Written Languaging, Learners' Aptitude and Second Language Learning

Masako Ishikawa

UCL Institute of Education, University College London
Thesis submitted for the degree of Doctor of Philosophy

I, Masako Ishikawa, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature_____

Abstract

Languaging (Swain, 2006), defined as learners' language use to make meaning, has been suggested and identified as a way to facilitate second language (L2) learning. Most of the research conducted so far has been on oral languaging, whereas the effectiveness of written languaging (WL) in promoting L2 development remains underexplored. To help to bridge this gap, this thesis examined (1) the impact of WL on L2 learning, (2) the relationship between the frequency/quality of WL and L2 learning, and (3) the associations between L2 learning through languaging and individual differences in aptitude and metalanguage knowledge.

The study used a pretest-posttest-delayed posttest design with individual written dictogloss as a treatment task. The participants were 82 adult EFL learners, assigned to three groups: +WL group, –WL group or a control group. The +WL group engaged in WL by writing about their linguistic issues when they compared their reconstructions and an original text, whereas the –WL group completed the same task without engaging in WL. The control group simply did the pre- and posttests. The assessments included an essay test, a grammar production test and a recognition test. The MLAT, LLAMA_F, and LABJ were employed as aptitude measures. A metalanguage knowledge test was also devised and administered to the participants. Finally, they completed an exit questionnaire.

Three main findings emerged. First, the +WL group outperformed the –WL group on the grammar production tests and, to a lesser degree, on the essay tests. Second, significant correlations were observed between the frequency/quality of WL and the gain scores on two grammar tests. Finally, a greater number of significant associations were identified between aptitude/metalanguage knowledge and L2 learning for the –WL group than the +WL group. These results are discussed with reference to previous research in second language acquisition and cognitive psychology.

Impact Statement

Learners' using language to reflect on their language use (i.e., languaging) facilitates second language (L2) learning (Swain, 2006). Ample evidence has been produced regarding the positive impact of oral languaging on L2 learning, whereas much less is known about written languaging (WL). In addition, although individual differences in language aptitude and metalanguage knowledge are likely to influence the effects of languaging, regardless of its modality, no research has been conducted yet.

Thus, the present thesis conducted an experiment to investigate the overall effects of WL on L2 learning and the relationship between L2 learning and individual differences. Two promising possibilities emerged. First, WL can function as a learning tool. Second, WL might level out individual differences in learners' language aptitude and metalanguage knowledge. These results seem to be attributable to two useful characteristics of WL. First, as WL is an individual activity, learners can engage in WL at their own pace, free from time pressure. Second, the process of WL produces products for learners to reflect on, presumably resulting in deeper understanding. An additional strength of WL is its high practicality. That is, as WL only requires pen and paper, it can be used anytime, anywhere by anyone, including learners who have no access to technology or those in disadvantaged areas.

Given the above, WL could be used as a domain-independent learning strategy in both L2 and non-L2 domains. In the domain of L2, as WL seems to benefit learners regardless of their individual differences, instructors might be able to utilise WL as one of their teaching techniques, especially in remedial programmes where learners are likely to benefit from additional assistance. What is more, the products of WL are expected to benefit learners and instructors alike, being a source of further reflection on their linguistic issues and of precious information regarding learners' development, respectively. Similarly, in non-L2 domains, WL is likely to facilitate learners reflecting on their domain-specific issues as a learning tool, leading to the development of domain knowledge as a consequence. Furthermore, WL may be employed as a facilitative tool even outside academia, such as for employee training in corporate settings and the development of skills in athletic fields. In addition, as an individual activity, WL may be used as a tool for self-improvement, benefitting people on a personal level.

This research project was started to make use of the study habit of learners who are quiet and reluctant to interact with others verbally but take notes frequently. Although WL is still a fairly new concept and has not been explored extensively, it seems to be a promising learning tool that could impact people in any settings, especially given its high practicality.

Table of Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

"Language is so tightly woven into human experience that it is scarcely possible to imagine life without it" (Pinker, 1994, p 3). Nonetheless, like air, language is so natural for us that we tend to overlook its significance, thinking of it as merely a communication tool. It has been identified, however, that language has more functions than that. People speak and write to clarify their thinking and remember things; namely, producing language helps people complete their thoughts and transform these thoughts into artefacts for further reflection, thereby mediating their cognition as a cognitive tool (Vygotsky, 1986, 1987). Informed by sociocultural theory (SCT), Swain (2006) introduced the term "languaging," referring to these mediating functions of language, and explained that it is "an action—a dynamic, never-ending process of using language to make meaning" (p. 96).

Although the importance of language production for language learning is now generally acknowledged, that was not the case until at least the middle of the 1980s. At that time, the Input Hypothesis, proposed by Krashen (1981), dominated the field of second language acquisition (SLA) research. It claimed that comprehensible input is all that is needed to achieve native-like proficiency in a second language. In such a context, Swain (1985) studied learners in French immersion programmes in Canada and noticed that many of them attained native-like proficiency level in terms of receptive skills (i.e., listening and reading), but not productive skills (i.e., speaking and writing) in spite of abundant input. Based on this, she proposed her initial version of the Output Hypothesis, emphasizing the importance of output (i.e., language production) in language learning.

More recently, Swain (2005, 2006) has highlighted the significance of not only task output (i.e., primary output) but also the language used to complete tasks (i.e., metalinguistic output, languaging). Swain argues that "languaging about language is one of the ways we learn a second language to an advanced level" (2006, p. 96). Furthermore, pointing to the fact that language use, such as think-alouds and stimulated recalls, is often used as a research tool to collect data (Ericsson & Simon, 1993), Swain claimed that it could also be a learning tool. She stated "If … verbalization has the impact I am suggesting on second language learning, then research tools such as think-alouds and stimulated recalls, need to be understood as part of the learning process, not just as a medium of data collection … They are part of what constitutes learning" (2005, p. 480). She further argues, "Speaking (and writing) are conceived of as cognitive tools—tools that mediate internalization; and that externalize internal psychological activity…" (p. 480).

Supporting Swain's (2005) arguments, SLA studies have to date provided ample evidence that languaging can contribute to second language (L2) learning (e.g., Brooks, Swain, Lapkin, & Knouzi, 2010; Knouzi, Swain, Lapkin, & Brooks, 2010; Swain, Lapkin, Knouzi, W. Suzuki, & Brooks, 2009; Yang, 2016). In addition to these studies, the following studies use other terms for languaging, for example, collaborative dialogue (e.g., Ammar & Hassan, 2017; Storch & Wigglesworth, 2010), metatalk (e.g., Storch, 2008; Swain, 1998) and private speech (e.g., Negueruela & Lantolf, 2006; Ohta, 2001; Yoshida, 2008). While these SLA studies do not use the term "languaging," they were conducted under the guise of languaging, supporting Swain's work.

Meanwhile, although not yet fully examined in L2-domains, in the field of cognitive psychology, "self-explaining," a type of language use that shares a similar concept with languaging, has been reported to have a positive impact on learning.

Defining self-explaining, Chi (2000) states that it is "the activity of explaining to oneself in an attempt to make sense of new information" (p. 164). As the definition demonstrates, while languaging is the act of producing language about language, and thus has been investigated in L2-domains, self-explaining refers to the act of explaining new information in general. Self-explaining, therefore, has been examined in various fields, such as biology and physics (e.g., Chi, 2000; McNamara, 2004, 2017) (see also Ericsson & Simon, 1993, for verbal protocols).

The first self-explaining studies were conducted by Chi and her colleagues starting in the late 1980s (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & La Vancher, 1994). The researchers conducted their research based on their belief that the acquisition of new knowledge could not be achieved solely by direct instruction (i.e., learning by listening to instructions and/or reading textbooks). They believed, instead, that learners should also be actively involved in the construction of their own knowledge, which led them to "the discovery of this learning strategy" (Chi, 2000, p. 162), i.e., self-explaining. Chi highlights the importance of learners' active involvement in learning, quoting Benjamin Franklin's (1706–1790) inspirational words, "Tell me and I forget. Teach me and I remember. Involve me and I learn" (p. 161).

As introduced briefly, the concepts of languaging and self-explaining are similar in that they both involve the act of externalising thoughts by using language. However, there are at least two differences between the two. First, as stated, languaging is the act of using language about *language*, whereas self-explaining can be used for any subject. Second, the perspective of SCT considers languaging to be a source of learning. That is to say, learning is continuously mediated by the articulation of thoughts. In addition, from an SCT perspective, externalised thoughts are transformed into artefacts, which enables learners to reflect on their thoughts even further. In contrast, from the

perspective of cognitive psychology, self-explaining triggers cognitive processes that are conducive to learning. Thus, self-explaining is one of the strategies for learners to use when they are faced with cognitively complex tasks. Despite these differences, languaging and self-explaining seem to share a fundamental concept, in that they assume that not only exposure to instruction (input) but also learners' active involvement, i.e., the externalisation of thoughts with language (output), plays a significant role in learning. Accordingly, researchers in both disciplines have been interested in how to promote learning by studying the ways in which learners verbalise their understanding and/or ask questions about instructional materials with others or oneself.

To date, both SCT and cognitive psychology have extensively examined oral versions of such verbal deliberation, whereas its written counterpart, i.e., written languaging (WL), has been underexplored. Compared to speaking, writing generally offers learners more time (Manchón, 2011; J. Williams, 2012). In addition, the product of their writing enables them to reflect on their thoughts more easily than when speaking, offering optimal conditions for learning (W. Suzuki, 2012). In spite of these positive factors, researchers have only recently begun to explore if WL might facilitate L2 learning (e.g., W. Suzuki, 2009a, 2009b, 2012). Although some WL studies have reported a facilitative impact on L2 learning (e.g., M. Ishikawa, 2013; W. Suzuki, 2012), such as a positive effect on revision (W. Suzuki, 2012) and better performances by some case study participants than those who did not engage in WL (M. Ishikawa, 2013), none of them have shown a direct link between WL and L2 learning.

1.1 Aims and Rationale of the Present Thesis

Against the background mentioned above, the present thesis had eight aims. First, it was intended to further explore the impact of WL on L2 learning. Previous SLA research findings on oral languaging (OL) suggest that the content of OL, i.e., language-related episodes (LREs), may influence L2 learning (e.g., Swain et al., 2009). (An LRE is defined as "any segment of the protocol in which a learner … spoke about a language problem he/she encountered" (Swain & Lapkin, 1995, p. 387).) This relationship, however, has not yet been fully examined in the written modality. Second, it aimed to investigate the frequency of WL episodes (WLEs), i.e., LREs written by learners, focusing on the target construction (i.e., T-WLEs), in relation to L2 learning (pre- and posttest scores). The third aim was to delve into the association between the quality of T-WLEs and L2 learning. Based on the distinction of the level of noticing by Schmidt (1990), the quality of T-WLEs was operationalised as the level of noticing reflected in T-WLEs. Fourth, the correlation between the frequency and quality of T-WLEs was examined in order to explore a possible link between the number of T-WLEs and the level of noticing/understanding displayed by learners. In other words, my aim was to reveal whether focusing on the target feature and producing more T-WLEs addressing it was linked to L2 learning. Although a significant correlation between the quantity and quality of OL has been reported (e.g., Swain et al., 2009), that for WL has not yet been investigated.

The fifth aim of the present thesis was to look into how individual differences in language aptitude might influence the impact of WL on L2 learning. Language aptitude has been recognised as one of the primary predictors of success in L2 learning (Dörnyei, 2010) and many studies have reported that the success of L2 learning can vary depending on language aptitude (e.g., Erlam, 2005; Li, 2013; Sheen, 2007a, 2007b; Stefanou & Révész, 2015; Y. Yilmaz, 2013). However, there is as yet no empirical

evidence for a potential association between aptitude and the impact of languaging (either oral or written) on L2 learning. Therefore, it was deemed important to investigate the extent to which learners with different aptitude profiles, i.e., varying characteristics of learner abilities, might benefit from WL. Sixth, this thesis examined the relationship between metalanguage knowledge and the benefits of WL on L2 learning. As is the case with aptitude, no research has been conducted to investigate this link. Given that WL provides learners with opportunities to verbalise their thoughts regarding linguistic issues in writing, their metalanguage knowledge is likely to influence the quantity and quality of WL and the extent to which learners benefit from it.

Seventh, this thesis attempted to identify how the frequency of T-WLEs relates to aptitude and metalanguage knowledge. Eighth, it also examined the relationship between the quality of T-WLEs and individual differences in aptitude and metalanguage knowledge. As stated, although the amount and quality of OL has been shown to have an impact on L2 learning (Swain et al., 2009), how language aptitude and metalanguage knowledge are related to the quantity and quality of languaging, regardless of its modality, has yet to be investigated.

The specific research questions addressed in this thesis were as follows:

1. To what extent does WL facilitate L2 learning?
2. To what extent does the frequency of T-WLEs relate to development in knowledge of the target construction?
3. To what extent does the quality of T-WLEs relate to development in knowledge of the target construction?
4. To what extent are the frequency of T-WLEs and the quality of T-WLEs related?

5. To what extent does language aptitude moderate the effect of WL on L2 learning?
6. To what extent does metalanguage knowledge moderate the effect of WL on L2 learning?
7. To what extent is the frequency of T-WLEs related to individual differences in language aptitude and metalanguage knowledge?
8. To what extent is the quality of T-WLEs related to individual differences in language aptitude and metalanguage knowledge?

With these research questions in mind, the present thesis included one independent variable, WL, which was operationalised as the provision or no provision of an opportunity for WL during the experiment (+/ –WL groups). The dependent variable was development in knowledge of the target construction, measured in terms of participants' performance on a posttest as compared to a pretest. The moderator variables are the results of aptitude tests and a metalanguage knowledge test. Therefore, the first research question was addressed by comparing the performances of a +WL group with those of a –WL group and a control group on pre- and posttests. The second research question investigated the frequency of T-WLEs in relation to the learning of the +WL group. Similarly, the third research question examined the quality of T-WLEs in relation to the learning of the +WL group. The fourth research question delved into the relationship between the frequency of T-WLEs and the quality of T-WLEs.

Meanwhile, the fifth research question was addressed by investigating the relationship between aptitude test results and the extent of development displayed by participants in the +WL and –WL groups. Similarly, the sixth research question was investigated by studying how the results of a metalanguage knowledge test were related

to the development demonstrated by the participants in both the +WL and –WL groups. The seventh research question was addressed by computing the correlation between the frequency of T-WLEs and the results of aptitude tests and a metalanguage knowledge test. Likewise, the eighth research question was investigated by analysing how the quality of T-WLEs relates to the results of aptitude tests and a metalanguage knowledge test.

1.2 Definitions of Terms

1.2.1 Languaging

While Swain (2006, 2010) is the researcher who brought the concept of languaging into focus, others had used the term "languaging" previously. Initially, Lado (1979) used the term loosely, referring to overall linguistic performance rather than individual linguistic features, such as a single word, a grammar rule or a mere pronunciation problem. Accordingly, for Lado, "in languaging our attention is not on the language" (p. 96). Similarly, Emig (1977) employed the term "languaging" in a broad sense, using it interchangeably with an adjective, "linguistic," and describing "listening, speaking, reading and writing" as "the four languaging processes" (p. 122). Following this line of thinking, Becker (1991a, 1991b) conceptualizes languaging as "an activity of human beings in the world" (1991a, p. 34) and states that "Languaging about language is as everyday as languaging about anything else" (1991b, p. 229). Nonetheless, this thesis focused on Swain's (2006) conceptualisation of languaging, defined as "the process of making meaning and shaping knowledge and experience through language" (p. 98). In addition, WL was defined as "any language noted by learners to reflect on their language use, with or without metalinguistic terminology," following Swain's (1998) interpretation of metatalk, a form of OL.

1.2.2 Self-explaining and Self-explanation

As stated above, self-explaining is a learning strategy proposed by Chi and her colleagues (Chi et al., 1989; Chi et al., 1994). It is almost thirty years since Chi et al.'s (1989) original study, and the term "self-explaining" has been used somewhat differently from study to study. Some researchers, for instance, use the two terms "self-explaining" and "self-explanation" synonymously (e.g., Griffin, Wiley, & Thiede, 2008; Ionas, Cernusca, & Collier, 2012), whereas others use the term "self-explanation" instead of "self-explaining" (e.g., Rittle-Johnson & Loehr, 2017; Tajika, Nakatsu, Nozaki, Neumann, & Maruno, 2007) to refer to the activity of self-explaining. According to Chi (2000), however, these terms refer to different aspects of self-explanation research. To be more specific, self-explaining refers to "the *activity* of generating explanations to oneself" (pp. 164–165), whereas "self-explanation" refers to "a unit of utterances produced by self-explaining" (p. 165). Moreover, she states that self-explanation in the plural form, i.e., "self-explanations," refers to the more general "entire corpus of collective utterances or verbal protocol data gathered from self-explaining in a particular study" (p. 165). In other words, according to her distinction, self-explaining is the process of actual verbalization, whereas self-explanation and self-explanations are the products of the process. Given that self-explanation research originated from Chi and her colleagues' study (Chi et al., 1989), this thesis adhered to their distinctions of self-explaining as an activity/process and self-explanation(s) as product(s).

1.2.3 Metalanguage Knowledge

The terms used for knowledge of metalinguistic terminology differs among researchers. For example, assuming that metalinguistic terminology is an essential part

of explicit metalinguistic knowledge, Alderson, Clapham and Steel (1997) define "metalinguistic knowledge" as "knowledge about language" and used the term (i.e., metalinguistic knowledge) to refer to both knowledge of metalanguage and knowledge of linguistic structures. In contrast, other researchers draw a clear distinction between the two types of knowledge and use separate terms for them (e.g., Berry, 2005, 2009; R. Ellis, 2004, 2009; Hu, 2011). Berry (2009), for instance, refers to metalanguage knowledge as "metalingual knowledge" and knowledge about language as "metalinguistic knowledge" (p. 114) (see Gutiérrez, 2013; Hu, 2011, for a similar distinction). Although R. Ellis (2004) also makes a distinction between the two, he uses the terms "metalinguistic knowledge" and "metalanguage knowledge" interchangeably.

Thus, in order to avoid confusion, the term "metalanguage knowledge" was used in the present thesis to refer to knowledge of metalinguistic terminology, i.e., the terminology used to describe language (R. Ellis, 2004) as opposed to metalinguistic knowledge, i.e., knowledge about language (Berry, 2014). Furthermore, metalinguistic knowledge and explicit knowledge were used synonymously in this thesis.

1.2.4 Second Language Acquisition, Learning, and Development

In DeKeyser's (2007) view, second language learning is essentially skill acquisition, where practice plays the role of facilitator in improving language skills. Regarding the process, he explains that learners initially acquire declarative knowledge, which is conscious and explicit. Then, a great deal of practice in using this knowledge transforms it into procedural knowledge, whose application is fast and automatic. Following his distinction, this thesis investigated changes in learners' interlanguage development in two dimensions, declarative knowledge and procedural knowledge. Although "acquisition" and "learning" are sometimes distinguished from each other

(Krashen, 1981), the terms *acquisition*, *learning*, and *development* are used interchangeably in this thesis, referring to any improvement in knowledge of the target construction.

The term *learning* was operationalised as an increase in learners' interlanguage knowledge of the target construction from pretest to posttest.

1.3 Structure of the Thesis

The remainder of the present thesis is organised as follows. Chapter 2 and Chapter 3 conduct a review of several strands of the research motivating this research endeavour. In Chapter 2, theoretical accounts of languaging and self-explaining, both oral and written, and their findings are mainly reviewed. In Chapter 3, individual differences in aptitude and metalanguage knowledge are discussed in relation to WL. In Chapter 4, the details of the pilot study are introduced and the findings reviewed. Chapter 5 begins with an overview of the research design of the main study. Then, an introduction to the participants and a description of the measures used to assess L2 learning, aptitude and metalanguage knowledge are presented, followed by the coding of WLEs. The chapter ends with explanations of the questionnaires, interview, case study and data analysis procedures. Chapter 6 is devoted to presenting the results of the quantitative analyses, while Chapter 7 presents the results of the qualitative analyses. Chapter 8 discusses all the results reported in Chapters 6 and 7 with regard to the specific research questions proposed in Chapter 3. Finally, Chapter 9 summarises the main findings and outlines some possible implications of the results and directions for future research.

CHAPTER II
LANGUAGING AND SELF-EXPLAINING

In this chapter, the theoretical background to languaging and self-explaining, both oral and written, is first reviewed from the perspectives of sociocultural theory and cognitive psychology, respectively, followed by an overview of relevant research findings. Then, the relationship between WL and the type of learning which it is likely to facilitate is considered.

2.1 Languaging

2.1.1 Theoretical Accounts of Languaging and Self-explaining

As mentioned in Chapter I, informed by SCT, Swain (2006, 2010) proposed the term and concept of languaging about a decade ago. Thus, languaging is still a fairly new concept and has not been fully explored. It should be pointed out, however, that even before the term was introduced, some studies were framed around the concept of languaging but using terms such as "collaborative dialogue" (e.g., Storch, 2001, 2002), private speech (Lantolf & Thorne, 2006; Ohta, 2001) and "metatalk" (e.g., Swain, 1998). Although the concepts of these terms are similar in that they all consider the act of using language as having a facilitative effect for learning, there are some differences as explained below.

As the name suggests, collaborative dialogue refers to "dialogue in which L2 learners are engaged in problem solving and knowledge building" (Swain & Lapkin, 1998, p. 102) *collaboratively*. In contrast, private speech is "oral language uttered not for communicative interaction with another, but for dialogue with the self" (Ohta, 2001, p. 14) in order to monitor and reflect one's mental activity. (The concept of private

speech is explained in more detail in the next section.) With respect to the two concepts, Swain, Kinnear, and Steinman (2011) describe collaborative dialogue and private speech as "talking with others and talking with the self" (p. 34), respectively, and explain that they are "two different forms of languaging" (p. 34).

Meanwhile, metatalk is a term and concept introduced by Swain (1998), who describes it as learners' talk to reflect on their own language use. The concept of metatalk originates from metalinguistic/reflective function of output, one of the three functions of output suggested in Swain's (1985) Output Hypothesis. (This function is described in detail in the next section.) Metatalk seems to be closer to collaborative dialogue, in that Swain highlights the importance of encouraging metatalk to learners when they are engaged in making meaning, "where the language being used and reflected upon through metatalk is serving a communicative function" (p. 69), i.e., talking with others.

