehension and (2) were able to accurately self-assess their language knowledge and reading comprehension in French both before and after their time at the language village, with little change in their capabilities before and after their time in the context. Using HLM frameworks, this study was also able to provide an individualized look at learners' abilities to self-assess demonstrating that there was variability among learners in their capacities to self-assess. Further, when expanding this model to include levels of proficiency of the learner, some of this variability in capabilities to self-assess was explained. As prior research had not explored gains in L2 language knowledge and reading comprehension, these results contribute to our understanding of possible gains in

such domestic immersion contexts. In addition, these findings, in light of prior work, contribute to our larger understanding of L2 self-assessment and reading comprehension.

5.2 Overarching Discussion

A consistent finding across these three experiments is the variability present in learners' reading comprehension abilities. This variability in performance scores across measures was evident even within contexts where learners were placed into the same class or level of study, demonstrating that more factors than proficiency are at play when learners are reading in a second language (Bernhardt, 2011; Brantmeier, et al., 2012; Grabe, 2009; Koda, 2005). Further, within these studies there were also clear differences on reading comprehension performance when measured by different tasks. Specifically, learners tended to be more successful on tasks such as sentence completion or multiple choice. Findings of this study, therefore, further validate that task or test method truly has an effect on learners' demonstrated L2 reading comprehension outcomes (Alderson, Bachman, Perkins, & Cohen, 1991; Alderson, Clapham, & Wall, 1995; Riley & Lee, 1996; Brantmeier, 2006b; Carrell, 1984a, 1984b, 1985; Shohamy, 1984). Consequently, this finding aligns with comprehensive understandings of reading comprehension assessment as Alderson (2000) indicates that more supportive cues, particularly when selecting the correct answer from a number of plausible distractors usually yield higher performance scores. In addition, this finding may align with the theory of Transfer Appropriate Processing (TAP) more commonly used in L1 psychological studies of learning (Morris, Bransford, & Franks, 1977). According to TAP, learners use their prior experiences with similar tasks to support their success (Morris, et al., 1977).

The second study of this dissertation also clearly demonstrates that L2 self-assessment training can benefit capabilities in self-assessment of L2 reading comprehension. Yet, when

examining these three investigations together, it is evident that learners studying English in China and those studying French in the immersion summer camp did not need this same training in order to accurately represent their abilities. Rather, for these learners their self-assessed ratings were significantly correlated with their actual performance on the measures without any specialized training in self-assessment. These findings align with prior studies of self-assessment instrumentation that indicate that untrained learners are able to accurately self-assess when the items are concrete, or criterion-referenced (Brantmeier & Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012; Little, 2005; Ross, 1998). Yet, the findings of *Chapter 3*: *Interactive Online Self-Assessment Training and Advanced Second Language Reading of Spanish at the University* do not align with these studies as learners were only able to accurately self-assess when they had been trained. However, findings like these do fit into our larger understandings of self-assessment as a complex, multi-faceted construct. For example, learner features such as anxiety (MacIntyre, Noels, & Clément, 1997), experiences (Dolosic et al., 2016; Suzuki, 2015), or proficiency (Heinlenman, 1990) may impact the accuracy of self-assessment. Therefore, these findings lay a foundation of future work uniting these variables in multivariate analyses in order to understand their effects comprehensively.

Further, across all three studies, self-assessment ratings and accuracy vary widely among individuals, with some able to accurately represent their abilities while others are distant from their expected values. Most saliently, in the *Chapter 4: An Individualized Approach to Self-Assessment with Readers in the French Village,* individual slopes and intercepts demonstrate that learners differed in their self-assessments. Such individuality aligns with prior findings of self-assessment which demonstrate that there is variability in the success of individuals to depict their own strengths and weaknesses (Brantmeier, et al., 2012; Ross, 1998).

In addition, findings of this dissertation suggest that there is a test method effect for the accuracy of self-assessment. Such findings align with the results of Brantmeier (2005a; 2006b) wherein learners were able to accurately depict their abilities in reference to some of the testing conditions but not all conditions. However, in this dissertation learners are most accurate on tasks that they are familiar with. These findings align more closely with Dolosic and colleagues (2016) and Suzuki (2015) who found that experience shaped learners' abilities to accurately self-assess with speaking. In the case of these reading comprehension self-assessments, such experience could be familiarity with a specific testing task, such as multiple choice. Again, such an interpretation would apply the L1 TAP model where learners are able to transfer similar experiences when answering (Morris et al., 1977). In order to fully develop the understanding of such transfer in L2 settings, further studies are needed.