As stated above, collaborative dialogue, private speech, and metatalk seem to be regarded as forms of languaging, on the ground that they all consider that using language about language facilitates learning. These terms, however, only refer to oral language use, which tends to result in a misinterpretation that the benefits of languaging refer only to speaking. Against this background, the term "languaging" was introduced by Swain (2006), who stresses that languaging includes both speaking and writing.

Meanwhile, as stated, in the field of cognitive psychology, self-explaining (Chi et al., 1989) emerged as a concept which, like languaging, is concerned with the impact of the externalisation of inner psychological activities with language on learning. The study of self-explaining has a longer history, with many empirical research findings (e.g., Chi et al., 1989; Chi et al., 1994; Chi & Chiu, 2014). Therefore, in this section,

theoretical accounts of both languaging and self-explaining, in oral and written modalities, are discussed.

2.1.1.1 Theoretical Accounts of OL

The concept of languaging originates from the perspective of the Russian sociocultural psychologist, Vygotsky's (1896–1934) SCT of mind, which views language as an essential mediator of cognition. According to Vygotsky (1987), "Thought is not merely expressed in words; it comes into existence through them" (p. 219). Reflecting on this perspective, sociocultural psychology assumes that humans develop through the internalization of language, and that language allows humans to develop and regulate their thinking and behaviour (Vygotsky, 1986, 1987) (see also Slobin, 1987, for his hypothesis of "thinking for speaking" for a similar concept). Although Vygotsky targeted his argument at the development of children, it has been argued to apply to adults as well (Duncan & Cheyne, 2001; John-Steiner, 1991; McCafferty, 1992). Vygotsky published his seminal works on language and thought, originally in Russian, in the 1930s, but they became widely recognised outside Russia after they were translated into English in the 1960s.

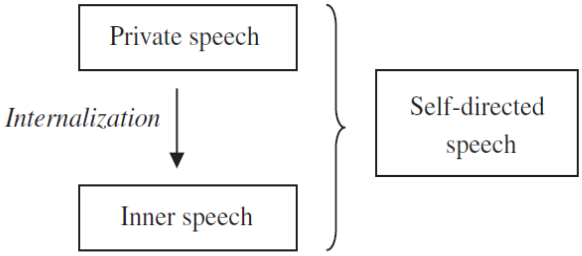
Vygotsky (1987) states that every psychological function appears twice; first between people on the inter-psychological level, then within the individual on the intra-psychological level. Namely, externalised thoughts are eventually internalized within each individual, and language is believed to play an essential role as a mediator of internalization. In Vygotsky's (1986, 1987) view, the earliest speech of children is essentially social, i.e., social speech. That is, they use language for inter-personal communication with people around them (e.g., their parents and caregivers), where language is expected to guide and regulate their behaviour and attention (other-

regulation). Then, children start to use language for behavioural self-regulation too, i.e., private speech, which is overt, audible speech directed to the self. Because of its overt nature, private speech is often mistaken as social or communicative, but it is for intra-personal communication and self-guidance. On that point, DiCamilla and Lantolf (1994) state that "private speech is directed by the self as speaker to the self, as listener for the purpose of organizing and directing the strategies (e.g., attention, planning, thinking, remembering, evaluating) entailed in mental activity" (p. 348) (see also Lantolf & Thorne, 2006). In this vein, Vocate (1994), basing much of her research on SCT, describes private speech as dialogue between "I" and "Me," in which "I" is a creative action taker, while "Me" monitors the action. Along the same lines, Winsler (2009) states that private speech enables children to "reflect better on their own thinking and behaviour and reach greater levels of control and mastery over their own behaviour" (p. 4).

According to Vygotsky, private speech remains an important self-regulatory tool throughout human development. Supporting his perspective, studies on private speech among adults demonstrate that they too use private speech, especially when faced with cognitively challenging tasks (e.g., Duncan & Cheyne, 2001; Duncan & Tarulli, 2009). Private speech is, however, not the end product of the transformation of social speech. As a final step, SCT assumes that private speech eventually evolves into inner speech, which is "fully internal, silent verbal thought—that is, speech fully inside one's head" (Winsler, 2009, p. 1), and functions as an internal regulator of behaviour and cognition. Referring to both private speech and inner speech as self-directed speech, Lidstone, Meins, and Fernyhough (2010) explain that this shift from overtness toward covertness in self-directed speech reflects the gradual internalization of private speech to inner speech (p. 439) (see Figure 2.1).

Figure 2.1

Conceptual Relations among Private Speech, Inner Speech, and Self-directed Speech



Adapted from Lidstone et al. (2010, p. 439)

The concept of internalization is a core principle of Vygotsky's approach to explaining human development. Extending the concept, Gal'perin (1902–1988), who was a contemporary of Vygotsky and one of the major figures in the sociocultural school of thought, conceptualizes internalization as "a transformation of *material* forms of individual external activity into *mental* forms of that same external activity" (Arievitch & Haenen, 2005, p. 158). Furthermore, applying his conceptualization of internalization to teaching and learning, Gal'perin devised his own teaching strategy based on the idea that learning any kind of knowledge can be extended to learning different kinds of actions (activities). In his view, "actions can be carried out with support at three different levels of abstraction, i.e., the material level, the verbal level and the mental level" (Lantolf & Thorne, 2014, p. 62). Pointing to the three levels, he explains that learners start acting at the material level where they learn with the help of physical objects, then move on to act at the verbal level, where physical objects are replaced by words and actions are executed verbally, finally reaching the mental level where actions are based on thought as a result of internalization. The most relevant to

the current thesis are the verbal and mental levels of action, which seem to correspond to private speech and inner speech, respectively.

Based on this conceptualization, Gal'perin (1992) proposed his theory of Systematic Theoretical Instruction (STI). With the help of his colleagues, he tested and modified his theory of STI, completing the stepwise instruction procedures (Lantolf & Thorne, 2014), where he emphasizes that it is essential for learners to verbalize instruction materials for them to internalize the materials. His contribution is significant, in that he put theory into practice by connecting instruction and action. What is also noteworthy is that Gal'perin aimed to enhance learners' potential to learn in the zone of proximal development (ZPD) by facilitating them shaping "the ability to cope independently with a new, previously unfamiliar task … i.e., the ability to learn from something new independently" (Gal'perin, p. 79). According to Vygotsky (1986), ZPD, one of the important concepts of SCT, is "the discrepancy between a child's actual mental age and the level he reaches in solving problems with assistance" (p. 187).

In the domain of L2 learning, supporting Vygotsky's argument that thinking is intimately related to language, Swain (2006) maintains that speaking and writing serve as "tools of the mind, mediating the cognition and re-cognition of experience and knowledge" (p. 106). In her view, learners' externalisation of thoughts while they work on tasks, i.e., languaging, leads to their L2 learning, which seems to run in parallel to the transformation of private speech (external language use for self-regulation) into inner speech (internal regulator of behaviour and cognition). (It should be pointed out, however, that languaging includes both interpersonal and intrapersonal communication because "talking with (or writing to) others and talking with (or writing to) oneself are connected theoretically and in practice" (Swain & Watanabe, 2013, p. 1).)

Swain states that the ways learners benefit from languaging are twofold. First, learners externalize their thoughts, and then these externalized thoughts provide them with objects to reflect on, enabling them to learn through the process as well as the product of their languaging. Put differently, languaging is expected to be a process which creates a visible and audible product about which one can language further (Swain, 2006), which seems to be compatible with the claims of the Output Hypothesis (Swain, 1985, 1995, 1998, 2005).

In light of the three functions of output proposed by the Output Hypothesis (Swain, 2005), i.e., noticing/triggering function, metalinguistic/reflective function, and hypothesis testing function, languaging is likely to enhance the first two functions of output. First, noticing has been regarded as a facilitator of L2 learning (Schmidt, 1990, 1993, 1994, 1995, 2001, 2010; Schmidt & Frota, 1986) since Schmidt's (1990) Noticing Hypothesis was proposed. In his view, noticing is a low level of awareness, which is "nearly isomorphic with attention, and seems to be associated with all learning" (1995, p. 1). (The notions of noticing and awareness are discussed in detail in section 2.3.2.) In terms of the noticing function of output, Swain (1995) states that task output may enable learners to "notice a gap between what they want to say and what they can say, leading them to recognize what they do not know, or know only partially" (pp. 125–126). Similarly, Swain and Lapkin (1995) explain that output can be "one of the triggers for noticing" (p. 373) a gap or a hole, which is when learners cannot express an idea in the target language (TL) (Swain, 1995). Given this, learners' languaging about the gaps or holes in their own language use (i.e., task output) is hypothesised to enhance their noticing even further. Namely, by the act of languaging, learners are more likely to notice their linguistic issues.

Second, in terms of the reflective function of output, Swain (1998) states that learners use language to reflect on language use, explaining that "learners' own language indicates an awareness of something about their own, or their interlocutor's, use of language" (p. 69). Swain (1995) further explains that "as learners reflect upon their own TL use, their output serves a metalinguistic function, enabling them to control and internalize linguistic knowledge" (p. 126). In line with Swain's view regarding the positive impact of reflection on learning, Izumi (2003) argues that "Reflection on language may deepen the learners' awareness of forms, rules, and form-function relationships" (p. 170), thus facilitating their L2 learning. More recently, concerning the reflective function of languaging, Swain (2005) made a similar claim, stating that "using language to reflect on language produced by others or the self, mediates second language learning" (p. 478). It is important to note that the process of reflection through output and OL assumes the existence of products of output and OL, respectively. Given this, engaging in OL on one's own output may be interpreted as yielding "double products."

In summary, Swain's argument that learners benefit from the noticing and reflective processes of OL appears to be compatible with the noticing and reflective functions of output.

2.1.1.2 Theoretical Accounts of WL

As mentioned above, "languaging" is a term introduced by Swain (2006), who stresses that languaging includes both speaking and writing. Theoretically, WL is likely to have an equivalent mechanism to that of OL, in that learners externalise their thoughts with language. That said, there are some fundamental differences that derive from their respective modalities. Namely, compared to speaking, writing tends to be

regarded as a more time-consuming and effortful process (Muñoz, Magliano, Sheridan, & McNamara, 2006). In the case of language learning, at least in the case of WL, however, these seemingly negative aspects of writing can be considered as potentially facilitative.

First, the more time-consuming process of writing can be considered a beneficial trait, in that it requires more time, i.e., a slower pace than speaking, which is an ideal condition for learning (W. Suzuki, 2012). In addition, unlike speaking, which usually involves interlocutors, writing is "intrinsically an individual enterprise" (Manchón, 2011, p. 76). Although research findings suggest that learners benefit more when they work with peers in a collaborative manner (Storch, 2013), engaging in writing individually can be advantageous in that learners can focus solely on the act of writing at their own pace, usually under minimal time constraints (J. Williams, 2012). Supporting this argument, from a "writing-to-learn" perspective, Manchón (2011) states that linguistic processing, including noticing and metalinguistic reflection, is "more likely to take place in writing than in speaking" (p. 70) because of "the greater availability of time in writing" (p. 71). Furthermore, she emphasizes that "writing … fosters a type of linguistic processing with potential learning effects" (p. 70), describing writing as "a tool for language learning" (p. 69).

Along the same lines, J. Williams (2008) points to two beneficial by-products of the slower pace of writing: planning time and access to explicit knowledge. Namely, the slower pace allows learners to plan what they write (in the case of WL, learners are likely to think about their linguistic issues). In addition, the slower pace of writing permits learners to consult their explicit knowledge, which is unlikely to be feasible while speaking because of time pressure. Given that writing is generally considered to be a learned behaviour acquired with the aid of formal and systematic instruction (Emig,

1977), the act of writing seems to be a conscious process by its very nature, probably making it easier for leaners to draw on their explicit knowledge when they write compared to when they speak. (The role of explicit knowledge in learning will be discussed in section 2.3.1.) In addition, according to Cumming (1990), writing "elicits an attention to form-meaning relations that may prompt learners to refine their linguistic expression—and hence their control over their linguistic knowledge" (p. 483) (see also Cumming, 1989, for a similar argument). This may be another by-product of the slower pace of writing. Thus, the more time-consuming nature of writing seems to be a facilitative factor in the case of language learning.

The other potential drawback of writing, i.e., it being a more effortful process than speaking, may derive from the fact that learners are likely to be forced to concentrate on improving the grammatical accuracy of their interlanguage, while paying attention to the language system of the target language. Moreover, when writing, they need to communicate without any external help from gestures, listeners or objects around them, which are available in the case of speaking (Sharwood Smith, 1976). However, these may be also interpreted positively, as learners are likely to engage in deeper processing (Craik & Lockhart, 1972), which is likely to heighten their language awareness more than would speaking. Regarding the above points, while emphasizing the importance of teaching writing, Sharwood Smith (1976) even warns that "learning only to speak the target language may actually retard the learning process" (p. 19). Thus, the negative factors of writing compared to speaking, i.e., it being more time-consuming and arduous, can be offset by the advantages offered by engaging in writing.

Furthermore, what is as important, if not more so, as the slower pace of writing is its product, which serves as a permanent record (W. Suzuki, 2012; J. Williams, 2012). Pointing to the nature of writing as process-and-product, Emig (1977) explains that

"information from the *process* is immediately and visibly available as that portion of the *product* already written" (p. 125). That is, unlike speaking, the process of writing offers its content in the form of a visible product simultaneously, without any additional recording device. In a similar vein, regarding the process and product of writing, Wigglesworth and Storch (2012) state that, in writing, "the words appear and remain on the page" (p. 368), which provides a valuable opportunity for learners to analyse their writing, and notice and reflect on any potential errors in it. Similarly, referring to both the slower pace and the products of writing, J. Williams (2012) states that while some claims for the value of writing and for the value of output in general may overlap, "they may be stronger for written production due to the more generous time constraints and permanent record of writing" (p. 323), emphasising that learners are more likely to notice their holes or gaps (Swain, 1998) in the written mode. If Williams' argument is applied to languaging, the expected reflective function of OL may be even stronger in the case of WL, given that the products of WL can be reflected on repeatedly by learners whenever and wherever they are, serving as permanent records (W. Suzuki, 2012). Taken together, the characteristics of writing, i.e., the slower pace and its function as a permanent record, seem to benefit learners.

Finally, the perspectives of Vygotsky (1986) and another contemporary, Luria (1902–1977), seem to provide additional insights into the heuristic nature of writing. Describing writing as written speech, Luria (1999) refers to the two aforementioned characteristics of writing and contends:

> Written speech … assumes a much slower, repeated mediating process of analysis and synthesis, which makes it possible not only to develop the required thought, but even to revert to its earlier stages, thus transforming the sequential chain of connections in a simultaneous, self-reviewing structure. Written speech thus represents a new and powerful instrument of thought. (p. 103)

Similarly, calling writing and speaking written speech and oral speech, respectively, Vygotsky (1986) compares the two and explains that "Written speech is a separate linguistic function, differing from oral speech in both structure and mode of functioning" (pp. 180–181). He further states that writing encourages people to express their ideas more explicitly and elaborately, explaining:

> Written speech is deployed to its fullest extent, more complete than oral speech. Inner speech is almost entirely predicative because the situation, the subject of thought, is always known to the thinker. Written speech, on the contrary, must explain the situation fully in order to be intelligible. The change from maximally compact inner speech to maximally detailed written speech requires what might be called deliberate semantics—deliberate structuring of the web of meaning. (p. 182)

In summary, SCT considers language to be a mediational tool and assumes that learners' externalisation of thoughts with language, i.e., languaging, both oral and written, mediates learning through internalization, which "can be viewed as the outside-in process of development" (W. Suzuki, 2009a, p. 22). In addition, when WL is compared to OL, there are two unique features, i.e., a slower pace and its function as a readily available permanent record, both of which are likely to create opportune conditions for learning.

2.1.1.3 Theoretical Accounts of Oral Self-explaining

From a cognitive psychology perspective, verbal protocols (Ericsson & Simon, 1993) and self-explaining (Chi, 2000) seem to be two major research paradigms on verbalization, i.e., the externalisation of thoughts by using language. While the primary focus of Ericsson and Simon's verbal protocols is to investigate cognitive processes through verbalization as a research tool, that of self-explaining is on learning as a result

of verbalization. Therefore, compared to verbal protocols, self-explaining seems to be closer to the concept of languaging, in that verbalization for self-explaining is assumed to be conducive to learning. Thus, although both are introduced below, this thesis is closer in focus to the research paradigm of self-explaining. (The validity of verbal protocols as a research tool has been questioned because of its potential reactivity (Bowles, 2010), which is discussed in the next section.)

2.1.1.3.1 Verbal Protocols

Verbal reporting procedures, such as protocol analysis, are a method of collecting verbal data introduced by Ericsson and Simon (1993), "where cognitive processes, described as successive states of heeded information, are verbalised directly" (p. 16) for the investigation of learners' cognitive processes. According to Ericsson and Simon, verbal report protocols consist of three levels. Level 1 verbalization is simply the reproduction of information in the form in which it was heeded, i.e., the verbalization of inner thoughts, which are in verbal code. Thus, individuals do not need to make any effort to communicate their thoughts. Meanwhile, "When the internal representation in which the information is originally encoded is not a verbal code, it has to be translated into that form" (p. 18), which is Level 2 verbalization. Therefore, Level 2 verbalization involves "recoding" of internal representation which is not originally in verbal code "into verbal code" (p. 18), such as verbal encoding and vocalization of scents or visual stimuli. In terms of Level 1 and Level 2 verbalization (i.e., nonmetalinguistic verbalization), individuals verbalize "only their thoughts entering their attention as part of performing the task" (p. xiii). Therefore, the sequence of thoughts is expected to be intact and unchanged by the requirements of verbalization (i.e., nonreactive).

In contrast, Level 3 verbalization (i.e., metalinguistic verbalization) involves verbalizing not one's spontaneous thoughts, but required information, such as reasons and explanations for one's actions or particular aspects of a situation that one would not ordinarily attend to, thus influencing and resulting in the changes to the sequence of thoughts (i.e., reactive). Reactive verbalization is expected to act as an extra task, "altering cognitive processes rather than providing a true reflection of thoughts" (Bowles, 2010, p. 14). Therefore, there is a clear difference between Levels 1 and 2 verbalization, where individuals are required to verbalize their thoughts per se, and Level 3 verbalization, where they need to verbalize specific information following instructions.

In light of the framework of Ericsson and Simon (1993), only Level 3 verbalization during task performance is likely to be reactive, resulting in possible verbal overshadowing, i.e., "negative consequences of verbalization" (Ericsson, 2002, p. 984). Admitting this problematic point as a research tool, Ericsson claims that "The reactive effects of 'verbal overshadowing' can be linked to the requirement of producing prescribed types of verbalizations" (i.e., Level 3 verbalization) "and are thus not caused merely by spontaneous verbal expression of one's thoughts" (p. 981) (i.e., Level 1 and Level 2 verbalizations). What is important to the present thesis is that Ericsson also states that reactivity can be positive. In his view, the process of verbalization can strengthen memory traces, which is likely to result in longer-term development. That said, as mentioned above, the focus of Ericsson and Simon's verbal protocols is on the use of verbalization for research purposes, and he does not go into detail regarding the possible positive impact of verbalization on learning.

Meanwhile, with respect to the potential reactivity of Level 3 verbalization, and more broadly think-alouds, some researchers have raised concerns regarding the use of

concurrent verbalization on the ground that it may alter learners' cognitive processes (e.g., R. Ellis, 2001; Jourdenais, 2001). Jourdenais, for instance, warns that learners' thinking aloud while performing a task might act as "an additional task" (p. 373) and change the task itself, thereby affecting their cognitive processes. Along the same lines, R. Ellis questions the validity of think-aloud protocols, arguing that they might result in "dual processing" (p. 14), thus overly taxing learners' attentional resources. From the need to identify the validity of concurrent think-aloud protocols as a research tool, a new strand of SLA research on reactivity emerged.

Starting from the first L2 study on reactivity by Leow and Morgan-Short (2004), reactivity research has been growing and the number of studies has been increasing (e.g., Bowles, 2008; Bowles & Leow, 2005; Rossomondo, 2007; Sachs & Polio, 2007; Yanguas & Lado, 2012). In her book published in 2010, Bowles conducted a meta-analysis including over ten reactivity studies that used verbal tasks and reported an overall "small effect" (p. 110), that "is not significantly different from zero" (p. 138), for think-alouds across tasks. That said, referring to various factors that are likely to affect the outcomes, she was cautious to conclude her book by saying that "the answer to the question of reactivity and think-alouds is not a simple 'yes' or 'no' but rather it is dependent on a host of variables" (p. 110).

2.1.1.3.2 Oral Self-explaining

In contrast to verbal protocols (Ericsson & Simon, 1993), self-explaining (Chi, 2000), another research paradigm on verbalization that involves learners' externalisation of thoughts by producing language, focuses on its impact on learning and views it as a strategy for effective learning and teaching (see Chi, 2000, for details). Self-explaining is usually defined as an activity involving "making sense of new information by

explaining to oneself" (Chiu & Chi, 2014, p. 91). Therefore, in the framework of Ericsson and Simon's verbal protocols, self-explaining can be considered as Level 3 verbalization, which is expected to trigger cognitive changes. Although not fully examined in the field of SLA, the impact of self-explaining has been investigated in a wide variety of non-SLA domains, such as biology, physics and maths since the late 1980s (e.g., Chi et al., 1989; McNamara, 2004, 2017). From the findings of self-explaining studies (e.g., Chi et al., 1994), at least three cognitive processes seem to account for the facilitative effect of self-explaining on learning, inference generation (Chi et al., 1994; Siegler, 2002), mental-model repair (Chi, 2000; Chi et al., 1994), and integration (Rittle-Johnson & Loehr, 2017).

First, as a "constructive inferencing activity" (Chi et al., 1994, p. 441), self-explaining is expected to enable learners to construct knowledge inferences. More specifically, in Chi et al.'s view, self-explaining encourages learners to infer new information that is missing from materials based on information included in those materials, leading to new information being encoded into memory and made available to facilitate subsequent performance. In line with their view, Siegler (2002) contends that "self-explanations are inferences about causal connections among objects and events … concerning 'how' and 'why' events happen" (p. 37).

Second, mental-model repair is another process that is likely to account for the positive impact of self-explaining on learning (Chi, 2000). That is to say, self-explaining is expected to enable learners to repair their flawed mental models when they encounter conflicts between new information in instructional materials and their prior knowledge, providing them with an "opportunity to self-repair" (Chi et al., 1994, p. 471) and helping them revise their understanding. Thus, self-explaining integrates new and repaired existing knowledge, enabling learners to encode this new integrated knowledge

into memory. The process of mental-model repair seems to be compatible with the noticing function of the Output Hypothesis (Swain, 2005), in that both claim to provide learners with opportunities to address their linguistic issues when they notice and confront issues with their language use (i.e., output and self-explaining.)