5.3 Future Directions

Building from the foundation of this dissertation, future studies should examine the multiple factors that shape L2 reading comprehension, particularly investigating the unexplained variance within the Compensatory Model (Bernhardt, 2011). For example, studies should examine L2 reading comprehension alongside variables said to shape the processes and products of L2 reading including topic familiarity, working memory capacity, anxiety, and testing methodology. Other facets such as metalinguistic knowledge and its contribution should also be explored. Then these studies should use multivariate techniques to uncover underlying relationships among these supportive or inhibitive impacts on ultimate comprehension. In addition, future work should study self-assessment training with L2 reading comprehension at all stages of acquisition and with varied lengths of intervention to find the possible benefits of self-assessment training. These interventions could particularly focus on interventions in beginning

and intermediate courses to develop abilities to self-assess L2 reading with accuracy in programs where literature is at the core of the major. Finally, there is a need for the use longitudinal designs to expand current understandings of the development of both L2 reading comprehension and self-assessment skills with lifelong language learners. Specifically, researchers should implement initiatives that follow learners both within courses and throughout a program of study so that progress in L2 reading comprehension can be better understood. All of this future work should be particularly taken up with adolescent and adult learners as they seek to use their L2 at advanced levels in personal and professional arenas that cross linguistic and national boundaries. Through these three lines of inquiry, future research will be able to better describe and understand L2 self-assessment and reading comprehension, providing the possibility of using self-assessment to its fullest and supporting successful second language reading at the highest levels.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Alderson, N. J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41-66.

Ashton, K. (2014). Using self-assessment to compare learners' reading proficiency in a multilingual assessment framework. *System, 42*, 105-119.

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

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14-29.

Baker, W. (2015). Culture and complexity through English as a lingua franca: Rethinking competences and pedagogy in ELT. *Journal of English as a Lingua Franca, 4*(1), 9-30.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual review of psychology, 52*(1), 1-26.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning, 52*(2), 323-363.

Barcroft, J. (2003). Effects of questions about word meaning during L2 Spanish lexical learning. *The Modern Language Journal, 87*(4), 546-561.

Barcroft, J. (2004). Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals, 37*(2), 200-208.

Barcroft, J. (2013). Input-based incremental vocabulary instruction for the L2 classroom. In Schwieter (Ed.) *Innovative research and practices in second language acquisition and bilingualism*, 107-138. Amsterdam, Netherlands: John Benjamins.

Barry, S., & Lazarte, A. A. (1998). Evidence for mental models: How do prior knowledge, syntactic complexity, and reading topic affect inference generation in a recall task for nonnative readers of Spanish?. *The Modern Language Journal, 82*(2), 176-193.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4} [Computer Software]. *Journal of Statistical Software, 67*(1), 1-48.

Bernhardt, E. (1991). *Reading Development In A Second Language: Theoretical, Empirical, And Classroom Perspectives*. New York, NY: Routledge.

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25,* 133-150.

Bernhardt, E. (2011) *Understanding advanced second-language reading*. New York, NY: Routledge.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417-444.

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39*(3), 313-338.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211-252.

Brantmeier, C. (2004). Building a comprehensive theory of adult foreign language reading: A variety of variables and research methods. *The Southern Journal of Linguistics, 27*(1), 1-7.

Brantmeier, C. (2005a). Nonlinguistic variables in advanced L2 reading: Learner's self-assessment and enjoyment. *Foreign Language Annals, 38*(4), 493-503.

Brantmeier, C. (2005b). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension. *The Modern Language Journal, 89*(1), 37-53.

Brantmeier, C. (2006a). Advanced L2 learners and reading placement: Self-assessment, computer-based testing, and subsequent performance. *System, 34*(1), 15-35.

Brantmeier, C. (2006b). The effects of language of assessment and L2 reading performance on advanced reader's recall. *The Reading Matrix: An International Online Journal, 6*(1), 1-17.

Brantmeier, C. (2006c). Toward a multicomponent model of interest and second language reading: Sources of interest, perceived situational interest, and comprehension. *Reading in a Foreign Language, 18*(2), 89-115.