Third, related to mental-model repair, "knowledge integration" is another process that is assumed to contribute to the positive impact of self-explaining on learning. To be more precise, self-explaining is expected to facilitate learning by connecting pieces of new information together or new information and learners' prior knowledge (Rittle-Johnson & Loehr, 2017). In terms of the processes of mental-model repair and knowledge integration, self-explaining is expected to assist learners in monitoring and controlling their thinking, resulting in greater learning and deeper understanding. As stated above, mental-model repair and knowledge integration are related in that both of them involve the integration of knowledge. They differ, however, in that the former assumes learners' incorrect mental models, whereas the latter does not.

In addition to these three mechanisms that are inherent to self-explaining, the use of two general cognitive processes is also expected to contribute to learning (Siegler, 2002). First, self-explaining is likely to induce greater depth of processing, "where greater 'depth' implies a greater degree of semantic or cognitive analysis" (Craik & Lockhart, 1972, p. 675), as learners need to think about materials deeply in order to generate explanations (i.e., make inferences and repair mental models). According to Craik and Lockhart, the strength of a long-term memory representation depends on the depth of processing with which the information was initially encoded. Compared to shallow processing, deep processing is assumed to yield more elaborate, durable and stronger memory representations.

Second, the beneficial effects of self-explaining are expected to be attributable to the generation effect proposed by Slamecka and Graf (1978), who identified that learners recall and/or recognize the items they generate better than items they simply read at a later point in time. Given that self-explaining requires learners to generate explanations, the generation effect is hypothesised to apply to learning via self-explaining as well. Supporting the claim of Slamecka and Graf, Chi (2000) states that a generation effect and a self-explanation effect, i.e., students learning better when they explain material to themselves (Chiu & Chi, 2014, p.92), can be expected when learners address constructive activities, further explaining that they are effective because "being generative means one is being more attentive and actively laying down memory traces" (Chi, 2000, p. 173).

In summary, engaging in self-explaining is expected to enhance the three cognitive processes inherent to self-explaining, i.e., inference generation (Chi et al., 1994; Siegler, 2002), mental-model repair (Chi, 2000; Chi et al., 1994), and integration (Rittle-Johnson & Loehr, 2017). In addition, self-explaining is likely to induce two general cognitive processes, i.e., greater depth of processing (Craik & Lockhart, 1972) and the generation effect (Slamecka & Graf, 1978), all contributing to learning.

2.1.1.4 Theoretical Accounts of Written Self-explaining

As for self-explaining in the written mode, learning is likely to occur involving all the aforementioned mechanisms. Namely, learners are expected to generate inferences regarding instruction materials, connect/integrate pieces of information and revise their mental models if necessary, resulting in deeper task engagement and greater learning, thanks to greater depth of processing (Craik & Lockhart, 1972) and the generation effect (Slamecka & Graf, 1978). In addition to these internal mechanisms that are processed

inside learners' minds, what is unique about written self-explaining is the function of external memory, i.e., "records … that are maintained in repositories that are external to their users" (Hertel, 1993, p. 665). When learners engage in written self-explaining, its products turn into objects that learners can reflect on at later points in time (i.e., external memory), thus relieving the demands on working memory (Baddeley, 2003). Furthermore, written self-explaining usually does not require learners to process information and produce language simultaneously as in oral self-explaining, resulting in less time pressure, which is also likely to relieve demands on working memory (Muñoz et al., 2006). Thus, although there are slight differences between interpretations of written self-explaining and WL in terms of "product" and "less time pressure" (cognitive psychology explains with working memory, whereas SCT explains with mediation and artefacts), what is stated about written self-explaining seems to be compatible with WL as well.

In summary, both oral and written self-explaining assume that the act of explaining something, such as understanding, problems or questions, to oneself triggers cognitive processes, such as inferencing and greater depth of processing, leading to greater learning as a consequence. As for written self-explaining, being consistent with the features of WL, the presence of external memory and a slower pace are its additional benefits. As reviewed above, theoretically, "languaging" and "self-explaining" are not equivalent because SCT assigns languaging a central role in learning, whereas internal cognition is the locus of learning from a cognitive perspective. To borrow W. Suzuki's (2009a) words again, learning from languaging may be considered as "the outside-in process" of development (p. 22), whereas learning from self-explaining might be viewed as "the inside-out process" of development. In spite of these differences in stance, self-explaining and languaging share similar concepts, in that both of them

assume learning derives from the externalisation of thoughts via language. Therefore, self-explaining seems to fall into Swain's (2010) definition of languaging, i.e., "the act of using language to mediate cognition—to bring thinking into existence" (p. 115). As such, the current thesis referred to the literature on both SCT and cognitive psychology.

2.1.2 Research Findings for OL and WL

2.1.2.1 Research Findings for OL

To date, SLA studies have produced ample evidence to support Swain's (2006) claim that OL can contribute to learners' L2 learning (e.g., Brooks et al., 2010; Swain & Lapkin, 2007; Swain et al., 2009; Yang, 2016). Although terms such as "metatalk" (e.g., Storch, 2008; Swain, 1998), "collaborative dialogue" (e.g., Ammar & Hassan, 2017; M. Suzuki, 2008; Swain & Lapkin, 1998), and "private speech" (e.g., Negueruela & Lantolf, 2006; Ohta, 2000) are used instead of "languaging" in many of these studies, researchers have unanimously found that OL is a facilitative mediator of cognitive development. It is also worth noting that OL is not necessarily the focus of all of these studies.

In addition, studies that have employed OL in the form of think-alouds and stimulated recalls as a research tool have generally reported a positive impact of such language use on learning (e.g., Nabei & Swain, 2002; Qi & Lapkin, 2001; Sanz, Liu, Lado, Bowden, & Stafford, 2009; Yanguas & Lado, 2012, but see Bowles & Leow, 2005; Sachs & Polio, 2007 for the reactivity of verbal protocols). A list of these studies is presented in Table 2.1, below, and the sections to follow present a more detailed discussion of their findings.

Table 2.1
List of Studies Using OL as a Learning or Research Tool

Topics	Studies
Collaborative dialogue	Ammar & Hassan, 2017; Kim & McDonough, 2008; Storch, 2001, 2002; Storch & Wigglesworth, 2010; M. Suzuki, 2008; Swain & Lapkin, 1998, 2002; Wigglesworth & Storch, 2012
Metatalk	Storch, 2008; Swain, 1998
Private speech	Negueruela & Lantolf, 2006; Ohta, 2000, 2001; Yoshida, 2008
Think-aloud, immediate report, stimulated recall	Adams, 2003; Bowles & Leow, 2005; Egi, 2007; Nabei & Swain, 2002; Qi & Lapkin, 2001; Rossomondo, 2007; Sachs & Polio, 2007; Sachs & Suh, 2007; Sanz et al., 2009; Swain & Lapkin, 2007; Yanguas & Lado, 2012
Oral languaging (OL)	Brooks & Swain, 2009; Brooks et al., 2010; Knouzi et al., 2010; Lapkin, Swain, & Knouzi, 2008; Swain et al., 2009; Yang, 2016

2.1.2.1.1 OL Intended as a Research Tool

SLA researchers have used OL as a research tool to investigate learners' cognitive processes, such as noticing and reflection (e.g., Qi & Lapkin, 2001; Storch, 2008). In addition, as mentioned earlier, in order to probe the validity of verbal protocols (i.e., OL) as a research tool (Ericsson & Simon, 1993), some researchers started to address the issue of reactivity by investigating the impact of think-alouds on L2 learning (e.g., Bowles & Leow, 2005; Yanguas & Lado, 2012). Some of these studies are reviewed below.

Focusing on the reflective function of output, Storch (2008) examined learners' metatalk in order to investigate the relationship between their level of engagement and L2 learning. The learners were instructed to reflect on their output by verbalizing their linguistic problems. The quality of engagement was measured on the basis of the content of verbalization; deliberation over the language items was referred to as elaborate engagement, and lack of deliberation as limited engagement. Although it was

found that engagement at both levels led to learner development in general, elaborate engagement was identified to have led to a better understanding of the target structures than limited engagement. Thus, pointing to the reflective function of learners' metatalk, Storch argues that its effect on L2 learning is not uniform, but can vary depending on learners' level of engagement. In addition, pointing to the result that engagement at both levels led to learners' learning, Storch states that "learners … may benefit from the opportunity to verbalise and deliberate about language" (p. 111), suggesting the very opportunity of languaging might have benefited the participants.

The facilitative impact of OL is also reported by Qi and Lapkin (2001), who employed think-alouds in order to examine the quantity and quality of noticing during and after a three-stage writing task and to explore whether noticing contributed to language learning. The quality of noticing was coded as perfunctory (i.e., noticing only) or substantive (i.e., noticing with reason), which seems to be compatible with the two levels of awareness, i.e., noticing and understanding, respectively, postulated in Schmidt's (1990) Noticing Hypothesis. The participants were two adult native speakers of Mandarin (one high-intermediate and the other low-intermediate in proficiency) who engaged in the following tasks individually. They first described a picture in writing and then proofread it while thinking aloud (Stage 1). The drafts were collected and reformulated by the researchers who corrected all the linguistic errors without changing the participants' original ideas. Four days later, each participant was given his or her original draft along with its reformulated version and was instructed to compare the two while thinking aloud (Stage 2) and participated in an immediate retrospective interview with the researcher. One week later, they revised their original text (Stage 3). They were allowed to use Mandarin, their first language (L1), for thinking aloud.

The analysis of think-alouds revealed that problems found in producing writing could trigger noticing of forms in the subsequent input and that noticing could contribute to the improvement of subsequent output, providing evidence in support of the Output Hypothesis (Swain, 2005) and the Noticing Hypothesis (Schmidt, 1990). It is important to note that the quality of noticing, perfunctory (noticing only) or substantive (noticing with reason), was identified to have a direct impact on the extent of improvement in subsequent output. In addition, differences in the quality of noticing were also found to be influenced by proficiency level. To be more specific, the more proficient learner demonstrated a larger proportion of substantive noticing than the less proficient learner. In contrast, with respect to the quantity of noticing, the results were rather mixed. That is, the more proficient learner produced considerably more instances of noticing in Stage 1, but the less proficient learner produced slightly more instances of noticing in Stages 2 and 3. In terms of the contribution of noticing to learning, 71% of substantive noticing of both participants in Stage 2 resulted in improvements in Stage 3. Not surprisingly, the ratios were much lower for LREs labelled as perfunctory (38% and 17% for the more and less proficient learners, respectively).

With respect to the facilitative impact of their think-aloud on L2 learning, supporting the statement of Storch (2008), Qi and Lapkin (2001) suggest that verbalization itself might be an effective strategy. Furthermore, they state that "asking learners about what they are thinking and the rationale for their grammatical decisions can promote metacognitive processing and lead to effective problem solving" (p. 296). The statement seems to be compatible with the notion of self-explaining (Chi, 2000) or Level 3 verbalization (Ericsson & Simon, 1993), in that the externalisation of thoughts with language is assumed to trigger cognitive change.

Let us turn to four reactivity studies that have also provided additional insights into the potential role of OL (Bowles & Leow, 2005; Sachs & Polio, 2007; Sanz et al., 2009; Yanguas & Lado, 2012). Partially replicating Qi and Lapkin's (2001) study, Sachs and Polio (2007) produced rather different findings from the original study in one of their two experiments. Sachs and Polio employed think-aloud to investigate learners' attentional processes in order to identify the effectiveness of two types of written feedback (written error corrections and reformulations) on an L2 writing revision task. The participants were 15 adult English as a second language (ESL) students with a high-intermediate proficiency level. They were divided into three groups depending on the experimental conditions: error correction, reformulation and reformulation plus think-aloud in English (i.e., their L2).

In contrast to Qi and Lapkin's (2001) findings, the analysis of the participants' think-aloud protocols revealed that the participants produced LREs of a shallower kind (84%) rather than those at a deeper level (15%). Also contradicting Qi and Lapkin's findings, Sachs and Polio (2007) found that the level of awareness (noticing with reasons and/or metalinguistic terms vs without reasons) did not have an impact on L2 learning. Furthermore, and more importantly, the think-alouds were found to be reactive. To be more precise, the participants in the reformulation condition produced significantly more accurate revisions than those who were in the reformulation plus think-aloud condition. Based on these results, Sachs and Polio warn that think-aloud protocols should be employed and interpreted with care. However, as the researchers admit, the use of L2 might have contributed to findings that were inconsistent with Qi and Lapkin's study, where the participants were allowed to use their L1. In addition, as pointed out by Bowles (2010), the lack of inclusion of a target construction could have

affected the measurement of L2 learning, which might have contributed to the observed reactivity. Given these factors, the results have to be interpreted with caution.

In line with the findings of Sachs and Polio (2007), the reactivity of think-alouds was also reported on by Bowles and Leow (2005). They addressed the issue of reactivity of both metalinguistic and nonmetalinguistic think-alouds (i.e., Level 3 verbalization, and Levels 1 and 2 verbalization, respectively, in Ericsson and Simon's (1993) framework) on text comprehension as well as learning of the target structure (i.e., the pluperfect subjunctive). The participants were 45 advanced learners of Spanish, who were randomly assigned to one of two think-aloud groups (i.e., metalinguistic or nonmetalinguistic) or to a control group. After reading a text embedded with the target structure, all the participants worked on the same comprehension and written production tasks. As predicted by Ericsson and Simon, no reactivity was evidenced in any comparisons between the control group and the nonmetalinguistic group (who engaged in Level 1 and Level 2 verbalization). Likewise, as predicted by Ericsson and Simon, Bowles and Leow found a significant difference between the two experimental groups (i.e., metalinguistic group and nonmetalinguistic group) in terms of their comprehension, producing evidence of reactivity of metalinguistic verbalization (i.e., Level 3 verbalization).

Interpreting the outcomes, Bowles and Leow (2005) concluded that it must have been difficult even for advanced learners to read and follow the meaning of the text while thinking out loud their thoughts and justifications. As proof of their conclusion, they introduced one of the participants' think-aloud protocols, which was "…I don't know what any of this is going on, because I'm really not paying attention. This is distracting to have to talk while I'm doing this" (p. 430). This seems to indicate that "tasks should be carefully selected on the basis of their compatibility with thinking

aloud" (Leow, Grey, Marijuan, & Moorman, 2014, p. 115). The most important finding for the current thesis, however, is that, although not reflected in the results of the tests, the participants in the metalinguistic condition generally demonstrated some awareness of the function of the unfamiliar target structure, whereas only the high-scorers in the nonmetalinguistic condition did so. As a plausible explanation for the result, Bowles and Leow state that the requirement to verbalize justifications of the production tasks might have enabled the participants to have a certain level of awareness, lending support to the claim of both Storch (2008) and Qi and Lapkin (2001) regarding the facilitative impact of OL on raising awareness.

In contrast to the above two studies by Sachs and Polio (2007) and Bowles and Leow (2005), positive reactivity has been reported in two more recent studies by Sanz et al. (2009) and Yanguas and Lado (2012) (see also Rossomondo, 2007, for positive reactivity). Sanz et al. conducted two experiments on reactivity with the same design, except for the explicitness of the treatment. In the first experiment, 24 English-speaking college students were divided into a think-aloud group and a silent group. The target structure was the case system of Latin, a case-marking language, unlike English. The participants, who had no prior knowledge of Latin, i.e., "naive learners of Latin" in the researchers' words, received a computerised treatment consisting of an explicit grammar lesson, practice and feedback. After the lesson, the participants in both groups took three kinds of tests, which did not show any statistically significant difference between the two groups, indicating no reactivity. The second experiment was conducted with the same design but with a less explicit version of the treatment, i.e., no inclusion of an explicit grammar lesson on the target construction. This time reactivity was found. In contrast to the first experiment, the participants in the think-aloud group scored

significantly higher than those in the silent group on the tests, indicating positive reactivity in their language development.

With respect to the inconsistent outcomes for the two experiments, Sanz et al. (2009) point to the differences in the nature of the treatment, i.e., differing level of explicitness, in the experiments. That is, in Experiment 1, the participants did not seem to have any room to enhance their awareness, thanks to the explicit instruction, because "pedagogically, every effort was made to level the field for the participants" (p. 65). Whereas in Experiment 2, the participants had to raise their awareness to compensate for the lack of an explicit grammar lesson (i.e., less explicit instruction), which "left learners more to their own devices" (p. 65). Given that the think-aloud group outperformed the silent group, think-aloud is likely to have functioned as part of their own devices. In addition, referring to the Output Hypothesis (Swain, 2005), the researchers explain that learners' verbalization (i.e., output) is likely to have facilitated their learning (p. 64), resulting in the provision of extra input and leading to the development of awareness as a consequence. Based on the results, Sanz et al. conclude that "requiring learners to perform think-aloud protocols has the potential to alter the very process they are meant to reflect" (p. 63), which is not a welcoming conclusion for think-aloud as a research tool, but additional support for it as a learning tool.

Similar results were obtained by Yanguas and Lado (2012), who investigated reactivity with a semi-guided writing task. The participants were 37 college students enrolled in classes called "Review of Oral and Written Spanish for Native Speakers Educated in the United States." As the title of the course suggests, the participants were bilinguals, in Spanish and English, in the United States whose heritage language was Spanish. According to the researchers, however, heritage language learners are generally better at processing language aurally than in written mode, which entails the use of

metalinguistic and explicit knowledge of the language. The participants were divided into two groups depending on the requirements of think-alouds, and wrote a story based on comic strips. The participants in the think-aloud condition were instructed to verbalise their thoughts while working on the writing task (i.e., nonmetalinguistic verbalization), whereas those in the non-think-aloud condition addressed the same task silently. Their language development was measured based on fluency (measured by number of words and number of words per T-unit), accuracy (measured by error-free T-units) and lexical variety. The analysis of the participants' writing in the think-aloud group revealed positive reactivity in terms of accuracy, but not fluency. (An almost statistically significant difference was found regarding lexical variety.)

Based on the results, Yanguas and Lado (2012) refer to three important points. First, regarding the improvement in accuracy, they argue that the "participants may have been made aware of the syntactic structure of their writing by simply speaking aloud" (p. 392), supporting the claim of the noticing function of the Output Hypothesis (Swain, 2005). In addition, like Sanz et al. (2009), based on the Output Hypothesis, they state that not just the process but the participants' think-alouds, i.e., intangible products, might have benefited the participants, providing them with additional input and expanding the opportunities to notice their linguistic issues, echoing the idea that OL/think-aloud benefits learners through its process and product. Second, also pointing to the positive reactivity, the researchers contend that verbalization benefits learners regardless of their language proficiency, such as naive learners in the study by Sanz et al. (2009) and native speakers in their own study. Third, regarding the results that positive reactivity was identified only in terms of accuracy, whereas no reactivity was found for fluency, Yanguas and Lado report that "reactivity depends on the type of measure investigated" (p. 391), supporting Bowles's (2010) statement that reactivity is

dependent on a host of variables and indicating the importance of employing multiple measures of learning (Norris & Ortega, 2003).

2.1.2.1.2 OL as a Platform for L2 Learning and as a Learning Tool

The studies reviewed above generally report a positive impact of OL on L2 learning (except for Bowles & Leow, 2005; Sachs & Polio, 2007), but they used OL as a research tool and none of them focused exclusively on its role as a learning tool. In contrast, some SLA studies, including the nine studies reviewed below, demonstrate that OL could work as a learning tool for L2 learners. Although the first four studies label OL differently, such as "metatalk" (Swain, 1998), "verbalizations" (Swain & Lapkin, 2007) and "collaborative dialogue" (Ammar & Hassan, 2017; Swain & Lapkin, 2002), their focus was, nevertheless, on OL. The latter five studies were conducted using the term "OL" (Brooks et al., 2009; Brooks & Swain, 2010; Knouzi et al., 2010; Swain et al., 2009; Yang, 2016).

For example, in Swain's (1998) study, the participants were 48 students in two classes of a French immersion programme. They were instructed to engage in metatalk (i.e., OL), which was explained as "talk about the language of the text they were reconstructing" (p. 70), while they worked on dictogloss in pairs. Three dictoglosses were prepared and the participants completed one dictogloss per week for three weeks. For the first two weeks, lesson and practice were conducted to familiarise the participants with the task. Only the data from the third week was analysed by focusing on LREs. The analysis of the LREs demonstrated that the learners solved many of their linguistic problems by metatalking with each other, enabling Swain to conclude that OL contributes to L2 learning.

Similarly, investigating the impact of learners' verbalization (i.e., OL) on their problem-solving, Swain and Lapkin (2007) identified that OL facilitates the comprehension of learners, contributing to L2 learning. As in Swain (1998), the participants were four students in the immersion programme in Canada who were instructed to think aloud while they worked on dictogloss. Two of them worked in a pair and the other two worked individually. The researchers reported that one learner reached an understanding of problematic phrases through writing them down and verbalizing what she wrote, saying, "Oh, I get it now!" Another learner was found to have talked herself into understanding that she did not understand, which may be interpreted as noticing the hole and seems to be "an important stimulus for noticing the gap" (Swain, 1998, p. 66). Although admitting that not all the verbalisation in the study led to problem-solving and learning, Swain and Lapkin concluded nevertheless that "one way in which language is acquired is through use: by producing language we can find out what it means, and of what it consists" (p. 316).

In another Swain and Lapkin's (2002) small-scale 5-stage study, two adolescent French immersion learners were instructed to write a story (Stage 1), compare their story and its reformulated version and notice differences between them (Stage 2), both talking collaboratively. The next week, they participated in a stimulated recall session, followed by individual posttest (Stage 4) and interview (Stage 5). Analysis of the participants' collaborative dialogue from Stages 1–3 revealed that they resolved their linguistic issues by "talking it through" (i.e., OL), providing empirical evidence that OL facilitates L2 learning. Moreover, a closer examination of the learners' collaborative dialogue in Stage 2 revealed that they accepted 65% of the reformulations, but rejected 35% of them (p. 294), demonstrating that learners' beliefs and attitudes toward feedback (i.e., reformulation) affect the uptake of feedback (their reaction to feedback).

Interestingly, however, it was found that they responded correctly at the same rate (three-quarters) to both accepted and rejected reformulations. Pointing to this result, the researchers state that "both acceptance and rejection of the reformulator's changes led to 'talking it through"' (p. 298), which is likely to have enabled each learner to individually draw on the knowledge they had previously jointly constructed, resulting in the internalization of new knowledge.