Brantmeier, C. (2013). Acquisition of reading in second language Spanish. In Geeslin, K. (Ed.). *Acquisition in the Spanish Classroom* (pp. 466-481). Hoboken, NJ: Wiley-Blackwell.

Brantmeier, C., & Dragiyski, B. (2009). Toward a dependable measure of metacognitive reading strategies with advanced L2 learners. In Brantmeier (Ed.) *Crossing languages and research methods: Analyses of adult foreign language reading* (pp. 47-72). Charlotte, NC: Information Age Publishing.

Brantmeier, C., Strube, M. & Yu, X. (2014). Scoring recalls for L2 readers of English in China: Pausal or idea units. *Reading in a Foreign Language, 26*(1), 114-130.

Brantmeier, C. & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System, 36*, 456-477.

Brantmeier, C., Vanderplank, R. & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System, 40*(1), 144-160.

Briggs, J. G. (2015). Out-of-class language contact and vocabulary gain in a study abroad context. *System, 53*, 129-140.

Bugel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal, 80*(1), 15-31.

Butler, Y. G. (2018). The role of context in young learners' processes for fesponding to self-assessment items. *The Modern Language Journal, 102*(1), 242-261.

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal, 90*(4), 506-518.

Butler, Y., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-31.

Byrnes, H. (2006). What kind of resource is language and why does it matter for advanced language learning. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsk*y (pp. 1-28). London, UK: Continuum.

Callahan, R. M., & Gándara, P. C. (2014). *The bilingual advantage: Language, literacy and the US labor market*. Bristol, UK: Multilingual Matters.

Carrell, P. L. (1984a). Evidence of a formal schema in second language comprehension. *Language learning, 34*(2), 87-108.

Carrell, P. L. (1984b). The effects of rhetorical organization on ESL readers. *TESOL Quarterly, 18*(3), 441-469.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19*(4), 727-752.

Chen, Q., & Donin, J. (1997). Discourse processing of first and second language biology texts: Effects of language proficiency and domain-specific knowledge. *The Modern Language Journal, 81*(2), 209-227.

Chen, Y. M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2), 235-262.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Chu, H. C. J., Swaffar, J., & Charney, D. H. (2002). Cultural representations of rhetorical conventions: The effects on reading recall. *TESOL Quarterly, 36*(4), 511-541.

Chun, D., Kern, R., & Smith, B. (2016). Technology in language use, language teaching, and language learning. *The Modern Language Journal, 100*, 64-80.

Cummins, J. (1981). Empirical and theoretical underpinnings of bilingual education. *Journal of Education*, *163*(1), 16-29.

de Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals, 42*(1), 158-178.

DeKeyser, R. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals, 43*(1), 80-92.

Ding, F., & Stapleton, P. (2016). Walking like a toddler: Students' autonomy development in English during cross-border transitions. *System, 59*, 12-28.

Dolosic, H.N., Brantmeier, C., Strube, M., & Hogrebe, M. (2016). Living language: self-assessment, oral production, and domestic immersion. *Foreign Language Annals, 49*(2), 302-316.

Donin, J., Graves, B., & Goyette, E. (2004). Second language text comprehension: Processing within a multilayered system. *Canadian Modern Language Review, 61*(1), 53-77.

DuBravac, S., & Dalle, M. (2002). Reader question formation as a tool for measuring comprehension: Narrative and expository textual inferences in a second language. *Journal of Research in Reading, 25*(2), 217-231.

Duque Micán, A., & Cuesta Medina, L. (2017). Boosting vocabulary learning through self-assessment in an English language teaching context. *Assessment & Evaluation in Higher Education, 42*(3), 398-414.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, *59*(4), 395-430.

Favreau, M. & Segalowitz, N. (1982). Second language reading in fluent bilinguals. *Applied Psycholinguistics, 3*(4), 329 – 341.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, *34*(10), 906.

Fletcher, T. (2010*). Psychometric: Applied Psychometric Theory* [Computer Software]. R package version 2.2.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* [Computer Software]. Thousand Oaks CA: Sage Publications.

Geisler, M., Kramsch, C., McGinnis, S., Patrikis, P., Pratt, M. L., Ryding, K., & Saussy, H. (2007). Foreign languages and higher education: New structures for a changed world: MLA ad hoc committee on foreign languages. *Profession*, 234-245.