More recently, Ammar and Hassan (2017) investigated the effect of collaborative dialogue (i.e., OL) on L2 learning and its relationship with target linguistic features (four French grammar morphology forms) and learners' proficiency levels with a pretest-intervention-posttest design. The participants were 79 Arabic-speaking learners of L2 French in four Grade 5 and 6 classes (two classes each), who were assigned to low- and high-proficiency subgroups based on pretest results as well as their teachers' judgement. The treatment task employed was zero-error dictation, dialogue-driven dictation. Like dictogloss, "its main objective is to make learners verbalize their language conceptions while writing" (p. 7). Unlike dictogloss, however, learners are instructed to write a sentence which a teacher dictates one word at a time, asking the teacher questions about any linguistic uncertainty or issues that they may have and talking with their peers (i.e., OL) in order to write a zero-error sentence. Thus, it is learners who initiate all focus-on-form episodes because "producing an error-free text is the objective, which motivates learners to ask questions and to pay attention to all the provided explanations" (p. 8). Of the four classes, two were assigned to an experimental condition (zero-error dictation), whereas the other two were assigned to a comparison condition where traditional dictation was conducted (without any collaborative dialogue). Both groups participated in weekly sessions for five weeks.

Analysis of the tests results produced three findings. First of all, the participants in the experimental group who engaged in collaborative dialogue were found to have outperformed their counterparts in the comparison group on a posttest, indicating the positive impact of OL. Interpreting the results, Ammar and Hassan (2017) point to "different built-in characteristics of collaborative dialogue" (p. 22), further stating that "in such dialogues, learners identify their own needs, act on them, and participate in the construction of the scaffold that assists them through the zone of proximal development" (pp. 22–23). In addition, they point to the abundance of metalanguage and the timing of its use as possible factors that might contribute to the effectiveness of collaborative dialogue. (These points are discussed in more detail in section 3.2.)

Second, analysis of the gain scores of the experimental group demonstrated that the participants' degree of learning varied across their proficiency levels. More specifically, lower-proficiency participants showed greater gains than higher-proficiency ones, while the gain scores between high- and low-proficiency learners were similar in the comparison group. The results indicate that "collaborative dialogue was beneficial for all students" (p. 25), but probably more so for lower-proficiency learners. Regarding the gap between the two levels, Ammar and Hassan (2017) explain that repeating the task over five weeks and engaging in collaborative dialogue in these sessions might have benefited lower-proficiency learners more. Supporting their speculation, most of the lower-proficiency participants gave positive comments regarding collaborative dialogue in interviews conducted after the experiment. In contrast, there were some higher-proficiency participants who expressed reservations in terms of the content of the collaborative dialogue, stating that it was of little use "because it did not offer them any information that they did not already know" (p. 26). Based on these comments, the researchers suggest that a task might have negative

effects depending on learners' proficiency levels, "if these learners fail to perceive the learning benefits of the task" (p. 26).

Third, a close examination of the gain scores of the experimental group also indicated that the participants' degree of learning varied across the four linguistic target structures. Albeit equally complex, Ammar and Hassan (2017) identify that the learning gains of one of the four types, i.e., the determiner-predeterminer agreement type, statistically surpassed the other target types. Regarding this outcome, the researchers state that the invariable nature of the determiners targeted in their study turned out to be simpler than the remaining three agreement types in their operationalization, resulting in different outcomes across the four target types.

As stated earlier, although the term "OL" is not used, the four studies reviewed above indicate that OL can be a platform for L2 learning. Meanwhile, the five OL studies introduced below were conducted with the term "languaging" (i.e., OL), indicating that OL can facilitate L2 learning. The first three studies to be reviewed addressed learners' individual OL (Brooks et al., 2010; Knouzi et al., 2010: Swain et al., 2009), while the latter two were conducted focusing on collaborative OL (Brooks & Swain, 2009; Yang, 2016).

Swain et al. (2009) examined the impact of OL on learning the concept of voice in French, based on Negueruela's (2008) study, where participants were assigned languaging tasks as their homework (Negueruela's languaging tasks comprised self-explaining three concepts taught in a Spanish class based on Gal'perin's (1992) STI.) Nine university students were instructed to read cards on the concept prepared by the researchers and explain the concept orally (i.e., OL). Five key metalinguistic terms (active voice, passive voice, middle voice, agent and patient) were provided to prompt the participants. A pretest was conducted to assess the participants' existing knowledge

of the grammatical concept of voice, which demonstrated that none of them were familiar with the concept or did not know the metalinguistic terms. The nine participants were divided into three groups (i.e., high-, middle-, low-languagers) depending on the amount of languaging they produced in the languaging stage. Their languaging units (LUs), "the cognitively complex on task talk arising from the explanatory text" (p. 10), were investigated qualitatively in relation to their scores on immediate and delayed posttests.

The LUs on the target concept were coded into three types: paraphrasing, inferencing and analysing. In addition, LUs that were not related to the concept, but seemed to have helped the participants, were further coded into two types: self-assessment and rereading. Although all the participants were found to have reached a better understanding of the target concept judging from the results of the immediate and delayed posttests, the researchers identified that, in general, the high-languagers performed better than middle- and low-languagers. Given that all the participants had similar levels of prior knowledge regarding the concept at the outset of the experiment, the results seem to indicate a positive correlation between the amount of OL and L2 learning. In addition, it was found that the languaging of high-languagers was different from that of middle- and low-languagers, not just quantitatively but also qualitatively. That is to say, high-languagers were found to have produced more inference and self-assessment LUs, attesting to Chi et al.'s (1994) claim concerning the mechanism of self-explaining as a trigger to make inferences. Moreover, self-assessment LU seems to run in parallel to the reflective function of output claimed by the Output Hypothesis (Swain, 2005), as well as the process of mental-model repair proposed by Chi (2000), which is expected to facilitate learners' monitoring and controlling of their own thinking.

As the amount of OL was found to have influenced the results of the tests in Swain et al.'s (2009) study, scrutinizing the data of two participants from the study, Knouzi et al. (2010) analysed the quality of a high- and a low-languagers' LUs. The analysis revealed that the high-languager self-scaffolded more by connecting new and prior knowledge, being examples of mental-model repair (Chi, 2000) and integration (Rittle-Johnson & Loehr, 2017), whereas the low-languager was hesitant and often reverted to thinking before languaging, presumably not understanding the "talking-it-through" aspect of languaging and not benefiting from it much as a consequence. Referring to the results, Knouzi et al. state that the discrepancies might reflect differences in the learners' ZPD.

Similarly, focusing on the data of two middle-languagers from Swain et al.'s (2009) study, Brooks et al. (2010) explored the role of OL in mediating learners' understanding of the grammar concept, making use of Vygotsky's (1986) distinction between spontaneous (everyday) and scientific (also often referred to as "academic") concepts. Two participants were chosen from the pool of five middle-languagers because their performance on delayed-posttests showed discrepancies between a written test (i.e., a limited production fill-in-the-blanks test) and an open-ended oral test (i.e., stimulated recall). (They scored the highest in the written test, but the lowest in the oral test.) Explaining the gap, Brooks et al. point to the differences in the nature of the two tests. That is, the highly-controlled written test probably did not require the use of more advanced grammatical knowledge, enabling the participants to treat the task as assessing spontaneous concepts which they already possessed. On the other hand, they needed to deploy their newly gained scientific concept in the case of the oral test (i.e., stimulated recall).

Although not raised by the researchers, the results may also be explained by developmental differences in the two dimensions of declarative and procedural knowledge employed in this thesis. Namely, the participants might have gained declarative knowledge of the concept but not have achieved a sufficient level of procedural knowledge to apply that knowledge in an open-ended oral test. Despite the different outcomes for the two types of tests, Brooks et al. (2010) contend that the mediating role of OL enabled the participants to make progress from no knowledge of the concept to a developing understanding, describing the participants' conceptual development observed during the stimulated recall as "development in progress" (p. 106).

As stated, the above three studies examined learners' individual OL, while the two OL studies below were conducted focusing on collaborative OL (Brooks & Swain, 2009; Yang, 2016). In one of the two studies, Brooks and Swain (2009) investigated the impact of different sources of expertise (i.e., peer, reformulation, researcher) on the process of learning. The participants were four adult ESL learners in Canada at an intermediate level of proficiency, and they worked on a picture description task in pairs. They were given a picture prompt and instructed to describe in writing what happened in the picture (collaborative writing task), while talking collaboratively (i.e., OL). After each pair finished writing a short story, they were given a reformulation of their story and instructed to discuss any changes which they noticed between their story and the reformulation (i.e., OL) (noticing task). The participants' OL during the writing task and the noticing task was recorded by a video camera and a tape-recorder. The recordings of the noticing task were employed in the next phase, i.e., augmented stimulated recall, where one of the researchers viewed the video of the noticing task with each pair and discussed what the participants noticed (i.e., OL).

One week later, a posttest was conducted, i.e., the participants were given a clean typed copy of their original story (without any reformulations) and asked to revise it individually. The recorded collaborative dialogue (i.e., OL) was coded depending on its focus (i.e., form or lexis), as well as on its degree of correctness (i.e., correct, incorrect and unresolved). When these OL episodes were examined, both pairs were found to have produced more form-based OL episodes than lexis-related ones. It was also identified that both pairs resolved most of the OL episodes correctly, with their peers, reformulations and the researcher functioning as expertise. Most importantly, an examination of the posttest (revision) results revealed that the participants' OL had a positive impact on learning, supporting Swain's (2006) claim that languaging can function as a learning tool.

More recently, Yang (2016) examined the effects of collaborative dialogue (i.e., OL) on revisions of collaboratively written stories in her six-week experiment. Instead of reformulations, model texts were employed for revisions in this study. The participants, four Chinese EFL university students with differing proficiency levels (one higher-, two middle-, one lower-proficiency) addressed a three-stage writing task (i.e., composing, comparing with a model, and revising) in pairs for two weeks. After reading a short story, they were instructed to rewrite the story from a different perspective each week.

Four findings were reported. First of all, it was found that most of the linguistic issues were correctly solved through collaborative OL and that OL was identified to have had a positive impact on the participants' revision. In addition, echoing the findings of the aforementioned study by Brooks and Swain (2009), the participants were found to have used the original story, the models and their peers as expertise. Second, unlike the Brooks and Swain study, the participants talked more about content-related

issues than linguistic-related ones, producing more content-related episodes (CREs) than LREs, probably because the task (i.e., rewriting the story from a different perspective) required the participants to think more about the content than the picture description task used by Brooks and Swain.

Third, L2 proficiency differences were found to influence the focus, but not the amount, of OL. That is to say, when a middle-proficiency participant paired with a lower-proficiency participant, they spent more time clarifying their understanding regarding the original story, producing more CREs. Conversely, when a middle-proficiency participant worked with a higher-proficiency participant, they focused more on language (i.e., linguistic features), producing more LREs. Not surprisingly, it was also found that the differences in proficiency levels influenced the rate of problem-solving when they faced linguistic issues. Namely, a middle-proficiency participant solved more problems when they worked with a higher-proficiency participant than with a lower-proficiency participant. Fourth, an important point for this thesis is that OL "occurred not only in peer interaction but also between learners and readings" (p. 252), i.e., individually. This finding is consistent with Swain and Watanabe's (2013) aforementioned statement that there are two types of languaging, interpersonal communication and intrapersonal communication.

2.1.2.1.3 Summary of OL Research

As briefly reviewed above, although OL is labelled differently in some studies, the following five findings emerge regarding OL. First and foremost, OL appears to have a positive impact on learners' L2 learning (e.g., Ammar & Hassan, 2017; Qi & Lapkin, 2001; Storch, 2008; Swain & Lapkin, 2007; Swain et al., 2009; Yang, 2016). However, some studies of reactivity suggest that OL can also negatively affect language

development, depending on the variables, such as the language of the report (L1 or L2) (e.g., Sachs & Polio, 2007), the compatibility of tasks and OL (e.g., Bowles & Leow, 2005) and the types of measures used (e.g., Yanguas & Lado, 2012). Second, OL facilitates L2 learning in both collaborative and individual situations (e.g., Swain & Lapkin, 2007; Yang, 2016). Third, learners' proficiency levels can influence the quality of OL, but not so much the amount of OL (e.g., Qi & Lapkin, 2001; Yang, 2016). In particular, compared to lower-proficiency learners, more proficient learners tend to produce OL with a higher level of awareness, resulting in better performance. Also, higher-proficiency learners tend to focus more on linguistic features than on planning content. That said, it is worth mentioning that the amount of OL can vary even when learners' proficiency levels are comparable (Swain et al., 2009), probably depending on factors, such as their ZPD (Knouzi et al., 2010) and individual differences in aptitude and attitude towards tasks (Ammar & Hassan, 2017; Swain & Lapkin, 2002).

Fourth, the amount of OL seems to correlate with learning (e.g., Swain et al., 2009). The more OL that learners engage in, the greater learning they are likely to achieve. Last but not least, the findings regarding the link between the quality of OL measured by the level of noticing (Schmidt, 1990) and the degree of L2 learning are rather mixed. Although higher quality of OL was reported to result in greater L2 learning by Qi and Lapkin (2001), no correlation between the two was reported by Sachs and Polio (2007). That said, the very opportunity of OL may benefit learners regardless of the quality of their OL (e.g., Sanz et al., 2009; Storch, 2008; Swain & Lapkin, 2002; Yanguas & Lado, 2012).

2.1.2.2 Research Findings for WL

As stated, languaging includes speaking and writing (Swain, 2006), but thus far, researchers have focused more on OL than WL. Furthermore, as was the case with OL, most of the previous SLA studies used learners' writing regarding their linguistic issues in order to investigate their cognitive processes. As such, although the studies involved the act of writing, which may be considered a form of WL, the products of writing were used as a research tool to collect data, not as a learning tool. Nonetheless, some studies have been conducted, focusing on the role of WL as a learning tool. Table 2.2 below provides a list of the studies conducted on WL, including studies that utilised writing as a research tool. The findings of some of these studies are reviewed in this section.

Table 2.2
List of Studies Using WL as a Learning or Research Tool

Topics	Studies
Research tool	
- learning journals	Mackey, 2006; Mackey, McDonough, Fujii, & Tatsumi, 2001; McDonough, 2005: Schmidt & Frota, 1986
- note-taking	García Mayo & Loidi Labandibar, 2017; Hanaoka, 2007; Hanaoka & Izumi, 2012
- metalinguistic journals	Simard, French, & Fortier, 2007; W. Suzuki & Itagaki, 2007; Yang, 2010
- questionnaires	Robinson, 1996; Simard, Guénette, & Bergeron, 2015
- learner portfolios	Antonek, McCormick, & Donato, 1997
Private writing	DiCamilla & Lantolf, 1994; Lee, 2008; Roebuck, 2000
Written languaging (WL)	M. Ishikawa, 2013, 2015, 2018; M. Ishikawa & W. Suzuki, 2016; Moradian, Miri, & Nasab, 2017; W. Suzuki, 2009a, 2009b, 2012, 2016, W. Suzuki & Itagaki, 2009; M. Yilmaz, 2016

2.1.2.2.1 WL Intended as a Research Tool in the L2 Domain

As mentioned above, many studies have been conducted using learners' writing to investigate their internal cognitive processes by means of elicitation tools, such as learning journals (e.g., Mackey, 2006; McDonough, 2005; Schmidt & Frota, 1986),

note-taking (e.g., García Mayo & Loidi Labandibar, 2017; Hanaoka, 2007; Hanaoka & Izumi, 2012), metalinguistic journals (e.g., W. Suzuki & Itagaki, 2007; Yang, 2010), questionnaires (e.g., Robinson, 1996; Simmard et al., 2015) and learner portfolios (Antonek et al., 1997). As shown in Table 2.3, below, these studies utilised WL to investigate learners' cognitive processes, i.e., how they process feedback (i.e., noticing), how they reflect on feedback and/or their own language production (i.e., metalinguistic reflection), in relation to their learning.

Table 2.3
List of Studies Using WL as a Research Tool

Topics	Studies
Research tool to investigate:	
- noticing	García Mayo & Loidi Labandibar, 2017; Hanaoka, 2007; Hanaoka & Izumi, 2012; Mackey, 2006; Mackey et al., 2001; McDonough, 2005; Schmidt & Frota, 1986
- metalinguistic reflections	Robinson, 1996; Simard et al., 2007; Simard et al., 2015; W. Suzuki & Itagaki, 2007; Yang, 2010

Mackey (2006), for example, employed learning journals as one of four measures in order to examine learners' noticing of feedback during classroom conversational interaction. The participants, 28 ESL learners, were instructed to fill in journals during class time. Operationalising noticing as "a learner's report indicating a mismatch between the target language form and the learner's non-targetlike production or comprehension" (p. 413), she used what the participants reported in the journals as evidence for noticing corrective feedback. Similarly, a learning journal was employed in the study by Schmidt and Frota (1986) to investigate the relationship between noticing and the learning of Portuguese. It should be pointed out, however, that Schmidt was a learner who kept a journal for five months during his stay in Brazil. As he used the journal as a research tool to find evidence of his noticing, he trained himself to notice

his noticing and write noticing episodes down, "which might in itself have had some effect on the outcome" (p. 313). Thus, he showed concerns about the validity of using his journal as evidence of noticing (i.e., experimenter bias), but did not refer to the possible positive impact of journal writing on his learning. Meanwhile, Hanaoka and Izumi (2012) employed note-taking in order to investigate learners' noticing. They instructed Japanese English as a foreign language (EFL) learners to write down what they noticed on comparing a model essay with their own, thereby using note-taking as a measure of their noticing (see also Hanaoka, 2007, for a similar procedure). Like Schmidt and Frota, Hanaoka and Izumi did not refer to the possible positive impact of note-taking on learning.

In addition to noticing, learners' metalinguistic reflections were investigated in some studies by examining their writing with data-collection tools, such as metalinguistic journals (e.g., W. Suzuki & Itagaki, 2007; Yang, 2010) and questionnaires (Robinson, 1996; Simard et al., 2015). In W. Suzuki and Itagaki's (2007) study, for instance, 108 Japanese EFL learners worked on a grammar exercise first. Then, they were asked "to write down in Japanese as fully and precisely as they could (written metalinguistic reflections) whatever they had thought about while they were performing the exercise" (p. 136), which was used to investigate their thinking processes. Similarly, Simard et al. (2015) employed what they termed "written verbalisations" to examine learners' metalinguistic reflections with respect to written corrective feedback. The participants, 49 francophone ESL high-school learners, were instructed to write an essay on a given topic and then worked on revising the essay upon receiving written corrective feedback. Immediately after the revision, a questionnaire was conducted to elicit verbalisation data regarding the participants' understanding and

perceptions in terms of corrected errors. This procedure was conducted four times over four months, collecting written verbalisation data every time.

To summarise, all the studies reviewed above employed participants' writing as a research tool in order to investigate their cognitive processes by means of data-collection tools, such as learning journals and questionnaires. Accordingly, in spite of the possible positive impact of their writing on learning, none of the studies examined the relationship between learners' writing and L2 learning, thus not focusing on the role of writing as a learning tool.

2.1.2.2.2 WL as a Learning Tool in a Non-L2 Domain

Although outside the L2 domain, some researchers have employed learners' writing with the concept of WL, i.e., WL as a learning tool, in the form of private writing (DiCamilla & Lantolf, 1994) and learner portfolios (Antonek et al., 1997). Therefore, although their focus was not on L2 learning, the findings of these studies seem to be relevant to the current study.

DiCamilla and Lantolf (1994), for example, examined private writing, which they define as "the written externalization of portions of one's inner dialogue with the self" (p. 351) when one is faced with a difficult task. The participants were English-speaking university students who were on a beginning-level writing course. Analysis of their first drafts of compositions demonstrated changes in their use of modality and references, indicating "increased self-regulation" (p. 358). In addition, a comparison of the private writing of the writer (Virginia Woolf) and that of the novices (the participants) demonstrated that both of them made use of private writing in order to clarify their thoughts and "most importantly, be clarified, informed, and guided by them" (p. 354) in such writing activities. Based on the results, the researchers contend that private writing

serves as “an instrument of thought” in seeking and planning the solution to a problem (Vygotsky, 1986).

Furthermore, comparing the private writing of Woolf and the participants closely, DiCamilla and Lantolf (1994) state that Woolf’s private writing was deliberately deployed, i.e., intentional, whereas the use of the novices’ private writing was unintentional (p. 354). In other words, experts are capable of using private writing intentionally as a strategy/tool, probably more efficiently and effectively, while novices are yet to learn how to use the strategy as skilfully as experts. Based on the results, the researchers conclude that private writing has a positive impact on learning, stating that “Writers, like speakers, utilize their linguistic systems to do more than just express themselves to, or communicate with, social others. They also use their language, in the form of private writing, to organize and direct strategic mental processes” (p. 365).

Meanwhile, Antonek et al. (1997) investigated the potential role of learner portfolios as a mediating tool for reflection and professional development in the Vygotskian theoretical framework. The participants were two student teachers who were in a foreign language education programme. They kept learner portfolios by recording and reflecting on topics of their choice, such as new teaching techniques they tried, class reaction and self-analysis for one semester (10 weeks). An examination of the two student teachers’ portfolios at the end of the semester revealed that the portfolios benefited them to a great extent, enabling them to perform reflective practice and construct their identities as teachers. Like the study by DiCamilla and Lantolf (1994), the focus of Antonek et al. was not on L2 learning. The study, however, provides evidence that writing can be a mediational tool, helping student teachers to “go beyond remembering teaching facts to reasoning about teaching practice” (Antonek et al., 1997, p. 17), which seems to be compatible with the concept of WL.

2.1.2.2.3 WL as a Learning Tool in the L2 Domain

Starting with W. Suzuki and his colleague's studies (W. Suzuki, 2009a, 2009b; W. Suzuki & Itagaki, 2009), several studies have been conducted to address the role of WL, specifically focusing on its role as a learning tool in the field of SLA with the term "WL." Table 2.4 provides a list of studies on WL with their focus, followed by a brief review of each one.