Genesee, F. (2008). Dual language global village in pathways to multilingualism evolving perspectives on immersion education. Clevedon: Multilingual Matters.

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 153-174).

Gernsbacher, M. A., & Kaschak, M. P. (2013). Text comprehension. In Reisberg (Ed.) *The Oxford Handbook of Cognitive Psychology* (pp. 462-474).

Goldberg, D., Looney, D., & Lusin, N. (2015). *Enrollments in languages other than English in United States institutions of higher education, Fall 2013*. New York, NY: Modern Language Association.

Goodman, Y. M. (1967). A psycholinguistic description of observed oral reading phenomena in selected young beginning readers (Doctoral dissertation, Wayne State University).

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge, UK: Cambridge University Press.

Hamilton H. E. & Cohen, A. D. (2004). Creating a playworld: Motivating learners to take chances in a second language. In J. Frodesen & C. Holten (Eds.) *The power of context in language teaching and learning* (p. 237-247). Boston, MA: Thomson/Heinle.

Hammadou, J. (1991). Interrelationships among prior knowledge, inference, and language proficiency in foreign language reading. *The Modern Language Journal, 75*(1), 27-38.

Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*(1), 12-20.

Hatch, E. M., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York, NY: Newbury House Publishers.

Heilenman, K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2), 174-201.

Horiba, Y. (1996). Comprehension processes in L2 reading: Language competence, textual coherence, and inferences. *Studies in Second Language Acquisition, 18*(4), 433-473.

Horiba, Y., & Fukaya, K. (2015). Reading and learning from L2 text: Effects of reading goal, topic familiarity, and language proficiency. *Reading in a Foreign Language, 27*(1), 22-46.

Hothorn, T., Bretz, F., Westfall, P., & Heiberger, R. M. (2008). multcomp: simultaneous inference for general linear hypotheses [Computer Software].

Hubbard, P. (2004). Learner training for effective use of CALL. In Fotos & Brown (Eds.) *New perspectives on CALL for second language classrooms* (pp. 45-68). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hung, Y. J., Samuelson, B. L., & Chen, S. C. (2016). Relationships between peer-and self-assessment and teacher assessment of young EFL learners' oral presentations. In Nikolov (Ed.) *Assessing young learners of English: Global and local perspectives* (pp. 317-338). Heidelberg, Germany: Springer International Publishing.

Isabelli-García, C., & Lacorte, M. (2016). Language learners' characteristics, target language use, and linguistic development in a domestic immersion context. *Foreign Language Annals, 49*(3), 544-556.

Jafarpur, A., & Yamini, M. (1995). Do self-assessment and peer-rating improve with training?. *RELC Journal, 26*(1), 63-85.

Jeon, E. H., & Yamashita, J. (2014). L2 Reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160-212.

Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly, 15*(2), 169-181.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44*(2), 137-166.

Kaplan, R. B. (1988). Contrastive rhetoric and second language learning: Notes toward a theory of contrastive rhetoric. In Purves (Ed.) *Writing across languages and cultures: Issues in contrastive rhetoric* (pp. 275-304). Newbury Park, CA: Sage Publications.

Kashy, D. A., Campbell, L., & Harris, D. W. (2006). Advances in data analytic approaches for relationships research: The broad utility of hierarchical linear modeling. In Vangelisti & Perlman (Eds.), *Cambridge Handbook of Personal Relationships* (pp. 73-89). Cambridge, UK: Cambridge University Press.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363-394.

Koda, K. (2005). *Insights into Second Language Reading: A Cross-Linguistic Approach*. Cambridge, UK: Cambridge University Press.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning, 57*(1), 1-44.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349-370.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449-468.

Krausert, S.R., 1991. Determining the usefulness of self-assessment of foreign language skills: post-secondary ESL students' placement contribution. (Doctoral Dissertation, University of Southern California).

Lai, C., & Gu, M. (2011). Self-regulated out-of-class language learning with technology. *Computer Assisted Language Learning, 24*(4), 317-335.

Lantolf, J., Thorne, S. L., & Poehner, M. (2015). Sociocultural Theory and Second Language Development. In B. van Patten & J. Williams (Eds.), *Theories in Second Language Acquisition* (pp. 207-226). New York: Routledge.

Leader, W. S. (2008). Metacognition among students identified as gifted or nongifted using the discover assessment. (Doctoral Dissertation, The University of Arizona).

LeBlanc, R., & Painchaud, G. (1985). Self-Assessment as a second language placement instrument. *TESOL Quarterly, 19*(4), 673-687.

Li, X., & Cutting, J. (2011). Rote learning in Chinese culture: Reflecting active Confucian-based memory strategies. In L. Jin & M. Cortazzi (Eds.), *Researching Chinese learners: Skills, perceptions and intercultural adaptations* (pp. 21-42). Basingstoke, UK: Palgrave MacMillan.

Linholm-Leary (2011). Student outcomes in two-way chinese immersion programs: Language proficiency, academic achievement and student attitudes. In Tedick, Christian, & Fortune (Eds.) *Immersion Education Practices, Policies, Possibilities*. Clevedon: Channel View Publications.

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing, 22(*3), 321-336.

Little, D. (2009). Language learner autonomy and the European language portfolio: Two L2 English examples. *Language Teaching, 42*(2), 222-233.

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference?. *System, 37(*3), 353-365.

Llanes, À., Tragant, E., & Serrano, R. (2012). The role of individual differences in a study abroad experience: The case of Erasmus students. *International Journal of Multilingualism, 9*(3), 318-342.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning, 47*(2), 265-287.

Maoying, X., & Aiwu, J. (2011). Comparison between the students' perceived and expected behavior of college English teachers. *Chinese Journal of Applied Linguistics*, *34*(4), 72-88.

Malzloomi, S., & Khabiri, M. (2016). Diagnostic assessment of writing through dynamic self-assessment. *International Journal of English Linguistics, 6*(6), 19-31.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*(4), 516-522.

McNamara, M. J., & Deane, D. (1995). Self-assessment activities: Toward language autonomy in language learning. *TESOL Journal, 5*(1), 17-21.

Met, M. (2008). Paying attention to language: Literacy, language, and academic achievement in Fortune & Tedick (Eds.) *Pathways to multilingualism evolving perspectives on immersion education*. Clevedon: Multilingual Matters.

Metcalfe, J. (1996). Metacognitive processes. In Bjork & Bjork (Eds.) *Memory* (pp. 381-407). Academic Press.

Montanari, S. (2014). A case study of bi-literacy development among children enrolled in an Italian–English dual language program in Southern California. *International Journal of Bilingual Education and Bilingualism, 17*(5), 509-525.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533.

Naeini, J. (2011). Self-assessment and the impact on language skills. *Educational Research, 2*(6), 1225-1231.

Nagy, W. (2007). Metalinguistic awareness and the vocabulary-comprehension connection. *Vocabulary acquisition: Implications for reading comprehension,* 52-77.

Nassaji, H. (2002). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language Learning, 52*(2), 439-481.

Nation, P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL journal*, *9*(2), 6-10.

Nguyen, L. T. C., & Gu, Y. (2013). Strategy-based instruction: A learner-focused approach to developing learner autonomy. *Language Teaching Research, 17*(1), 9-30.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation, 15*(12), 2-9.

Oskarsson, M. (1978). *Approaches to self-assessment in foreign language learning.* Strasbourg, France: Council for Cultural Cooperation.

Padilla, A. M., Fan, L., Xu, X., & Silva, D. (2013). A Mandarin/English two-way immersion program: Language proficiency and academic achievement. *Foreign Language Annals, 46*(4), 661-679.

Pan, J. (2007). Facts and considerations about bilingual education in Chinese universities. In Feng (Ed.), *Bilingual education in China: Practices, policies and concepts* (pp. 200-218). Multilingual Matters: Clevedon, UK.

Paris-Kidd, H., & Barnett, J. (2011). Cultures of learning and student participation: Chinese learners in a multicultural English class in Australia. In L. Jin & M. Cortazzi (Eds.), *Researching Chinese Learners: Skills, Perceptions and Intercultural Adaptions* (169-187). Basingstoke, UK: Palgrave Macmillan.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Blackwell Publishing Ltd.: Malden, MA.

Plonsky, L. (2015). *Advancing quantitative methods in second language research*. New York, NY: Routledge.

Poehner, M. E. (2012). The zone of proximal development and the genesis of self-assessment. *The Modern Language Journal, 96*(4), 610-622.

Pulido, D. (2007). The effects of topic familiarity and passage sight vocabulary on L2 lexical inferencing and retention through reading. *Applied linguistics, 28*(1), 66-86.