Table 2.4
List of WL Studies by Focus

Focus	Studies
WL and feedback	Moradian et al., 2017; W. Suzuki, 2009a, 2009b, 2012, 2016; M. Yilmaz, 2016
WL and proficiency	M. Ishikawa, 2015; W. Suzuki & Itagaki, 2009
WL and L2 learning	M. Ishikawa, 2013, 2018; M. Ishikawa & W. Suzuki, 2016; Simard et al., 2007
Quality/type of WL	W. Suzuki, 2016

Although termed differently, the study conducted by Simard et al. (2007) attempted to identify the possible effect of written metalinguistic reflections (i.e., WL) on L2 learning (i.e., the acquisition of English grammar and vocabulary). The participants were 29 Canadian L1 French elementary school students in an intensive ESL class where a strong form of communicative teaching was implemented. They were required to keep a metalinguistic journal regarding what they learned in class every week for three months. Their learning was measured by pre- and posttests, consisting of a grammar accuracy test and two vocabulary tests. The researchers identified that the participants achieved statistically significant improvement on the three tests over the 3-month period. It was found, however, that a close examination of the test items and the participants' metalinguistic reflections on these items demonstrated no significant

correlations, making it impossible for Simard et al. to discern if the observed improvement was attributable to the participants' written metalinguistic reflections.

Referring to previous studies that reported a positive impact of oral metalinguistic reflections on L2 learning (e.g., Storch, 2001; Swain, 1995) in contrast to their findings, Simard et al. (2007) speculate that the highly communicative learning context might have influenced the results, suggesting that a traditional learning context where more focus is on form might have produced different outcomes (see Elder & Manwaring, 2004; Hu, 2011, for a similar perspective). Moreover, as one of the limitations, Simard et al. point to "the lack of explicit training in reflecting about language per se" (p. 519). Also, as another limitation, they refer to the open-ended questions used to elicit the participants' written metalinguistic reflections. That is, they stated that the type of questions might have contributed to the fact that the participants "only rarely referred to the grammatical features examined in the test" (p. 516).

In addition, although the researchers did not refer to this possibility, the age of the participants, who were elementary school students aged 10–12, might have contributed to the outcome. Namely, in studies that reported a facilitative effect of oral metalinguistic reflections (e.g., Storch, 2001; Swain, 1995), the participants were adult learners (Storch, 2001) or slightly older students (Swain, 1995), who were probably more cognitively developed and capable of metalinguistic reflections. Because of the shortcomings of the study, it was not feasible for the researchers to identify a direct link between the participants' written metalinguistic reflections and their L2 learning. They stated, however, that "participants' metalinguistic activity may have nevertheless affected their L2 learning … enhancing acquisition by helping them to notice elements of the input they might not have otherwise noticed" (p. 519). Given the improved tests scores, their speculation seems highly plausible.

Meanwhile, in one of the first WL studies by W. Suzuki and Itagaki (2009), 141 Japanese EFL learners at two different levels were encouraged to engage in WL in their L1 while they worked on a translation task and checked a model translation afterward. Of the 141 participants, 73 were high school students and 68 were university students, who were assigned to low-intermediate and high-intermediate proficiency groups, respectively. Analysis of these WLEs revealed that grammar-related WLEs were in the majority in both groups, contradicting previous findings of studies that reported that learners focused mainly on lexis (e.g., Hanaoka, 2007; Hanaoka & Izumi, 2012; J. Williams, 2001). Furthermore, it was found that the higher-proficiency participants languaged more and produced more grammar-related WLEs than the lower-proficiency participants.

In W. Suzuki's (2009a) study, he examined the impact of indirect feedback and WL on revisions of essays in a 3-week experiment. In Week 1, the participants, 24 Japanese EFL learners, were instructed to write essays based on a prompt provided for 30 minutes. In Week 2, they received their draft from the previous week with indirect feedback (i.e., incorrect words and phrases were underlined in red ink) and were asked to engage in WL, i.e., to explain in writing "why their linguistic forms (e.g., grammar, lexis) had been incorrect/wrong" (p. 83) for 25 minutes in their L1, Japanese. They were allowed to say "I don't know" when they did not know why their writing was corrected. In Week 3, the participants received a draft of their essays from the first week (without any feedback) and were asked to revise them for 20 minutes.

The analysis, which compared their first drafts from Week 1 and revisions from Week 3, revealed that the average number of linguistic errors decreased significantly in the revisions. When the relationship between the WLEs which included the participants' explanations for the reasons for their errors and their revisions was examined, more than

half (53.5%) of them were found to be successfully revised. What is noteworthy is that even the "don't know WLEs," where the participants did not explain the reasons for their errors, were found to have resulted in successful revisions 42% of the time, suggesting a positive impact of WL. These results seem to echo the findings by Swain and Lapkin (2002), who identified the same rate for participants who rejected feedback but answered correctly on a posttest as those who accepted feedback, probably as a result of the experience of "talking-it-through." In the case of this study, writing about errors, even if they were merely noticing errors (not at the level of understanding), might have facilitated their revision (i.e., writing-it-through). Based on these results, W. Suzuki concluded that the study provided empirical evidence that "learners improved their linguistic accuracy by languaging about indirect feedback on their linguistic errors" (p. 87). It should be pointed out, however, that he did not employ any measures to assess the participants' L2 learning besides the revisions of their essays.

Next, in order to investigate the relationship between *direct* feedback and WL as well as the effects of WL on L2 writing revision, W. Suzuki (2012) conducted another study (see details for his thesis, W. Suzuki, 2009b). With a little modification to the procedures followed in W. Suzuki (2009a), the experiment was conducted over two weeks. In Week 1, as in his previous study (2009a), the participants, also 24 Japanese EFL learners, were instructed to write essays based on a prompt for 30 minutes. Then, copies of the participants' essays were prepared and their native English instructor provided direct feedback by correcting all the linguistic errors that he noticed by providing the correct forms or structures, deleting any unnecessary parts, and inserting missing words or phrases, i.e., direct written corrective feedback. The following week, upon seeing the direct feedback, the participants were instructed to explain the reasons why their linguistic forms were corrected in their writing (i.e., WL) for 30 minutes and

worked on a questionnaire for 20 minutes. Soon after this, they were given copies of their first drafts from the previous week (without feedback) and worked on revisions, also for 20 minutes (immediate revision).

Two findings that are consistent with previous WL studies were reported. First, in terms of the revisions of the essays, the numbers of errors decreased significantly, echoing the findings of W. Suzuki (2009a) and suggesting a positive impact of WL on L2 learning. Second, an analysis of the types of WLEs revealed that the participants engaged in grammar-related languaging the most, replicating the results of W. Suzuki and Itagaki (2009). Based on these results, W. Suzuki concluded that WL has a facilitative impact on L2 learning. Also, interpreting the result that grammar-related WLEs were in the majority, he points to two aforementioned facilitative characteristics of WL for learning. Namely, WL generally allows learners more time, freeing them from the pressure to make timely responses to their interlocutors. In addition, WL provides a form of external memory that learners can reflect on when necessary (see also J. Williams, 2012). Although the study provides evidence to demonstrate the positive impact of WL on revision, suggesting a positive impact of WL on L2 learning as well, it was "of an exploratory nature, rather than a carefully controlled experimental study" (W. Suzuki, 2012, p. 1128). Therefore, as the researcher admits, there were some limitations.

First, the study did not include a control group, leaving possibility for speculation that the observed positive effect of WL might have been due to direct feedback. Second, although WL was identified to have contributed to revisions immediately after the treatment, the study did not investigate the impact of the participants' WLEs on their L2 learning, as was the case with his previous study (2009a). Moreover, the long-term effect on revision was not examined. Third, although the study reported that the

participants produced grammar-related WLEs the most, it is worth mentioning that the instructor might have made more corrections to grammar errors than to other linguistic issues, making grammar-focused WLEs the most frequently produced WLEs. Despite these limitations, the study is noteworthy in that it was the first attempt to examine the relationship between WL and direct feedback and inspired other researchers, resulting in the following two replication studies.

Replicating W. Suzuki's (2012) study, M. Yilmaz (2016) conducted his study in a Turkish EFL context. The participants were 17 Turkish-speaking EFL university students enrolled on an elementary-level course. In the first week, they wrote paragraphs regarding their best or worst days of their lives. The following week, they received direct feedback on their paragraphs from the instructor/researcher, and engaged in WL, i.e., they explained the reasons for the corrections in Turkish. As in W. Suzuki's original study, they were allowed to write "I don't know" when they could not explain the reasons. Echoing the results found in the original study, improvements in the revisions and the highest number of WLEs on grammar of all the WLEs were identified. Also consistent with the original study, the same limitations were, however, observed. That is, a control group was not included and no measure of learning other than immediate revision was employed. Given that more than half (57%) of the "don't know WLEs" turned out to be successful on immediate revision, which is a higher proportion than in W. Suzuki's original study (42%), one could argue that the improvement may be attributable to memory, rather than learning. In addition, the same claim could be made regarding the result that grammar-related WLEs were in the majority. That is, there could be a possibility that feedback on grammar might have been in the majority.

In the other replication study, circumventing the issue of the lack of a control group in W. Suzuki's (2012) original study, Moradian et al. (2017) included a control

group to tease out the effect of WL from that of corrective feedback in an Iranian EFL context. Thirty-eight Iranian EFL learners were randomly assigned to two groups, direct feedback plus WL and direct feedback only. Employing dictogloss instead of essay-writing as a treatment task, the experiment was conducted over four weeks. In the first week, a practice session in dictogloss was run. In the second week, the main dictogloss task was performed. Direct feedback was given to the participants' reconstructed texts, which were returned to the participants the following week (Week 3). Upon receiving the reconstructed texts with direct feedback, the participants in the +WL condition were instructed to engage in WL, while the participants in the other group were simply asked to read "their writings and number their errors" (p. 8). Although the time limit was 25 minutes, the average time that the participants actually spent for this task was 21 and 14 minutes for the direct feedback plus WL group and direct feedback only group, respectively.

Instead of immediate revision as conducted in W. Suzuki (2012), all the participants engaged in the revision in the final week, receiving clean copies of their reconstructed texts from the second week (i.e., no direct feedback added). Following the procedure followed by W. Suzuki, the reconstructed texts from Week 2 and the revisions from Week 4 were analysed, and these demonstrated that the group which engaged in WL outperformed their counterparts by improving their accuracy significantly. Therefore, the researchers argue that their findings lend support to W. Suzuki's claim that WL has a positive impact on learning. Although the study attempted to circumvent the shortcomings of the original study, some limitations were left unaddressed. That is, revisions were the only measure of the participants' linguistic improvement, as in W. Suzuki's original study. In addition, the difference between the

two groups could be attributed to a gap in terms of time on task, as the researchers admit.

Against this background, coining the term “metanote” as opposed to metatalk for WL, M. Ishikawa (2013) conducted an experiment with a pre-posttest design over three weeks in order to investigate the impact of WL on L2 learning and its possible enduring effect. The participants were 14 Japanese university students who were lower-intermediate EFL learners. In the first week, a pretest on tense consistency, the target construction, was conducted. In Week 2, the participants worked on a three-stage Japanese-English translation task. They were instructed to write about any problems or questions that occurred to them as they translated (Stage 1) and about whatever they noticed while checking the model translation (Stage 2), with both of these being examples of WL. The effect of WL was analysed by comparing the results of the participants’ pre- and posttests with those of the participants who performed the same task but without WL. In Stage 3, they did a posttest. In the third week (after a 4-week interval), a delayed-posttest was conducted.

Although the analysis of the results of the pre-and posttests produced no statistically significant evidence, from an SCT perspective, the case studies that were conducted indicated that WL positively influenced L2 learning. Namely, the two participants who produced more WLEs (i.e., two higher metanote takers) showed not only a positive attitude towards WL, but also improvement in their test scores. It should be pointed out, however, that the two participants who produced fewer WLEs (i.e., two lower metanote takers), either rejected the model translation or did not focus on the target form, thus not improving their test scores. It is important to note that, in contrast to the findings of W. Suzuki and Itagaki (2009) and W. Suzuki (2012), lexis-related WLEs were identified as the majority of WLEs in both stages. Given that the higher-

proficiency participants languaged more and produced more grammar-related episodes than the lower-proficiency participants in W. Suzuki and Itagaki's study, proficiency levels were assumed to be a factor influencing this result.

In order to address the issue, M. Ishikawa (2015) investigated the impact of learners' proficiency levels on their types of WL, which were divided into the following three categories according to linguistic foci: grammar, lexis and other. The participants were 24 Japanese EFL learners from two English classes with differing proficiency levels and they did the translation task used in M. Ishikawa (2013). An analysis of the WLEs and task outcome (i.e., accuracy in translation) revealed that the participants' proficiency levels influenced their WL in terms of focus on the target form (i.e., tense consistency). To be more specific, higher-proficiency participants paid considerably more attention to the target form, echoing the findings of Yang's (2016) study on OL, where the learners with intermediate proficiency focused more on form when they worked with a higher-proficiency participant than with a lower-proficiency one. Such was not the case, however, with regard to the types of WLEs. That is, higher-proficiency participants produced slightly more grammar-focused WLEs than did lower-proficiency participants, but not significantly so.

Echoing the results of M. Ishikawa (2013) and previous SLA studies (e.g., Hanaoka & Izumi, 2012; J. Williams, 2001), but in contrast to W. Suzuki and Itagaki (2009), W. Suzuki (2012) and M. Yilmaz (2016), both groups produced lexis-focused notes the most frequently. Interestingly, all the participants, regardless of their proficiency levels, produced more WLEs when their translations were incorrect. Similar results were obtained in the aforementioned Qi and Lapkin's (2001) study, where most of the problems that were noticed but not resolved in the first stage were noticed again in the next stage. Explaining the results, Qi and Lapkin state that "This sense of lack of

fulfilment … may push a learner to look out for any future relevant information available" (p. 289), which was likely to be the case in M. Ishikawa's (2015) study despite the difference in modality. Although this study shed some light on the relationship between learners' proficiency levels and their WL, it did not address the role of WL as a learning tool.

In such a context, replicating the study by Swain et al. (2009) on OL, M. Ishikawa and W. Suzuki (2016) examined the effects of WL on L2 learning in order to identify whether the findings for OL (positive impact of OL on learning) could be applied to WL. Forty Japanese EFL learners were assigned to three groups, written languaging (+WL), no WL (–WL) and a control. The experiment was conducted over three weeks, with the experimental procedures consisting of a pretest, treatment and two posttests. In Week 1, a pretest was administered to all groups. In Week 2, after reading a text on the target construction (i.e., the present counterfactual conditional), the +WL group was instructed to write about their understanding of the relevant rule (i.e., WL) and the –WL group worked on a grammar exercise related to this structure. A posttest followed, a delayed posttest was conducted one week later. The participants in the control group only took the pre- and posttests. In order to measure the participants' learning, two types of assessment tests were devised, multiple-choice recognition tests and production (Japanese-English translation) tests. Although both the +WL and –WL groups outperformed the control group on all the posttests, no statistically significant difference between the two treatment groups was observed. It should be pointed out, however, that only the +WL group scored significantly higher than the control group in the delayed production posttest, hinting at the facilitative effect of WL. No such difference was identified with respect to the recognition tests.

More recently, W. Suzuki (2016) examined the relationship between the quality of WL and participants' L2 learning (measured by improvement in revisions) with data from his 2012 study. WLEs were divided into three categories, noticing only (i.e., explanation without reasons or metalinguistic terminology), noticing with reasons (i.e., explanation with reasons and/or metalinguistic terminology) and uncertainty (i.e., the "I don't know" episode in his study in 2012). In terms of the quality of WL, analysis of the WLEs revealed that 75% of them were coded as "noticing with reasons," whereas much lower ratios of WLEs were categorised as noticing only and uncertainty (13% and 12%, respectively). Pointing to the results, the researcher states that "the participants were more likely to express deeper levels of awareness about direct corrections of their errors" (p. 17). With respect to the relationship between the quality of WL and the accuracy of revision (i.e., an indicator of L2 learning), the results revealed that both noticing only and noticing with reasons contributed to accuracy improvement to a greater degree (both over 90%) compared to uncertainty (71%), thus attesting to the claims of Schmidt's (2001) Noticing Hypothesis in that noticing contributes to learning. That said, the outcome that both noticing only and noticing with reasons WLEs resulted in similar degrees of accuracy improvement contradicts Schmidt's perspective that a higher level of noticing leads to greater learning than a lower level. In addition, it is not consistent with the findings of Qi and Lapkin's (2001) study, where a higher level of noticing, i.e., substantial noticing in their terms, resulted in greater learning than a lower level of noticing, i.e., perfunctory noticing.

This discrepancy might be explained by the research design of W. Suzuki's (2012) study. Noted as one of the limitations, improvement was only measured by immediate revision, which might have contributed to the similar results for noticing at both levels. Especially, given that over 70% of the uncertainty WLEs resulted in

successful revision, as stated regarding the result of M. Yilmaz's (2016) study, it is plausible that the participants addressed revision from memory rather than understanding. Interpreting the results with these limitations in mind, W. Suzuki (2016) is cautious to conclude that WL could function as an effective mediating or retrospective tool for problem-solving in L2 learning, without referring to a possible direct link between WL and L2 learning.

2.1.2.2.4 Summary of WL Research and Current Issues

As the brief review above shows, WL is a growing area of research and several studies on WL have been conducted since the pioneering works by W. Suzuki (2009a, 2009b) and Simard et al. (2007). These studies have investigated WL in relation to direct feedback (W. Suzuki, 2009b, 2012; Modarian et al., 2017; M. Yilmaz, 2016), indirect feedback (W. Suzuki, 2009a), learners' proficiency levels (M. Ishikawa, 2015; W. Suzuki & Itagaki, 2009), its possible impact on L2 learning (M. Ishikawa, 2013, 2018; M. Ishikawa & W. Suzuki, 2016; Simard et al., 2007) and the potential impact of its quality on revision (W. Suzuki, 2016), generally producing promising results regarding the effect of WL on L2 learning (see, however, M. Ishikawa, 2015; Simard et al, 2007; W. Suzuki & Itagaki, 2009). That said, there remain many issues still to be resolved.

First, compared to OL, WL is still an underexplored area of research and the number of studies conducted on it is limited. Second, although a facilitative effect of WL on learning has been reported, the results are only indicative because of no inclusion of pre- and posttests (W. Suzuki, 2009a, 2009b, 2012) and a control group (W. Suzuki, 2009a, 2009b, 2012; M. Yilmaz, 2016); also, learners had unequal time on task (Modarian et al., 2017) in previous studies, and a direct link between WL and learning

has not yet been investigated. Therefore, the reported positive impact of WL on L2 learning merits empirical validation.

Moreover, studies conducted with a pre-posttest design did not find statistically significant differences between a treatment group and a comparison group (M, Ishikawa, 2013) or between two treatment groups (M. Ishikawa & W. Suzuki, 2016). The most plausible reason to account for the outcome in M. Ishikawa's study is that the participants, at least the two low-metanote takers, i.e., participants who produced fewer metanotes as mentioned above, might not have understood the concept of WL because of insufficient practice (only one 10-minute session). In addition, given that the +WL group seemingly outperformed the –WL group on the posttests, the small number of participants (i.e., seven in each group, 14 in total) might have made it difficult to obtain statistically significant results, indicating a need to include more samples to increase the statistical power. Meanwhile, as for M. Ishikawa and W. Suzuki's study, the treatment task for the –WL group (i.e., a grammar exercise on the target structure), while the +WL group engaged in WL activity (i.e., writing about their understanding of the target structure), probably enhanced the –WL participants' explicit knowledge with respect to the target construction as much as the +WL condition, presumably contributing to no significant differences between the two groups. Taken together, the shortcomings of the research designs and their operationalisation might have masked the possible facilitative impact of WL on L2 learning.

Third, given the results of the case studies in M. Ishikawa (2013), which demonstrated the differences in learners' attitudes toward WL, individual differences, such as aptitude and motivation, are likely to influence the impact of WL on L2 learning. However, no studies have yet been conducted to address this issue. Fourth, similar to the third point, considering the findings of W. Suzuki and Itagaki (2009) and

M. Ishikawa (2015), learners' proficiency levels seem to have a significant impact on the type of WLEs, further influencing their L2 learning. Once again, however, a possible link is yet to be investigated. Fifth, judging from the findings of OL by Knouzi et al. (2010), the quality of WL is also likely to have an impact on L2 learning. Although the quality of WL was examined by W. Suzuki (2016), it has not been examined in relation to L2 learning. Therefore, the findings obtained to date should only be considered suggestive until confirmed by future research that addresses these issues.

2.1.3 Research Findings for Oral and Written Self-explaining (Non-L2 Domain)

As no SLA research has yet examined self-explaining (languaging) from a cognitive perspective, the research findings in non-L2 domains are reported in this section. First, the findings for oral self-explaining are introduced below, followed by those for written self-explaining.

2.1.3.1 Research Findings for Oral Self-explaining

In their pioneering work on self-explaining, Chi et al. (1989) investigated the relationship between learners' oral self-explanations and learning in physics. The participants were 10 university students with similar levels of (i.e., a little) domain knowledge at the outset of the experiment. They were instructed to self-explain after they studied examples in a physics book in a knowledge-building phase while also working on problems in a problem-solving phase. The participants were later categorized as "Good" and "Poor" self-explainers based on their task performance in the problem-solving phase. It was found that Good self-explainers' self-explanations were different from those of Poor ones, both quantitatively and qualitatively. Namely, Good self-explainers not only produced more self-explanations, but also generated more

inferences, monitored their comprehension, justified their explanations and connected new knowledge to their prior knowledge more frequently than Poor self-explainers, who mainly produced paraphrasing self-explanations. The results provide evidence to support the proposed mechanisms of inference generation (Chi et al., 1994; Siegler, 2002), mental-model repair (Chi et al., 1994) and knowledge integration (Rittle-Johnson & Loehr, 2017). Based on the results, the researchers contend that self-explaining facilitates learning, arguing that the more self-explanations that learners produce, the greater the learning and the deeper understanding they achieve.

Building on Chi et al. (1989), Chi et al. (1994) conducted another study in order to further investigate the relationship between self-explaining and subsequent understanding, employing a biology text as material this time. Moreover, on the assumption that the number of self-explanations and the amount of learning correlate, the researchers devised instructional prompts in order to enhance the elicitation of self-explanations. More specifically, the participants were instructed to self-explain after reading each sentence of a passage on the human circulatory system. The participants were 24 eighth graders with a range of abilities in terms of scores on the California Achievement Tests (CAT), widely used tests of basic academic skills for children from kindergarten to grade 12. These participants were recruited in order to enable the researchers to examine the relationship between learners' ability differences and the impact of self-explaining on learning. The participants were divided into two groups depending on a requirement to self-explain, a prompted group that was required to self-explain and an unprompted group that was not.