R Core Team (2017). R: A language and environment for statistical computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing.

Revelle, W. R. (2017). *psych: Procedures for personality and psychological research* [Computer Software]. Evanston, Illinois.

Riley, G. L. (1993). A story structure approach to narrative text comprehension. *The Modern Language Journal, 77*(4), 417-432.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language testing, 13*(2), 173-189.

Rivers, W. P. (2001). Autonomy at all costs: An ethnography of metacognitive self-assessment and self-management among experienced language learners. *The Modern Language Journal*, *85*(2), 279-290.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20.

Savage, B. L., & Hughes, H. Z. (2014). How does short-term foreign language immersion stimulate language learning?. *Frontiers: The Interdisciplinary Journal of Study Abroad*, *24*, 103-120.

Schultz, L. M. (2017). Affect with Chinese learners of English: Enjoyment, self-perception, self-assessment, and abilities across levels of language learning. *Quarterly Journal of Chinese Studies*, *5*(2), 65-81.

Shahrakipour, H. (2012). On the impact of self assessment on EFL learners' receptive skills performance. *International Research Journal of Arts and Humanities, 40*(40), 143-164.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147-170.

Smith, K., & Craig, H. (2013). Enhancing the autonomous use of CALL: A new curriculum model in EFL. *CALICO Journal, 30*(2), 252-278.

Snijders, T. & Bosker, R. (2012*). Multilevel analysis: An introduction to basic and advanced level modeling* (2nd edition). Los Angeles, CA: Sage Publishing.

Steffensen, M. S., Joag-Dev, C., & Anderson, R. C. (1979). A cross-cultural perspective on reading comprehension. *Reading research quarterly*, *15*(1), 10-29.

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing, 32*(1), 63–81.

Swain, M. (2009). Languaging, agency and collaboration in advanced second language proficiency. In Byrnes (Ed.) *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95-108). London, UK: Continuum.

Sweet, G. & Mack, S. (2017, March). *Self-assessment and learner agency: A new approach.* Paper presented at American Association for Applied Linguistics, Portland, OR.

Sweet, G., Mack, S., & Olivero-Agney, A. (2017). Self-assessment in language courses: Does in-class support make a difference? In Alexander, I.D. & Poch, R.K. (Eds.) *Innovative Learning and Teaching: Experiments Across the Disciplines*. St. Paul, MN: Center for Educational Innovation of the University of Minnesota.

Swender, E., Conrad, D., & Vicars, R. (2012) American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines.

Tedick, D., Christian, & Fortune (2011). The future of immersion education: An invitation to "Dwell on Possibility". In Fortune, Tedick, & Christian (Eds.) *Immersion Education Practices, Policies, Possibilities*. Clevedon: Channel View Publications.

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*(4), 668-678.

Tschirner, E. (2009). The development of oral proficiency in a four-week intensive immersion program in Germany. *Unterrichtspraxis [Teaching German], 40*, 111-117.

Urquhart, A. H., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York, NY: Routledge.

Uso-Juan, E. (2006). The compensatory nature of discipline-related knowledge and English-language proficiency in reading English for academic purposes. *The Modern Language Journal, 90*(2), 210-227.

Van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Wickham, H. (2009) *ggplot2: Elegant graphics for data analysis* [Computer Software]. New York, NY: Springer-Verlag.

Wilson & Kamana (2011). Insights from indigenous language immersion in Hawai'i. In Fortune, Tedick, & Christian (Eds.) *Immersion Education: Practices, Policies, Possibilities*. Clevedon: Channel View Publications.

Wolf, D. F. (1993). A comparison of assessment tasks used to measure FL reading comprehension. *The Modern Language Journal, 77*(4), 473-489.

Wu, S. (2016). *The use of L1 cognitive resources in L2 reading by Chinese EFL learners*. London, UK: Routledge.

Yamagata, S. (2018). Comparing core-image-based basic verb learning in an EFL junior high school: Learner-centered and teacher-centered approaches. *Language Teaching Research*, *22*(1), 65-93.

Yoon, E., & Lee, H. K. (2013). Do effects of self-assessment differ by L2 language level? A case of Korean learners of English. *The Asia-Pacific Education Researcher, 22*(4), 731-739.

Yoshida, M. (2012). The interplay of processing task, text type, and proficiency in L2 reading. *Reading in a Foreign Language, 24*(1), 1-29.

Zhang, S., & Thompson, N. (2004). DIALANG: a diagnostic language assessment system. *The Canadian Modern Language Review, 61*(2), 290-293.

Zheng, X., & Borg, S. (2014). Task-based learning and teaching in China: Secondary school teachers' beliefs and practices. *Language Teaching Research, 18*(2), 205-221.

Ziegler, N. A. (2014). Fostering self-regulated learning through the European language portfolio: An embedded mixed methods study. *The Modern Language Journal, 98*(4), 921-936.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational Psychologist*, *25*(1), 3-17.

Appendix A

1. I can understand in detail a wide range of long, complex texts, if I can reread difficult sections.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

2. I can read short stories and follow the flow of thoughts and actions, understanding overall meaning and details.

1	2	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Appendix B

Self-Assessment

1. I can use my knowledge of grammar to understand stories and events in the past, including descriptive details.

1 2 3

2. I can use complex, challenging vocabulary to express my thoughts clearly and succinctly when I am talking with someone.

1 2 3

Language Knowledge

15. Monique _____ (être) une femme _____ (gentil).

20. Je _____ (se coucher) à 8h ce soir.
    a. me couche
    b. te couches
    c. coucheme
    d. couche

Reading Comprehension

22. What kinds of excavation techniques were used to discover this surprise?
    a. Step-by-step excavation
    b. Non-invasive scanning
    c. Rapid excavation for a new highway led to the initial discovery
    d. Rapid excavation for a new building led to the initial discovery

Appendix C

As discussed in the main body of Chapter 4, in order to get at the individualized nature of self-assessment this study used a hierarchical linear modeling approach, nesting multiple measures for each individual in an analysis (Snijders & Bosker, 2012). Hierarchical modeling provides a flexible approach to handling variables that are within-person and between-person simultaneously, capturing the complexity that is understood in self-assessment research (Kashy, Campbell, & Harris, 2006; Brantmeier, et al., 2012). In the present study's two-level analysis, multiple measures of reading comprehension and self-assessment were collected from each individual. Scores were extracted from multiple sections of the pre-tests and the post-tests, including subsections of grammar, vocabulary knowledge, and reading comprehension for both self-assessment and performance assessment. These smaller sections were found to be reliable measures of students' abilities and self-assessments (Cronbach's $\alpha > 0.70$ for all measures). Sections of the performance assessment were designed to be able to be analyzed separately as measures of their constructs. To validate the separation of the self-assessment items into subscales, a single confirmatory factor analysis was conducted. Through this analysis it was determined that the theorized separate dimension of self-assessment for grammar and vocabulary knowledge and three measures of reading comprehension were able to be separated with reasonable validity (TLI > 0.85). Therefore, for each individual 24 separate measures are included in the analysis, with 12 coming from the pre-test and 12 coming from the post-test. These measures were organized into a new data file and HLM analysis was then conducted within the open-source statistical program, R version 3.3.3 (R Core Team, 2017), using the lmer package (Bates, Machler, Bolker, & Walker, 2015).

In order to capture self-assessment accuracy from before and after participation in the immersion camp, a dummy coded no-intercept model was used. For this model, two dummy codes are used, one with 1 representing the pre-test while 0 represents the post-test and another with 0 representing the pre-test while 1 represents the post-test. All predictors were standardized. This approach allows modeling an intercept and slope separately for pretest and posttest for each participant in the study. Much like typical regression analyses, these intercepts and slopes represent the relationship between self-assessment and performance. Importantly, these slopes and intercepts can vary across individuals and that variability is available to be accounted for by individual characteristics. However, covariances among these variances were held to zero to overcome high correlations between pre-test and post-test scores. The formula for this analysis is:

$$Performance_{ij} = \beta_{1i}D_{1ij} + \beta_{2i}D_{2ij} + \beta_3 SA_{ij}D_{1ij} + \beta_4 SA_{ij}D_{2ij} \quad \text{(C.1)}$$

$$\beta_{1i} = \gamma_{10} + \mu_{1i} \quad \text{(C.2)}$$

$$\beta_{2i} = \gamma_{20} + \mu_{2i} \quad \text{(C.3)}$$

$$\beta_{3i} = \gamma_{30} + \mu_{3i} \quad \text{(C.4)}$$

$$\beta_{4i} = \gamma_{40} + \mu_{4i} \quad \text{(C.5)}$$

$Performance_{ij}$ is reading comprehension or language knowledge performance of a particular individual on a particular subscale and SA_{ij} is the corresponding self-assessment for that individual and that subscale. The dummy variables, D_{1ij} and D_{2ij}, dictate whether the particular performance and self-assessment pair represents pretest measures or posttest measures. Consequently, the model provides, for each participant, intercepts for the pretest (β_{1i}) and posttest (β_{2i}) as well as individual slopes for the pretest (β_{3i}) and posttest(β_{4i}).