The researchers reported four important findings. First, analysis of the posttest results demonstrated that the learners in the prompted group outperformed their counterparts in the unprompted group who read the same text twice without self-

explaining, providing evidence for the facilitative impact of self-explaining on learning. Furthermore, the differences in learning between the learners in the two conditions were more noticeable with more difficult questions, suggesting that generating self-explanations enabled the learners to understand the text more deeply than the learners who just read it, allowing them to address more challenging questions. Second, with respect to the number of self-explanations, high-explainers, i.e., learners who produced more self-explanations, were found to have attained a deeper understanding of the text than lower-explainers, i.e., learners who produced fewer self-explanations, lending support to the findings of Chi et al. (1989). Third, echoing their earlier study, regarding the quality of self-explanations, the participants who generated inferences and integrated their prior knowledge with new knowledge demonstrated more remarkable gains than those who simply reread the sentences. Fourth, in terms of the relationship between the participants' ability and learning, the researchers found that all the self-explainers attained similar gains on posttests regardless of their CAT scores, indicating that self-explaining helps learners on all academic levels.

As stated above, the positive impact of self-explaining is confirmed. In addition, the correlation between amounts of self-explanations and of learning, as well as the relationship between the quality of self-explanations and learning, is identified. Furthermore, it is found that learners do not generate productive self-explanations spontaneously (Conati & VanLehn, 2000; Renkl, 1997). Based on these findings, researchers started to conduct self-explaining studies with improved prompts and/or including training in order to elicit greater numbers of and higher quality of self-explanations. For example, using explanatory texts on complex causal phenomena from the natural or social sciences, Griffin, Wiley, and Thiede (2008) instructed participants to self-explain explicitly, asking questions such as "What new information does this

paragraph add? How does it relate to previous paragraphs? Does the paragraph raise new questions in your mind?" (p. 97). In addition to the prompts, they prepared a short example text with sample self-explanation comments for each sentence for the participants. It was found that those self-explainers outperformed their counterparts, who just read the text once or twice, on posttests administered after the treatment, providing evidence to support the previous findings of Chi et al. (1994).

Meanwhile, in terms of training, McNamara (2004) developed a programme called self-explaining combined with reading training (SERT) and examined its effect on the comprehension of a difficult biology text on cell mitosis and the improvement in self-explanation quality of 42 university students who were divided into two groups, a SERT condition and a control condition. The experiment consisted of five one-on-one sessions conducted in a controlled laboratory setting. It was found that SERT improved low-knowledge students' ability to self-explain and comprehend a difficult text. More recently, building on her study in 2004, McNamara (2017) also examined the impact of SERT in the field of biology. In contrast to her earlier study, the experiment was conducted with a larger number of participants (i.e., 265), who were university students in biology classes in a classroom setting. In addition, instead of the comprehension of a text, learning was examined by their performance on exams. Despite the changes, echoing the findings of her earlier study, SERT was found to have benefited low-knowledge students (i.e., students with less knowledge about biology at the beginning of the course). That is, it was found that low-knowledge students, with the aid of SERT, performed comparably to high-knowledge students on tests taken throughout the course, whereas low-knowledge students without SERT did not perform as well as high-knowledge students. Interpreting these results, McNamara states that self-explaining has a positive impact on learning. Moreover, pointing to the result that SERT did not benefit

high-knowledge students, she adds that SERT benefits low-knowledge students more by allowing them to improve their deep-level comprehension of texts, whereas high-knowledge students probably do not need such facilitation because they already have sufficient knowledge readily available to understand the texts. It should be pointed out, however, that as she did not separate the self-explaining and reading training conditions, the effect of self-explaining alone is not clear.

2.1.3.2 Research Findings for Written Self-explaining

As for self-explaining in the written mode, learners have been asked to self-explain either by handwriting (e.g., Kastens & Liben, 2007) or typing (e.g., Hausmann & Chi, 2002; Muñoz et al., 2006). For example, Kastens and Liben (2007) instructed fourth graders to write down their reasoning every time they made a response to a field-based map skills task. It was found that written self-explainers performed significantly better than their counterparts who performed the task without written self-explaining, indicating the positive impact of written self-explaining on learning. Meanwhile, growing numbers of studies, including three studies reviewed below, have been conducted, focusing on typed self-explaining, presumably because of practical benefits. That is, compared to resource-intensive traditional approaches which involve human instructors, computer-based environments can provide one-on-one lessons without having human instructors (Chiu & Chi, 2014, p. 94).

Hausmann and Chi (2002), for example, examined the impact of typed self-explaining on learning to explore the possibility of utilising computers for self-explaining. In order to facilitate comparisons with previous studies on oral self-explaining, the learning and assessment materials from Chi et al. (1994) were employed with only minor changes. For instance, as the participants in Chi et al.'s study were

eighth graders, the text on the human circulatory system was revised to make it more challenging for the participants who were 20 psychology major university students. They were divided into two groups, i.e., a typing self-explaining group and a read-only group, depending on the requirement to type self-explaining while reading the text.

Considerably different results from those of previous studies on oral self-explaining were observed. That is, the participants in the typing self-explaining group were identified to have generated far fewer self-explanations than Chi et al.'s (1994) participants who engaged in oral self-explaining. In terms of the types of self-explanations, the participants were found to have predominantly paraphrased the text, which is not expected to support comprehension as much as other types of self-explaining, such as inferencing and knowledge integration (Chi et al., 1994; McNamara, 2004, 2017). Moreover, no statistically significant difference was identified between the typing self-explaining group and the read-only group on posttests administered after the treatment. Interpreting these results on the assumption that the number of self-explanations could correlate with the amount of learning, Hausmann and Chi stated that the few typed self-explanations contributed to the outcome. In addition, explaining the reason why the participants produced significantly fewer typed self-explanations compared to previous findings on oral self-explaining (Chi et al., 1994), the researchers pointed to two possibilities, i.e., typing skills and avoidance of errors.

First, Hausmann and Chi (2002) state that bad typists might have found it difficult to engage in typed self-explaining, resulting in fewer self-explanations. However, given that the participants were "average typists" (p. 6), according to their self-reports, the second possibility seems to be more plausible. That is, the participants were aware that their typed self-explanations, which were going to be transformed into log files, would be analysed later as permanent records by the researchers. Therefore, they might have

preferred to avoid errors by not generating many self-explanations and producing paraphrased self-explanations, which were not likely to result in errors but also not in learning, either. According to McNamara (2004), paraphrasing "does little … to aid comprehension" (p. 19). As mentioned earlier, the function of permanent records is considered to be a facilitative feature of written self-explaining and WL (W. Suzuki, 2012; J. Williams, 2012). Nonetheless, given the account by Hausmann and Chi, learners may find it face-threatening depending on how written self-explaining is employed. In addition, although not pointed out by the researchers, the participants might simply have found it troublesome to type their explanations. Supporting this speculation, in the aforementioned study by Simard et al. (2015), some participants demonstrated a negative attitude towards writing, stating "I don't like writing much."

In contrast to Hausmann and Chi's (2002) study, which demonstrated considerable differences between self-explaining in two modalities, little difference was found by Muñoz et al. (2006). They investigated the impact of typed self-explaining by directly comparing with that of thinking aloud (i.e., oral self-explaining) while students read science texts. Two experiments were conducted with around 50 university students as participants in each experiment. In addition to the modality of self-explaining (i.e., typed vs oral), the reading skill of the participants (i.e., skilled vs less skilled) and types of reading strategy (paraphrasing vs inferencing) used for self-explaining were examined. The results obtained from the two experiments demonstrated no significant differences due to the modality of self-explaining to indicate comprehension of the texts. Moreover, the modality was not found to have a significant impact on the types of reading strategies utilised by the participants in either group, only demonstrating "subtle" differences, in contrast to the "dramatic" differences observed in Hausmann and Chi's study (p. 216).

However, when the two modalities were compared in relation to the level of the participants' reading skills, some differences emerged. To be more specific, what is noteworthy for this thesis is that, when less-skilled readers typed self-explanations, they were found to have used inferencing as much as skilled readers, whereas less-skilled readers in the oral condition did not use the strategy as much as their more-skilled counterparts. Given that making inferences is one of the cognitive processes that contribute to learning more than simply paraphrasing texts (Chi et al., 1994; McNamara, 2004, 2017), the results seem to indicate that less-skilled readers benefited more from typing than speaking, which may be attributable to the fact that typing offers less-skilled readers more time for reflection and puts less demand on working memory than speaking. With respect to this point, i.e., greater time for reflection offered by typing, Muñoz et al. (2006) state that "Less skilled readers seemed to have reaped the benefits of this more reflective response mode, and they were able to express more inferences when typing than when speaking" (p. 216).

Another important point for this thesis is that, unlike Hausmann and Chi's (2002) findings, Muñoz et al. (2006) found that the participants in the spoken self-explaining condition produced more paraphrasing self-explanations than those in the typed condition. Explaining the result, the researchers described it as "*economy of expression* in typing" (p. 216). That is, the participants probably knew that paraphrasing would not facilitate their understanding of the text as much as inferencing. Therefore, in order to be economical (i.e., to save effort and time typing), they did not type as many paraphrasing self-explanations as those in the spoken condition. According to Muñoz et al., thinking aloud, which can be compared to oral self-explaining, "is less effortful and less time consuming, and thus less 'costly' for verbalizing a thought" (p. 216). Conversely, writing can be more effortful and time-consuming, which may be another

reason why the participants in Hausmann and Chi's study did not produce as many typed self-explanations as the participants in the previous study on oral self-explaining. They might have produced fewer typed self-explanations in order to save effort and time, i.e., to be more economical.

Similarly, typed self-explaining was examined by McNamara, O'Reilly, Best, and Ozuru (2006), but their focus was on enhancing the quality of typed self-explanations along with training in reading strategies. Building on the success of her aforementioned human-delivered programme, i.e., SERT (McNamara, 2004, 2017), McNamara and her colleagues developed a computer-based reading strategy training program called the Interactive Strategy Trainer for Active Reading and Thinking (iSTART). In the program, animated pedagogical agents (i.e., a teacher-agent and student-agents) provide learners with reading strategy instructions by interacting with each other and with the user (i.e., learner) in order to elicit the types of self-explanations that are likely to facilitate knowledge-building, i.e., a global level of comprehension that goes beyond a given text (McNamara et al., 2006, p. 156). The participants were 39 adolescents in the United States (average age: 12.7 years), who were divided into two groups, an iSTART group and a control group. Their conditions differed in that they received different types of instruction before the treatment. Namely, the participants in the iSTART group completed the iSTART program, consisting of three modules (i.e., introduction to self-explaining, demonstration of reading strategy, and training in typed self-explaining using the strategy), while those in the control group were shown only the first module of iSTART (i.e., introduction to self-explaining). After the participants completed the pre-treatment instruction under their assigned conditions, all the participants read a low-cohesive, i.e., complex, science text and the impact of the iSTART was examined by comparing the comprehension of the text of the two groups.

Three findings relevant to the present thesis were reported. First, the participants in both groups improved their understanding of the text, indicating the positive impact of typed self-explaining (i.e., WL), even without training. Second, the participants in the iSTART group outperformed their counterparts in the control group, providing evidence for the facilitative effect of iSTART. The results also support Chi et al.'s (1994) claim that the effects of self-explaining can be enhanced depending on the prompts and training. Third, a closer examination of the participants in the iSTART condition revealed that those with less prior knowledge about reading strategies performed significantly better on text-based questions (i.e., simpler questions), whereas those with more prior knowledge (high-strategy knowledge participants) improved their comprehension of inference questions (i.e., more challenging questions). Interpreting the results, McNamara et al. (2006) state that iSTART benefited all learners, regardless of their prior knowledge, but probably within their ZPD (Vygotsky, 1978).

2.1.3.3 Summary of Oral and Written Self-explaining

To summarise oral and written self-explaining studies, four points have been identified. First of all, although no self-explaining studies have been conducted in the field of SLA, self-explaining has been identified as facilitating learning in a wide variety of fields, such as biology, physics and geography, indicating that it can be a domain-independent learning strategy (e.g., Chi et al., 1994). Second, the amount of self-explaining generally correlates to the amount of learning, high self-explainers benefiting more than low self-explainers (e.g., Chi et al., 1989; Chi et al., 1994; Muñoz et al., 2006). Third, it is not just the quantity but the quality of self-explaining that influences the amount of learning. More specifically, self-explaining that involves inferences and the integration of new and prior knowledge contributes to greater

learning than self-explaining in the form of paraphrasing or repetition (e.g., Chi et al., 1989; Chi et al., 1994). Given the successful results of SERT (McNamara, 2004, 2017) and iSTART (McNamara et al., 2006), providing strategy training to improve the quality of self-explanations appears to be beneficial in both modalities. Fourth, in terms of learners' levels, the findings are rather mixed. Some researchers have found that self-explaining benefits all learners irrespective of their level of knowledge (academic skills) (e.g., Chi et al., 1994), whereas others have found that self-explaining benefits lower-knowledge learners more (e.g., McNamara, 2004, 2017; Muñoz et al., 2006). This discrepancy may stem from the variables, such as differences with respect to difficulties with measurement instruments (McNamara et al., 2006) and inclusion or not of training (McNamara, 2004; McNamara et al., 2006).

Finally, although the findings for self-explaining overlap across modalities, some are inherent to the written mode. First, echoing the findings of WL studies, a positive impact on learning is reported, probably deriving from the characteristics of written self-explaining, i.e., more time and external memory (i.e., fewer demands on working memory) (Muñoz et al., 2006). At the same time, however, it is also reported that the function of external memory can be a negative affective factor, in that learners may find it intimidating to make mistakes as they are aware that their written self-explanations may be recorded permanently (Hausmann & Chi, 2000). Second, depending on individual differences, some learners may find writing more trouble than speaking, thus not producing self-explanations because of economy of expression (Muñoz et al., 2006).

2.2 WL and Learners' Proficiency Levels

As introduced above, several languaging and self-explaining studies in both modalities have examined learners' proficiency levels and level of knowledge/ability in

relation to two factors, (1) their learning and (2) the content (quantity and/or quality) of their languaging or self-explanation. In terms of learning, Ammar and Hassan (2017) found that lower-proficiency learners achieved greater gains from collaborative dialogue (OL) than did higher-proficiency learners. Similarly, McNamara (2017) reported that lower-knowledge learners demonstrated similar levels of performance to higher-knowledge learners under a SERT (i.e., oral self-explaining and training) condition, stating that SERT facilitated and benefited lower-knowledge learners' advancement more than higher-knowledge learners, who were not likely to need such facilitation. Also, Muñoz et al. (2006) found that typed self-explaining (WL) assisted lower-knowledge learners more by enabling them to use cognitively demanding but effective strategies probably because of less time pressure than with oral self-explaining. It should be pointed out, however, that Chi et al. (1994) found that self-explaining benefits learners similarly regardless of their academic levels. A similar claim is made by Yanguas and Lado (2012), who contend that OL can benefit naive and native learners alike, regardless of their proficiency levels. Meanwhile, McNamara et al. (2006) report that the levels of learners' prior knowledge influence their achievement, indicating some interaction between learners' knowledge level and attainment.

As for the relationship between learners' proficiency levels and the content of their languaging, Qi and Lapkin (2001) found that, for OL, the LREs of a higher-proficiency learner indicated a higher level of noticing compared to those of a lower-proficiency learner. Similar results were obtained in the OL study by Yang (2016) and the WL study by M. Ishikawa (2015). Meanwhile, W. Suzuki and Itagaki (2009) report that higher-proficiency learners produced not only more WLEs, but more WLEs on grammar. Given that Swain et al. (2009) and Chi et al. (1989) identify significant correlations between the number of OL/self-explanations and learning, higher-

proficiency learners may be expected to achieve greater learning, conflicting with the findings of Ammar and Hassan (2017), McNamara (2017) and Muñoz et al. (2006). However, no WL studies have examined the possible impact of learners' proficiency level on their L2 learning.

2.3 WL and Explicit Learning

As mentioned earlier, writing is a conscious process (DiCamilla & Lantolf, 1994; Luria, 1999) that usually allows learners more time than speaking, thereby creating optimal conditions for explicit learning, i.e., learning with awareness (Leow, 2015), and enabling learners to draw on their explicit knowledge (J. Williams, 2008). Languaging is also a conscious process. It enables learners to reflect on their linguistic issues, presumably leading to deeper thoughts as a consequence. Therefore, WL, a combination of both writing and languaging, may be expected to further facilitate explicit learning. In this section, first, the nature of explicit/implicit learning and knowledge is discussed. Then, the role of attention and awareness in L2 learning is considered, followed by a discussion of the interface between explicit and implicit knowledge and the relationship between WL and explicit learning. Finally, the debate on the merits of written corrective feedback (WCF) is discussed in relation to explicit knowledge.

2.3.1 Explicit/Implicit Learning and Knowledge

There has been a great deal of discussion regarding the roles of explicit learning and implicit learning (DeKeyser, 2003; Krashen, 1981) in SLA. The term "implicit learning" was first employed by Arthur Reber (1967), the pioneer of implicit learning research (DeKeyser, 2003). Reber conducted two experiments in order to investigate the learning of an artificial grammar. These revealed that learners acquire knowledge about

complex, artificial grammar implicitly, that is, without intending to and without becoming aware of the knowledge they have acquired. More than 50 years have passed since Reber's account of implicit learning, and definitions of the relationship between implicit and explicit learning remain a thorny issue in SLA research.

For example, as Dörnyei (2009) describes, explicit learning is "characterized by the learner's conscious and deliberate attempt to master some material or solve a problem" (p. 136), whereas implicit learning is unconscious and automatic. In line with this, DeKeyser (2003) states that "awareness" is an essential feature of the difference between implicit and explicit learning and defines implicit learning as "learning without awareness of what is being learned" (p. 314). Using the term "awareness," Leow (2015) simply describes implicit learning as "learning without awareness," while, as stated above, explicit learning as "learning with awareness." In a similar vein, but also including the factor of being able to verbalize one's knowledge, Hulstijin (2005) explains that "explicit and implicit knowledge differ in the extent to which one has or has not (respectively) an awareness of the regularities underlying the information one has knowledge of, and to what extent one can or cannot (respectively) verbalize these regularities" (p. 130). Echoing Hulstijin's view, R. Ellis (2004, 2006, 2009) also states that being verbalizable is one of the key characteristics of explicit knowledge, stating that "explicit knowledge is potentially verbalizable" (2004, p. 239).

Taken together, all these definitions, although not identical, seem to share the fundamental concept of explicit-implicit learning and knowledge, that is, learning and knowledge with or without awareness. (Although the terms "awareness" and "consciousness" have been used in the SLA domain and the non-SLA domain, respectively (Leow, 2015), these two terms are used interchangeably in this thesis.) In addition to awareness, being verbalizable appears to be another key factor that is likely

to have significant importance for WL, given that it requires learners to verbalize their thoughts in writing.

2.3.2 Attention, Awareness, and L2 Learning

The role of awareness in L2 learning has been a controversial issue in the field of SLA. According to Krashen (1981), there is a clear difference between the processes of conscious "learning" and unconscious "acquisition." He claims that second language acquisition is primarily an unconscious process and that knowledge learnt consciously is of limited use. Opposing Krashen's claim, Schmidt (1990) proposed his Noticing Hypothesis. Schmidt argues that learning is not possible without noticing, i.e., "the registration of the occurrence of a stimulus event *in conscious awareness* and its subsequent storage in long-term memory" (1994, p. 166, italics added), based on his own aforementioned experience as a learner of Portuguese (see Schmidt & Frota, 1986, for details). He further argues that noticing at the level of awareness is a "necessary and sufficient condition for the conversion of input to intake" (Schmidt, 1994, p. 209). Later Schmidt (2001) weakened his original claim concerning the role of noticing as a necessary prerequisite for learning and stated that noticing is likely to have a facilitative impact on L2 learning. He still rejected the idea, however, that learning L2 features without noticing them (i.e., conscious awareness) is possible.

It is important to note that Schmidt also made a distinction between noticing, i.e., awareness only at a very low level of abstraction (Schmidt, 2001, p. 5), and understanding, a higher level of awareness. Regarding the two, he explained:

> I use "noticing" to mean conscious registration of the occurrence of some event, whereas "understanding" as I am using the term, implies recognition of a general principle, rule or pattern. Noticing refers to surface level phenomena and item learning, while understanding refers to a deeper level of abstraction related to

> (semantic, syntactic, or communicative) meaning, system learning. (Schmidt, 1995, p. 29)

In contrast to Schmidt's (1990, 1994) position, Tomlin and Villa (1994) claim that awareness is not a prerequisite for L2 learning. In their view, the construct of attention comprises three distinct but interrelated attentional processes: alertness, orientation and detection. Alertness is defined as "an overall, general readiness to deal with incoming stimuli or data" (p. 190), whereas orientation involves the focusing of attention on a stimulus. Finally, detection functions as the cognitive registration of stimuli. Regarding the three constructs, the researchers argue that detection alone is necessary for further cognitive processing of input, the other two functioning as facilitators for detection to occur, but they are not necessary processes. It is worth noting that Tomlin and Villa consider that none of the components/processes of attention require awareness, stating that "awareness requires attention, but attention does not require awareness (p. 194). Redefining Schmidt's concept of noticing as "detection within selective attention" (p. 199), Tomlin and Villa separate awareness and learning as related but distinct processes.

Bringing together ideas from both Schmidt's (1990, 1994) and Tomlin and Villa's (1994) points of view, Robinson (1995) defines noticing as "detection plus rehearsal in short-term memory, prior to encoding in long-term memory" (p. 296). According to his model, detection is regarded as the first step in the learning process, prior to noticing, and it is also expected to be responsible for encoding language stimuli in short-term memory. As such, Robinson agrees with Tomlin and Villa's view in that detection plays an essential role for learning to occur, but he places it at a different stage of the learning process. In Robinson's view, detection occurs during the initial stages of learning, whereas Tomlin and Villa posit that detection follows alertness and orientation (i.e., later stages). In addition, in contrast to Tomlin and Villa's perspective but in line with

Schmidt's, Robinson claims that detection has to be accompanied by awareness for learning to occur, acknowledging the importance of awareness. Therefore, Robinson's model takes in aspects from both Schmidt's Noticing Hypothesis and Tomlin and Villa's model of attention. Relevant to this thesis is that he recognizes the role of awareness for L2 learning.

As briefly reviewed above, attention is generally acknowledged as a prerequisite for L2 learning in the field of SLA, while the role of awareness remains controversial, probably because of the difficulty of investigating the exact role of awareness in L2 learning. Concerning this issue, Leow, Johnson and Zárate-Sández (2011) contend that the construct of awareness "is undoubtedly the slipperiest to operationalize and measure" (p. 61). In addition, some studies report that learning without awareness is possible (e.g., J. N. Williams, 2004, 2005). Nonetheless, research on the role of awareness in L2 learning has provided empirical support for the positive impact of awareness on language learning (e.g., Leow, 1997; Rosa & Leow, 2004). Many researchers agree that awareness is likely to have a facilitative impact on L2 learning.

2.3.3 Interface or No Interface?

Like the role of awareness in L2 learning, whether there is an interface between explicit and implicit knowledge is a controversial issue. Krashen (1981) is one of the researchers who holds the view that implicit and explicit knowledge do not interface (i.e., no interface position). His Monitor Theory distinguishes between learning which involves consciousness, resulting in learnt/explicit knowledge, and acquisition which does not involve consciousness, resulting in acquired/implicit knowledge. It is important to note that a fundamental claim of the Monitor Theory is that learning functions only as a "monitor" to inspect and edit one's language (p. 156). Accordingly,

as mentioned earlier, Krashen considers consciously learnt knowledge to be of limited use, prioritizing acquired implicit knowledge over learnt explicit knowledge. Referring to his theory, the researcher argues that there is no interface (connection) between implicit (acquired) and explicit (learnt) knowledge, which has triggered a great deal of discussion/debate regarding this issue.

Now, researchers seem to generally acknowledge that explicit knowledge can facilitate the acquisition of implicit knowledge (Dörnyei, 2009; R. Ellis, 2004, 2005, 2008) (i.e., weak interface position). R. Ellis (2004) states that explicit knowledge contributes "indirectly to the acquisition of implicit knowledge by facilitating attention to form in the input" (p. 228). Some researchers assume an even stronger connection between the two types of knowledge (i.e., strong interface position) (DeKeyser, 1998, 2003, 2007; Sharwood Smith, 1981), claiming a direct interface between explicit and implicit knowledge. For example, in DeKeyser's (2003) view, "even though implicitly acquired knowledge tends to remain implicit, and explicitly acquired knowledge tends to remain explicit, explicitly learned knowledge can become implicit in the sense that learners can lose awareness of its structure over time" (p. 315). Thus, he states that proceduralized explicit knowledge may be considered functionally equivalent to implicit knowledge. DeKeyser (2007) further argues that L2 skills can be acquired through practice, referring to skill acquisition theory, which claims that:

> …the learning of a wide variety of skills shows a remarkable similarity in development from initial representation of knowledge through initial changes in behaviour to eventual fluent, spontaneous, largely effortless, and highly skilled behaviour, and that this set of phenomena can be accounted for by a set of basic principles common to the acquisition of all skills. (p. 97)

According to DeKeyser (2007), the theory posits three stages of development, declarative (initial representation of knowledge), procedural (initial changes in behaviour) and automatic (eventual fluent, spontaneous, largely effortless, and highly skilled behaviour). In the first stage, learners acquire knowledge verbally from others who possess the knowledge or through perceptive observation and analysis of others who are engaged in skilled behaviour, i.e., declarative knowledge. In the next stage, learners act on the knowledge, "turning it into a behaviour, turning 'knowledge that' into 'knowledge how' or, in more technical terms, turning declarative knowledge into procedural knowledge" (p. 98). In DeKeyser's view, the proceduralization of knowledge is "not particularly arduous or time consuming" (p. 98) for learners with relevant declarative knowledge who can make use of it to execute the target behaviour. In contrast, the final stage, automatization of knowledge, is expected to be a long and time-consuming process, which is acquired as a consequence of "a large amount of practice" (p. 98).

There seems to be a tendency to prioritise implicit knowledge among SLA researchers (N. Ellis, 2005). However, given the possibility of an interface between explicit knowledge and implicit knowledge, explicit knowledge is also likely to be important. Moreover, pointing to empirical findings regarding the benefits of explicit learning (e.g., Bowles, 2003; Leow, 2001), Leow (2015) contends that "there is really no argument or debate concerning the beneficial role of awareness in L2 development or the fact that explicit learning does promote L2 development" (p. 198).

Finally, it is also worth noting that although the declarative and procedural distinction is often used interchangeably with the explicit and implicit distinction, and they are related, they are not identical. While declarative knowledge is relatively, but not always, accessible to awareness, explicit knowledge always involves awareness

(DeKeyser, 2007, 2009). Therefore, there may be instances when declarative knowledge contains information that is not explicit. For the purposes of this thesis, however, explicit and declarative knowledge are used synonymously.

2.3.4 WL and Explicit Learning

As mentioned above, explicit learning is expected to result in explicit knowledge, which is conscious, declarative and potentially verbalizable (R. Ellis, 2004, 2006, 2009; Hulstijin, 2005). According to R. Ellis (2004), "any language task that a learner finds difficult may naturally result in an attempt to exploit explicit knowledge" (p. 239). Meanwhile, WL is a conscious process (DiCamilla & Lantolf, 1994) that entails the articulation of one's thoughts in writing "to mediate cognitively complex activities" (Swain & Deters, 2007, p. 822). As stated earlier, the slower pace of writing is considered to allow learners to consult their explicit knowledge (J. Williams, 2008). Therefore, explicit learning and WL seem to be compatible, in that they both involve consciousness and verbalization, which are likely to be used when learners are faced with challenging tasks.

Explaining explicit knowledge from the perspective of SCT, R. Ellis (2009) states that "explicit knowledge can be viewed as a 'tool' that learners use to mediate performance and achieve self-control in linguistically demanding situations" (p. 13). Given this, learners are likely to make use of "the tool," i.e., explicit knowledge, by engaging in WL in order to achieve self-control in linguistically demanding situations. Although "no perfect tests or procedures exist for distinguishing the results of implicit and explicit learning" (DeKeyser, 2003, p. 320), WL is likely to facilitate explicit learning processes, promoting the establishment of explicit knowledge.

Explicit learning can take place in both instructed and uninstructed settings. Nonetheless, the nature of the end product, i.e., explicit knowledge, is likely to differ depending on the setting. Namely, when learners are provided with grammar rules by their instructor or textbook and make use of them, they are likely to acquire "externally provided explicit knowledge" (Roehr-Brackin, 2014, p. 775). Meanwhile, when they deduce and identify grammar rules in the language input that they encounter on their own, the knowledge obtained in this way is expected to result in "internally-derived explicit knowledge" (p. 775). Given that WL is a process that provides learners with opportunities to write down their thoughts regarding their linguistic issues (i.e., an inside-out learning process), WL is likely to facilitate the acquisition of internally-derived explicit knowledge.

Comparing implicit and explicit learning, Roehr-Brackin (2014) states that explicit learning is potentially faster and more efficient, depending less on frequency of exposure to input than does implicit learning (p. 776). Considering that WL is expected to assist learners when engaging in explicit learning processes, WL may be hypothesized to contribute to even faster and more efficient explicit learning than learning without WL. Furthermore, regarding the development of procedural knowledge, DeKeyser (2007) stresses that the "sequence of proceduralization and automatization cannot get started if the right conditions for proceduralization are not present (the declarative knowledge required by the task at hand, and a task set-up that allows for use of that declarative knowledge)" (p. 100). Put differently, in order to allow learners to proceduralize and further automatize their knowledge, a task that encourages learners to make use of their declarative knowledge appears to be necessary. Considering that WL is a conscious process that is likely to require learners to use their declarative knowledge, the opportunity to use WL may create "the right conditions."

2. 3. 5 Written Corrective Feedback, WL, and Explicit Knowledge

As stated, Krashen (1981) triggered a great deal of controversy, stating that implicit and explicit knowledge do not interface. Truscott (1996) was another researcher who prompted a great deal of debate from the same perspective (i.e., no interface position) in the field of WCF. Discussing feedback from an SCT perspective, Bitchener and Storch (2016) state that "feedback, including written CF, constitutes a form of assistance" (p. 69). They further explain that WCF can scaffold learners' development, with language functioning as a primary mediational tool. Moreover, referring to languaging studies (e.g., Knouzi et al., 2010; W. Suzuki, 2012; Swain et al., 2011), Bitchener and Storch state that self-directed language enables learners to engage in self-scaffolding. That is to say, learners are likely to solve their linguistic issues with their self-directed language, such as self-directed questions and self-explanations. Given this, although WL is different from WCF, WL may assist language learning as a form of self-feedback. In this section, therefore, the debate started by Truscott in the field of WCF, referring to explicit/implicit knowledge, is reviewed.

In his controversial review article, Truscott (1996) argued that "grammar correction has no place in writing classes and should be abandoned" (p. 361), thus denying the facilitative role of WCF in learners' development. As stated above, his argument triggered a great deal of controversy and many empirical studies were conducted in order to counter his argument (e.g., Chandler, 2003; Ferris, 1999, 2004). Ferris (1999), among other researchers, challenged his claim in her study entitled "The case for grammar correction in L2 writing classes: A response to Truscott (1996)" and produced findings to prove that WCF can help learners write more accurately. Empirical findings to support her argument have been reported since her study (e.g., Bitchener, 2008; Sheen, 2007a; Stefanou, 2014). For example, in Bitchener and Knoch's (2010)

longitudinal study, the participants who received WCF outperformed those in a control group on an immediate posttest and three delayed posttests conducted over ten months. More recently, Shintani, R. Ellis, and W. Suzuki (2014) also reported that participants receiving WCF improved their accuracy significantly more than those in a control group on some immediate posttests and delayed posttests conducted two weeks after the treatment (see Bitchener & Storch, 2016, for a review of these studies).

Against this background, now, researchers seem to generally acknowledge that WCF is beneficial to learners. Like Krashen's (1981) claim regarding the interface/no interface debate, therefore, the field of SLA research seems to have moved on from Truscott's (1996) argument regarding the ineffectiveness of WCF. It should be pointed out, however, that Truscott's argument is based on his aforementioned standpoint (i.e., non-interface position) regarding explicit and implicit knowledge. Therefore, arguing that WCF does not contribute to the acquisition of "the important type" (p. 346) of knowledge, i.e., implicit knowledge, Truscott admits that WCF may be effective for the acquisition of explicit knowledge. Along the same lines, Polio (2012) states that "I have assumed … that written corrective feedback will increase explicit knowledge" (p. 386). Similarly, but with more certainty, Bitchener (2012) maintains that "we do know that written CF can play a role at least in terms of developing explicit knowledge" (p. 361). As such, an ongoing issue seems to be whether WCF can facilitate the acquisition of implicit knowledge as well.

What is relevant to the current thesis is that, as stated above, WCF is expected to facilitate the acquisition of explicit knowledge (Bitchener, 2012; Polio, 2012). Supposing the aforementioned speculation that WL might function as a kind of self-feedback is correct, the facilitative impact of WCF in developing explicit knowledge

may support the hypothesis of this thesis, i.e., WL is likely to facilitate the acquisition of explicit knowledge.

CHAPTER III

APTITUDE AND METALANGUAGE KNOWLEDGE

In this chapter, focusing on learners' individual differences in aptitude and metalanguage knowledge, how these differences might influence the extent to which WL affects L2 learning is discussed. Finally, the research questions that guided the present thesis are presented.

3.1 Aptitude

3.1.1 Definitions of Aptitude

Of all the individual differences, aptitude has been recognised as one of the primary variables in the field of SLA research (Dörnyei, 2010) and identified to correlate strongly with L2 proficiency (Ehrman & Oxford, 1995). Carroll (1981), who was a cognitive psychologist and has been a dominating figure in language aptitude research (Skehan, 2012), defined foreign language learning aptitude as including the characteristics of learners which control the rate of their progress in learning a foreign language. Similarly, focusing on the rate of progress that learners are likely to make but also referring to the final language-learning outcome, Wesche (1981) describes aptitude as the "ability to learn a new language quickly and to a high degree of proficiency" (p. 119).

Meanwhile, Dörnyei and Skehan (2003) define aptitude as "a specific talent for learning foreign languages that exhibits considerable variation between learners" (p. 240). Likewise, Robinson (2005) states that "L2 learning aptitude is characterized as strengths individual learners have—relative to their population—in the cognitive abilities" (p. 46). As is clear from this statement, in Robinson's view, aptitude

comprises multiple strengths and abilities instead of a single strength or ability. He further states that learners draw on their strengths "during L2 learning and performance in various contexts and at different stages" (p. 46). Supporting Robinson's perspective, Kormos (2013) argues that "language-learning aptitude is not a unitary construct, but rather a conglomerate of different abilities that can assist in the different stages and processes of language learning" (p. 140).

Similarly, stating that aptitude consists of multiple components, Skehan (2002) relates SLA processes to aptitude components (see Table 3.1). In his view, the role of aptitude components alters in relation to learners' L2 development. As shown in Table 3.1, he suggests nine developmental stages on four main levels: noticing (Stage 1), patterning (Stages 2–5), controlling (Stages 6–8), and lexicalising (Stage 9).

Defining aptitude is no easy task. According to Dörnyei and Ryan (2015), when scholars refer to language aptitude, "what is really meant by the concept is 'the results of a language aptitude test'" (p. 46). While Robinson (2013) states that aptitude is measurable like height, he also states that, unlike height, language aptitude cannot be directly observed but has to be inferred from performance on psychological tests designed to measure it. Against this background, aptitude tests have played a significant role in aptitude research. Therefore, aptitude is further discussed by introducing the major aptitude tests created so far, with the background to their creation, in the next section.

Table 3.1
SLA Processing Stages and Potential Aptitude Components

SLA Processing Stage	Aptitude Component
1. noticing	auditory segmentation attention management working memory phonemic coding
2. pattern identification	fast analysis/working memory grammatical sensitivity
3. extending	inductive language learning ability
4. complexifying	grammatical sensitivity inductive language learning ability
5. integrating	restructuring capacity
6. becoming accurate, avoiding errors	automatisation proceduralisation
7. creating a repertoire, achieving salience	retrieval processes
8. automatising rule-based language, achieving fluency	automatisation proceduralisation
9. lexicalising, dual-coding	memory, chunking, retrieval processes

Reproduced from Skehan (2002, p. 90)

3.1.2 Measurement Instruments for Aptitude

Although the concept of foreign language aptitude was discussed as early as 1575 by a Spanish physician (Carroll, 1981), tests to measure it were not devised until the 20th century. The first language aptitude tests were developed in the United States in the 1920s and ’30s, with the aim being to “increase the cost effectiveness of language education” (Dörnyei & Ryan, 2015, p. 48). Around that time, the government of the

United States was spending a great deal of money to train large numbers of people in foreign languages. As such, creating aptitude tests to predict the rate of learning was of great importance and had practical value for the government in order to select those who were likely to be quick and successful learners.

Regarding the use of aptitude information, Skehan (2012) points to four educational aims that such information might be used to achieve: (1) selection, (2) counselling, (3) remediation, (4) instructional modification. Of the four aims, given the aforementioned background, early aptitude tests seem to have focused solely on selection.

3.1.2.1 The Modern Language Aptitude Test

In the 1950s and ’60s, “the golden period of scientific language aptitude testing” (Rees, 2000), more aptitude tests were developed with the same aim as the early ones (i.e., selection). The Modern Language Aptitude Test (MLAT) developed by Carroll and Sapon (1959) is one of the tests from this period and is probably the most famous aptitude instrument ever devised. Although it is still widely used even today, its initial purpose was to screen candidates for foreign language instruction at the Foreign Service Institute in the United States (R. Ellis, 2008).

Reflecting on the developmental processes of the MLAT, Carroll and Sapon (1959) state that the MLAT was the outcome of their five-year research study (p. 3). According to Skehan (2012), Carroll made contributions to the field of aptitude research in two areas. First, as a theorist, he proposed his four-factor view of aptitude. His second and more practical contribution was the development of the MLAT with Sapon. It should be pointed out, however, that his contributions did not occur in that order. More specifically, as the reflection above demonstrates, after five years of piloting sub-

tests, Carroll and Sapon completed the MLAT, keeping five subtests that produced good results (i.e., number learning, phonetic script, spelling clues, words in sentences, paired associations). Subsequently, examining data obtained from the MLAT, Carroll (1981) identified four constituent abilities of aptitude, i.e., phonetic coding ability, grammatical sensitivity, rote learning ability and inductive language learning ability, which is "post hoc theorizing" in Dörnyei and Ryan's (2015, p. 51) words (see VanPatten & Smith, 2015, for a similar discussion). Below is a description of the four major components and subtests of the MLAT:

1. phonetic coding ability – the ability to identify distinct sounds, form associations between those sounds and the symbols representing them, and retain these associations
   (Part 1: number learning, Part 2: phonetic script, Part 3: spelling clues)
2. grammatical sensitivity – the ability to recognize the grammatical functions of words (or other linguistic entities) in sentence structures
   (Part 4: words in sentences)
3. rote learning ability for foreign language materials – the ability to learn associations between sounds and meaning rapidly and efficiently, and to retain those associations
   (Part 5: paired associations)
4. inductive language learning ability – the ability to infer or induce the rules governing a set of language materials, given samples of language materials that permit such inferences

Although Carroll identified four components, not all of them are represented in

the MLAT. To be more specific, the fourth construct, inductive language learning ability, is not included in any subtests of the MLAT. Meanwhile, Skehan (1998) suggests that Carroll's two components, grammatical sensitivity and inductive language learning ability, can be combined into one, i.e., language analytic ability, and defines this as "the capacity to infer rules of language and make linguistic generalizations and extrapolations" (p. 204).

3.1.2.2 Pimsleur's Language Aptitude Battery

Another famous battery of aptitude tests developed in the golden period is Pimsleur's Language Aptitude Battery (PLAB) (Pimsleur, 1966). Following Carroll and Sapon's (1959) procedure, Pimsleur, "a foreign language teacher with a strong background in educational psychology and testing" (Carroll, 1981, p. 94), created the PLAB by first trying out instruments which he devised and then constructing a theory of language aptitude from data he collected. The PLAB, however, is different from the MLAT on some points.

First, the main target is high school students. Second, as mentioned above, although inductive language learning ability, one of the four constructs identified by Carroll, is not represented in the MLAT, the PLAB does specifically measure this construct. Third, it "places greater emphasis on auditory factors, and less on memory" (Dörnyei & Skehan, 2003, p. 594), which is clear from the three factors given by Pimsleur in conceptualizing an aptitude for language learning, i.e., verbal intelligence, motivation and auditory ability (Dörnyei & Ryan, 2015). The decision to emphasize auditory factors stems from Pimsleur's hypothesis that some high school students under-achieve in language courses because of their deficient auditory skills. He believes that "use of the PLAB should enable early diagnosis of remedial learning difficulties in

high-school foreign language programs" (Dörnyei & Skehan, 2003, p. 594), which seems to correspond to the aforementioned educational aims proposed by Skehan (2012), i.e., remediation and counselling.

3.1.2.3 Aptitude Tests in the 1970s, '80s and Aptitude Renaissance (CANAL-FT and LABJ)

In the 1970s and '80s, aptitude research lost momentum (Dörnyei & Ryan, 2015, Skehan, 2002, 2012), probably for the following three reasons. First, aptitude was considered "anti-egalitarian" (Skehan, 2002, p. 72), as making decisions based on abilities that people are born with was deemed to deny the value of individual effort. Second, aptitude was associated with outmoded instruction styles, i.e., audiolingual methodologies, which were prevalent when the MLAT and the PLAB were created. (The audiolingual approach is characterised by drills, rote learning and the development of explicit grammar knowledge/teaching.) Third, related to the second point, as aptitude tests were developed to predict the speed of learning in *instructed contexts*, some researchers, such as Krashen (1981), claimed that aptitude is only a relevant concept for learning in classroom settings, not in natural ones. Against this background, the publication of "Language Aptitude Reconsidered," by Parry and Stansfield (1990), triggered the "aptitude renaissance" (Dörnyei & Ryan, 2015, p. 59) that lasted for the next 10 years or so. Referring to the fact that the aptitude tests widely used then (e.g., the MLAT, the PLAB) were becoming "out of date" (p. 2), Parry and Stansfield contended that new insights revealed in the field of cognitive psychology into human learning processes, language learning processes in particular, should be taken into account. They further argued that the notion of aptitude needed to be expanded and refined, relating to factors other than learners, such as the language to be learned and the

level of proficiency to be obtained. At that time, new types of aptitude tests whose focus was not solely on the prediction of language learning success, as suggested by Parry and Stansfield, were being developed, including the two tests described below.

One of these two tests, the "Cognitive Ability for Novelty in Acquisition of Language as Applied to Foreign Language Test" (CANAL-FT), was developed by Grigorenko, Sternberg, and Ehrman (2000) based on theory instead of post hoc theorising relied on for the development of the MLAT and the PLAB (Dörnyei & Ryan, 2015). As the name of the test implies, it aims to measure how people cope with novelty and ambiguity in their learning from the perspective of cognitive psychology. For this aim, language aptitude is measured based on the analysis of acquisition processes of Ursulu, an artificial language (Skehan, 2012). Although Grigorenko et al. argue that their "work should be viewed as a foundation for further development rather than as a completed effort," no tests to develop their CANAL-FT further have been created so far. Interpreting this situation, Dörnyei and Ryan (2015) give two possible reasons. First, the CANAL-FT is not widely available. Second, it seems to perform similarly to earlier tests, i.e., it does not yield better statistical predictions than the MLAT.

The other test, the Language Aptitude Battery for Japanese (LABJ), is also one of the aptitude tests developed during that period by a Japanese researcher, Sasaki (1991, 1996). Rather than merely predicting learners' rate of learning, she was more interested in elucidating the relationships among L2 proficiency, aptitude and intelligence. Therefore, defining aptitude as "a set of cognitive abilities related to both foreign language learning success and to the process itself" (1996, p. 8), Sasaki investigated the relationships among the three factors (i.e., L2 proficiency, aptitude and intelligence) among 160 Japanese university students with an average of 7.3 years of foreign language (English) learning experience, and identified that the three factors are not

identical, but correlate closely. The LABJ is an aptitude test created for the study. It is a Japanese translation of the MLAT and the PLAB for L1 Japanese participants. As the current study project is also conducted with Japanese university students with similar previous foreign language learning experience (English with 7.3 years of experience on average), part of the LABJ is employed as a measure of aptitude.