As is reported in the main text, this model in which self-assessment is included in this individualized way, is an improvement over a null-model where no self-assessment is included ($\chi^2 = 17.91$, $p = 0.001$). The results of this model are reported in Table C.1.

Table C.1

Results of No-Intercept, Dummy Coded Model: Self-Assessment and Performance Measures

	Fixed Effects	
	β (SD)	*t*
Pre-Test	0.02 (0.11)	0.18
Post-Test	-0.01 (0.11)	-0.08
Pre-Test Self-Assessment	0.25 (0.09)	2.87
Post-Test Self-Assessment	0.19 (0.06)	2.98
	Random Effects	
	Variance (SD)	
Residual	0.39 (0.63)	
Pre-Test	0.43 (0.66)	
Post-Test	0.43 (0.66)	
Pre-Test Self-Assessment	0.06 (0.14)	
Post-Test Self-Assessment	0.02 (0.63)	

Follow-up tests examining differences among pretest and posttest coefficients were conducted using the multcomp package (Hothorn, Bretz, & Westfall, 2008). However, these tests demonstrated no significant differences between pretest and posttest intercepts and slopes ($p > 0.05$).

HLM also allows modeling variability at Level 2 as a function of between-participant variables, as can be seen in the formulas presented below. As a result, the variability in the intercepts and slopes can be accounted for by between-person moderators.

$$Performance_{ij} = \beta_{1i}D_{1ij} + \beta_{2i}D_{2ij} + \beta_3 SA_{ij}D_{1ij} + \beta_4 SA_{ij}D_{2ij} \quad \text{(C.6)}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}Level_i + \mu_{1i} \quad \text{(C.7)}$$

$$\beta_{2i} = \gamma_{20} + \gamma_{21}Level_i + \mu_{2i} \quad \text{(C.8)}$$

$$\beta_{3i} = \gamma_{30} + \gamma_{31}Level_i + \mu_{3i} \quad \text{(C.9)}$$

$$\beta_{4i} = \gamma_{40} + \gamma_{41}Level_i + \mu_{4i} \quad \text{(C.10)}$$

The new variable introduced at Level 2, referred to as "Level," is based on the ACTFL proficiency of students, as determined by the instructors as a team in order to confer grades with accuracy in the high school credit program. Due to sample size limitations, this predictor was entered into the model as a standardized continuous variable, despite the few levels present. The results of this model are presented in Table C.2.

With this new predictor in the model, comparisons were made to understand if this model was an improvement over the previous model. Such comparisons indicated that this model was an improvement over the initial self-assessment model ($\chi^2 = 113.74$, $p < 0.001$). Using multcomp, comparisons were also conducted within the model. Again, no significant differences existed between pre-test and post-test slopes or intercepts. Yet, results of this analysis indicated that Level did contribute to the model overall, providing an important factor within this model.

Table C.2

Fixed Effects of No-Intercept Dummy-Coded Model of Self-Assessment, Level, and Performance

	Fixed Effects	
	β (SD)	*t*
Pre-Test	-0.01 (0.06)	-0.17
Post-Test	-0.04 (0.73)	-0.59
Pre-Test Self-Assessment	0.09 (0.07)	1.47
Post-Test Self-Assessment	0.09 (0.06)	1.61
Pre-Test: Level	0.69 (0.06)	10.70
Post-Test: Level	0.60 (0.07)	8.11
Pre-Test: SA: Level	0.02 (0.06)	0.41
Post-Test: SA: Level	0.08 (0.06)	1.38
	Random Effects	
	Variance (SD)	
Residual	0.39 (0.63)	
Pre-Test	0.06 (0.25)	
Post-Test	0.13 (0.36)	
Pre-Test Self-Assessment	0.07 (0.10)	
Post-Test Self-Assessment	0.00 (0.00)	