3.1.2.4 Aptitude Tests in the 21st Century (LLAMA and Hi-LAB)

Since the aptitude renaissance, "aptitude research has been in hibernation for a long period of time" (Li, 2013, p. 637) and has once again become a "relatively neglected area" (Robinson, 2013). Meanwhile, Skehan (2012) is more optimistic, stating that "aptitude research has recently been one of the revitalized areas in applied linguistics and SLA" (p. 381), especially in contrast to his own remarks in 2002, that is, "research and theorising on foreign language aptitude has languished over the last thirty years" (p. 69). Although it is not clear where aptitude research currently stands, two of the aptitude tests created since the turn of the century, the LLAMA and the High-level Language Aptitude Battery (Hi-LAB), seem to be worth mentioning. In his recent review of aptitude research, Skehan (2015) divides aptitude tests into two categories, ones that follow the aptitude structure proposed by Carroll (1981) and others that draw on contemporary research in SLA and cognitive psychology (p. 369). According to this categorisation, the LLAMA falls into the former category and the Hi-LAB (and the CANAL-FT) into the latter.

As stated above, the LLAMA (Meara, 2005) is a set of aptitude tests based on the MLAT. What is innovative about the LLAMA, however, is that it was created by "technologically updating the classic Carrollian approach" (Dörnyei & Ryan, 2015, p. 58), i.e., it is computer-based, which was not possible when the MLAT was created in

1959. In addition, it is available as a free download (http://www.lognostics.co.uk/tools/llama/) to anyone with a computer and Internet access. It is a revised version of the Swansea Language Aptitude Test, which "included materials loosely based on Polish and Turkish that made stimuli more familiar to test-takers who spoke Hungarian or Azeri" (Granena, 2013, p. 107). Unlike its antecedent, the LLAMA is language independent, because an artificial language is used in this version of the test so that it is accessible to people with any L1. It consists of four sub-tests: LLAMA_B, a test of vocabulary learning, LLAMA_D, a test of sound recognition that requires previously heard sound sequences to be identified in new sequences, LLAMA_E, a test of sound-symbol associations, and LLAMA_F, a test of grammatical inferencing. As WL/self-explaining is expected to involve inferencing, LLAMA_F was employed in the current thesis.

In contrast to the LLAMA, the Hi-LAB (Doughty, Campbell, Bunting, Bowles, & Haarmann, 2007; Linck, Hughes, Campbell, Silbert, Tare, Jackson, & Doughty, 2013) was created based on research findings from cognitive psychology and highlights the role of memory (i.e., phonological short-term memory, associative memory and long-term memory), implicit learning and auditory discrimination. There are two other distinctive features that make the Hi-LAB stand out. First, the Hi-LAB, as the name of the test implies, was designed to identify a group of cognitive and perceptual abilities with the aim of discerning individuals capable of attaining high-level proficiency. Namely, positing aptitude as "a measurable ceiling on language learning" (Doughty, 2013, p. 153), the Hi-LAB intends to predict learners' ultimate attainment. It is in sharp contrast to the LLAMA and other aptitude tests widely used today (e.g., the MLAT, the PLAB), which are designed to measure mainly the rate of learning in the early stages. Another notable difference is its omission of material to assess the learning of patterns

(Skehan, 2015), which is not surprising given that patterning is associated with the early processing stages according to Skehan's (2002) aptitude profile model.

3.1.3 Aptitude Treatment Interactions

As stated above, initially, the main goal of aptitude tests was to predict learners' rate of progress for selection purposes. However, a shift in the focus of aptitude research eventually occurred and researchers started to explore ways to connect language aptitude to other pedagogical issues in SLA. As reported already, Pimsleur (1966) put emphasis on auditory skills in his PLAB to identify learners who were likely to be under-achievers. In this way, he intended to offer instruction to help avoid later problems, which is an example of the use of aptitude information for remediation, one of the aims proposed by Skehan (2012). Other researchers started to think of ways to use aptitude information to offer treatments that match learners' aptitude profiles in order to maximise the effect of treatments. This line of research has become known as research on aptitude treatment interactions (ATI) in educational psychology. It seems to resonate with another educational aim proposed by Skehan (2012), instructional modification.

In his seminal paper entitled "How can instruction be adapted to individual differences?" Cronbach (1967) made four suggestions to answer his own question. Namely, he proposed the adaptation of instruction: (1) within a predetermined programme, (2) by matching goals to the individual, (3) by erasing individual differences and (4) by altering the instructional method. In his view, an individual's learning rate should vary, depending on the nature of the instruction. Therefore, pointing to the significance of the last suggestion, Cronbach contended that adapting instructional techniques to learners' aptitude profiles should be essential, rather than

merely altering the duration of exposure in the long run. He further argued that "aptitude information is not useful in adapting instruction unless the aptitude and treatment interact" (p. 30), emphasising the importance of ATI.

Similarly, in the field of SLA, the significance of ATI has been increasingly recognised (DeKeyser, 2012; Robinson, 2002, 2012) (see Li, 2015; Skehan, 2015; Vatz, Tare, Jackson, & Doughty, 2013, for recent reviews). In his pioneering research on ATI, Robinson (2002) states that a major aim of pedagogically oriented language aptitude research is to profile individual differences in cognitive abilities and match these profiles to effective instructional options, such as types of pedagogic tasks, interventionist "focus on form" techniques, and more broadly defined learning conditions. So far, SLA researchers have conducted studies to explore the interaction between aptitude and instruction and produced some evidence of ATI, demonstrating that certain aptitude abilities facilitate L2 learning, depending on the types of instructional treatment (e.g., Erlam, 2005; Li, 2013; Sheen, 2007a, 2007b; Stefanou & Révész, 2015; Wesche, 1981; Y. Yilmaz, 2013). Some key studies are summarised below.

One of the earliest ATI studies was conducted by Wesche (1981) in the context of a Canadian government French language training programme for civil servants. Learners were usually placed into one of the three types of language instruction, i.e., an audio-visual method, an analytical approach, and a functional approach, in such a way that the instruction type matched their aptitude profiles based on the results of the MLAT and the PLAB. For the purposes of this study, however, a group of learners who were diagnosed as analytical were divided into two groups. One group was assigned to instruction that matched the learners' aptitude profiles (i.e., analytical approach) and the other was assigned to instruction that mismatched their profiles (i.e., audio-visual

method). Achievement tests conducted after three months of training demonstrated a statistically significant improvement for the group that received instruction that matched the learners' profiles, whereas comparable results were not observed for the group that received mismatched instruction. Furthermore, the learners in the former group expressed greater satisfaction regarding their instruction. Based on these results, Wesche concluded that learners achieve significant improvement when they receive instruction that matches their aptitude profiles.

Similarly, a significant correlation between aptitude and a certain instruction method was identified by Erlam (2005), who also employed the MLAT and the PLAB as aptitude measures for grammatical sensitivity (which she broadly refers to as language analytic ability), phonemic coding ability and working memory. The participants were high school students in New Zealand who were learning French and assigned to three instruction groups (deductive instruction, inductive instruction and structured input instruction). Before and after receiving three 45-minute lessons on the target construction, i.e., French direct object pronouns, all participants took pre-, post- and delayed posttests over three time periods, which were used to measure their learning. The association between gain scores and aptitude tests results was examined. Of the three groups, the deductive instruction group, who were provided with grammar rules as well as opportunities to produce language, demonstrated fewer statistically significant correlations compared to the other two groups. Interpreting this result, Erlam states that deductive instruction seemed to "minimize or level out any effect that individual differences in language aptitude may have with respect to instructional outcomes" (p. 163) because of its explicit nature. To borrow Cronbach's (1967) words, deductive instruction might have erased learners' individual differences. It should be pointed out, however, that the aptitude tests were conducted six months after a delayed

posttest. Therefore, the participants were not assigned to the instructional methods they received based on their aptitude profiles, which might be considered a weakness in terms of the research design (Vatz et al., 2013).

Some studies have produced evidence to support Erlam's (2005) findings (e.g., Li, 2013; Stefanou & Révész, 2015; Trofimovich, Ammar, & Gatbonton, 2007). Trofimovich et al., for example, probed the relationship between the effects of computerized recasts (i.e., implicit feedback) and grammatical sensitivity (broadly referred to as language analytic ability) measured by a French translation of the MLAT Part IV. The participants, 32 French-speaking adult ESL learners in Canada, worked on a picture description task that targeted possessive determiners (i.e., his, her). An examination of the results of posttests and aptitude tests revealed a positive correlation between the participants' grammatical sensitivity and their ability to benefit from recasts. Put differently, the higher the grammatical sensitivity the participants possessed, the greater the effects of recasts (implicit feedback) in learning grammar items.

Similar results were obtained by Stefanou and Révész (2015), who investigated the effectiveness of two types of direct written corrective feedback in relation to grammatical sensitivity measured by a Greek translation of the MLAT with a pre- and posttest design. The participants were 89 EFL Greek high school students, who were divided into three groups depending on the type of feedback they received, a direct feedback only group, a direct metalinguistic group and a comparison group who received no feedback. The direct feedback only group literally received exactly that, whereas the direct metalinguistic group received direct feedback as well as relevant metalinguistic information. Therefore, both types of feedback were explicit but direct metalinguistic feedback was more explicit than direct feedback only. After completing a written text summary task, the participants received feedback on their non-target-like

article use with specific and generic plural referents (i.e., definite plural, bare plural, demonstrative plural). Examining the results of pre- and posttests and aptitude tests, the researchers identified statistically significant correlations between grammatical sensitivity and the direct feedback only group and the control group, but not the direct metalinguistic group. Interpreting the results, Stefanou and Révész state that more explicit feedback due to additional metalinguistic information might have "enabled the participants to compensate for their potentially weaker ability to recognise grammatical functions" (p. 277). The results obtained by Trofimovich et al. (2007) and Stefanou and Révész lend further support to Erlam's (2005) claim that explicit feedback might neutralise learners' individual differences in language aptitude.

Meanwhile, aptitude has also been found to interact with instruction in other studies (e.g., Sheen, 2007a, 2007b; Y. Yilmaz, 2013), but the findings are somewhat different from those reported above, in that learners with higher aptitude were observed to benefit more from explicit instruction. Sheen (2007a), for example, explored the effect of two types of written corrective feedback, direct-only correction and direct metalinguistic correction (both explicit but to different degrees, as in the study by Stefanou and Révész (2015)), in relation to learners' inductive language learning ability (broadly referred to as language analytic ability in her study). The direct-only group received a traditional error correction strategy that consists of indicating the location of an error and provision of the correct form, whereas the direct metalinguistc correction group received metalinguistic explanations in addition to the same treatment as the direct-only group. The aptitude measurement instrument employed was a language analysis test created by Ottó (2002). The participants were asked to find English translations for sentences in an artificial language with the help of a glossary consisting of words and sentences in that language.

The target structure was English definite and indefinite articles, *a* as the first mention and *the* as anaphoric reference. Regarding the rationale of the choice of structure, Sheen (2007a) explains that the participants, 91 ESL learners, were not explicitly taught articles in the semester and errors regarding articles were corrected infrequently because of their non-saliency and the complicated rules associated with them. The participants completed narrative retelling tasks that elicited both definite and indefinite articles. The researcher found a stronger correlation between aptitude measures and gains on immediate and delayed posttests for the direct metalinguistic correction (more explicit) group than for the direct-only (less explicit) group.

Similar results were obtained in another study by Sheen (2007b), which compared the effects of two corrective feedback types, i.e., recasts (implicit feedback) and metalinguistic correction (explicit feedback) in relation to inductive language learning ability (broadly referred to as language analytical ability). The same aptitude test, target structure and treatment tasks were used as in Sheen (2007a), and the participants were 80 ESL learners. Echoing the findings of Sheen (2007a), analysis of the results of pre- and posttests demonstrated a significant correlation between the participants' gain scores and aptitude for metalinguistic correction (explicit instruction), but not for recasts (implicit instruction).

Similar findings emerged from a more recent study by Y. Yilmaz (2013), who probed the correlation between the effect of feedback type (explicit and implicit) and learners' inductive language learning ability (referred to as language analytical ability) measured by the LLAMA_F. The participants were 48 adult native speakers of English who had no previous exposure to the target language, i.e., Turkish. They were instructed to learn 59 Turkish words online individually before the experiment, and then they carried out tasks that targeted the Turkish plural morpheme and locative case

morpheme. During the tasks, they received feedback in the form of recasts or explicit correction, which was operationalised as "the direct rejection of the learner's production followed by the target-like reformulation of the erroneous segment of the learner's production in a direct manner" (p. 352). Echoing the findings of Sheen (2007a, 2007b), a statistically significant correlation between posttest performance and language analytical ability was identified for the explicit feedback group but not for the recast group. It is worth mentioning, however, that no metalinguistic information (rules or terminology) was provided, probably making the feedback less explicit than that used in Sheen (2007a) and Stefanou and Révész (2015). To be more precise, the explicit correction in Y. Yilmaz's study corresponds to direct only and direct feedback only in Sheen (2007a) and Stefanou and Révész (2015), respectively, which were categorised as "less explicit feedback" in those studies.

As mentioned above, conflicting findings were reported regarding the correlation between learners' aptitude and type of feedback (i.e., implicit vs explicit feedback). Some researchers have reported a significant correlation between learners' aptitude (i.e., grammatical sensitivity, inductive language learning ability) and implicit feedback (e.g., Erlam, 2005; Stefanou & Révész, 2015; Trofimovich et al., 2007), but no significant correlation with explicit feedback. These results may not be surprising given that learners do not need to rely on their high aptitude under explicit learning conditions where they are provided with explicit instruction or feedback, such as grammar rules and metalinguistic information. In contrast, high aptitude is likely to be an influential factor under implicit learning conditions where no such instruction is provided and learners need to compensate for the lack of the external help with their higher aptitude in order to achieve gains. Nonetheless, others have identified a positive correlation

between aptitude ability and explicit feedback (e.g., Sheen, 2007a, 2007b; Y. Yilmaz, 2013).

An interpretation to account for this discrepancy is proposed by Li (2013), who also found statistically significant correlations between aptitude and implicit corrective feedback, supporting the findings of Erlam (2005) and other researchers. He investigated the interactions between the effects of implicit or explicit feedback (i.e., recasts and metalinguistic correction, respectively) and grammatical sensitivity (which he broadly refers to as language analytic ability) measured by the subtests of the MLAT. The participants were 78 Chinese learners from two universities in the United States, who were given feedback in response to their non-target-like use of Chinese classifiers. The analysis of pre- and posttests scores revealed a significant correlation between grammatical sensitivity and implicit feedback (i.e., recasts), which is in line with the findings of Erlam (2005), Trofimovich et al. (2007), and Stefanou and Révész (2015), but differs from those of Sheen (2007a, 2007b) and Y. Yilmaz (2013).

With respect to this inconsistency, Li (2013) points to the difficulty of linguistic targets as a possible contributor, stating that "whether language analytic ability [= grammatical sensitivity] influences the effects of implicit feedback (recasts) is also constrained by the extent to which the linguistic target is within learners' processing capacity" (p. 647). According to the researcher, the target structures used in his and Trofimovich et al.'s (2007) studies (i.e., Chinese classifiers and English possessive determiners, respectively) were not likely to have involved complex form-meaning mapping, i.e., they were within the participants' processing capacity. Therefore, the participants with higher aptitude were likely to have recognised the grammatical functions with the mere help of implicit instruction, using their higher aptitude as an additional resource and resulting in a significant correlation between aptitude and

implicit feedback. Meanwhile, no such correlation between higher aptitude and explicit feedback was observed, probably because higher aptitude was not likely to be needed for simpler target forms under the explicit feedback condition. Furthermore, for those with lower aptitude, implicit instruction was not likely to be enough to develop their interlanguage, demonstrating no significant correlations as a consequence. This reasoning seems to apply to the target structure employed in the study by Erlam (2005), as her participants were intermediate-level learners (high school students enrolled on a French course) who were expected to have some knowledge of the target structure.

In contrast, Li (2013) speculates that the target structure employed by Sheen (2007a), English articles, might have been too difficult even for high aptitude learners, only with the aid of implicit feedback, to extract rules from. In his view, they probably needed both higher aptitude and explicit feedback in order to achieve improvement, resulting in a significant correlation between grammatical sensitivity and explicit feedback. Given that the same target structure was employed in Sheen (2007b), it seems natural to assume that the same interpretation could be applied to that as well. It should be pointed out, however, that English articles were also used as the target construction in the study by Stefanou and Révész (2015), who identified a significant correlation between the participants' aptitude and implicit feedback, but not explicit feedback. This discrepancy may be explained by the difference in the participants' prior knowledge of the target construction. As stated earlier, on the one hand, the participants in the Stefanou and Révész study were high school students who had had over seven years of formal education and were expected to have some knowledge of the target. On the other hand, the participants in Sheen's studies were ESL learners who were not taught the target explicitly, suggesting that learners' prior knowledge of the target construction may be another key.

Regarding the study by Y. Yilmaz (2013), in addition to the possible difficulty of the target structures (i.e., Turkish locative and plural), given that the participants had no prior knowledge of the target language (Turkish), extracting the rules of the new language might have been beyond their processing capacity. Therefore, in order to achieve gains, they are likely to have needed both high aptitude and explicit feedback, resulting in a significant correlation between aptitude and explicit feedback. (It should be noted, as mentioned earlier, that the "explicit" feedback in Y. Yilmaz' study was not as explicit as that used in previous studies in its operationalization, which might also have affected the result.)

Based on the findings of his own and previous studies, Li (2013) hypothesises that:

> Other things being equal, language analytic ability is implicated in implicit conditions in the learning of easy, transparent structures that are within one's processing capacity, and in explicit conditions in the learning of hard, opaque structures where the internalization of available metalinguistic information sets heavy processing demands on internal cognitive resources. (p. 648)

Some supporting evidence for Li's hypothesis comes from Yalçin and Spada (2016), who investigated the association between language aptitude and the learning of two target structures with different grammatical difficulty, i.e., the past progressive (easy structure) and passive voice (difficult structure). The participants were 66 EFL learners in the eighth grade in Turkey, who received the same instruction on both target structures (i.e., explicit information about the formal properties of the target features and opportunities to practise them). Their language aptitude was measured by all the subtests of the LLAMA, but what is most relevant to the current thesis is that the researchers identified a significant correlation between grammatical inferencing ability

(measured by LLAMA_F) and the passive voice (difficult structure), but not the past progressive (easy structure) in spite of the same instruction they received. Analysing the results, Yalçin and Spada explain that the participants were likely to have needed higher language aptitude for the difficult target structure to achieve gains, but it was probably not the case with the easy structure, resulting in no significant correlation. Based on their findings, the researchers state that the nature of the target structure may interact with the type of L2 instruction and language aptitude, supporting Li's hypothesis.

In a similar vein, referring to the relationship between the difficulty of target structures and types of feedback (implicit vs explicit), Skehan (2015) explains that both higher language aptitude and explicit feedback seem to be needed in order to improve interlanguage knowledge for difficult structures because of factors such as redundancy and/or non-saliency, whereas either higher language aptitude or explicit feedback is likely to be sufficient for easy structures. Regarding this issue, he highlights the importance of paying attention to the "joint effects" of both the nature of target structures and types of corrective feedback. Given these findings, ATI may not be enough to consider the possible interaction thoroughly. More specifically, future research may need to consider ATTI, i.e., aptitude treatment and *target structure* interaction.

3.1.4 Summary of Aptitude Research and WL

As reported above, initially, aptitude information only had practical use. That is to say, the main use of aptitude information was for the prediction of learning rate for selection purposes (Dörnyei & Ryan, 2015). Over the years, however, it expanded and started to be used for pedagogical purposes, such as counselling, remediation, and instructional modification, as suggested by Skehan (2012). Regarding this point, Li

(2015) states that language aptitude can be defined in two different ways, i.e., "as a variable that is predictive of ultimate L2 attainment and one that interacts with contextual factors in affecting L2 outcomes" (p. 388) in his review of aptitude research. The ATI research has shown that the impact of instructional techniques can be enhanced or diminished depending on learners' language aptitude profiles (e.g., Erlam, 2005; Sheen, 2007a, 2007b; Stefanou & Révész, 2015: Y. Yilmaz, 2013). Moreover, the difficulty of target structures has been suggested to influence the aptitude and treatment interaction (Li, 2015; Skehan, 2015). Given that WL is intended to be a kind of instructional technique, it was deemed important to examine how WL might interact with aptitude profiles, in order to explore whether learners with certain types of learner aptitude profiles are more likely to benefit from engaging in WL than others. To this end, it was decided to examine a possible association between language aptitude and learning from WL.

As stated, language aptitude is believed to consist of multiple abilities/components (Kormos, 2013; Robinson, 2005; Skehan, 2002). Of these abilities, inductive language learning ability appears to be relevant to WL, considering that WL is expected to involve inference generation. Also, given that WL requires learners to solve linguistic problems on their own, grammatical sensitivity is likely to be another key factor influencing the impact of WL on L2 learning. Therefore, as already mentioned, these two abilities, i.e., inductive language learning ability and grammatical sensitivity, were investigated in the current study. As discussed above, none of the subtests of the MLAT measure inductive language learning ability. Therefore, the subtests of the LLAMA and the LABJ were employed to examine inductive language learning ability. With respect to grammatical sensitivity, a version of the MLAT adapted for Japanese was used (see section 5.5.1 for details).

3.2 Metalanguage Knowledge and WL

As the definition of WL in this thesis, i.e., "any language noted by learners to reflect on their language use, *with or without metalinguistic terminology*" demonstrates, the current thesis does not require the participants to use metalinguistic terminology when they engage in WL. For example, referring to an indefinite article in the original text used in the treatment of this thesis (I could find *a* good job), one participant wrote *hitotsu dakara "a"* "one (job), so 'a'" without using the term "indefinite article," which was coded as one WLE. However, given that WL involves learners' verbalisation of their thoughts regarding linguistic issues in writing, familiarity with metalinguistic terminology is deemed to be an influential factor in the extent to which they can successfully engage in WL. That is, metalanguage knowledge is expected to influence the quality of WL, further influencing L2 learning.

According to R. Ellis (2004, 2009), explicit knowledge consists of analysed knowledge (i.e., conscious representations of linguistic structures, which involves awareness) and metalanguage knowledge. Referring to the difficulty of measuring explicit knowledge accurately, he contends that there should be a distinction between explicit knowledge as analysed (potentially aware) knowledge and as metalanguage (2004, p. 265). In addition, pointing to one of the characteristics of explicit knowledge, i.e., being verbalizable, he explains that verbalizing a rule need not entail the use of metalanguage (2009, p. 13). He further states, however, that although metalanguage knowledge is not essential for explicit knowledge, they seem to be closely related. In his view, an increase in the depth of explicit knowledge can occur hand in hand with the acquisition of more metalanguage, "if only because access to linguistic labels may help sharpen understanding of linguistic constructs" (2004, p. 240), suggesting the significance of metalanguage knowledge for explicit learning. In line with this thinking,

Elder (2009) states that metalanguage knowledge "is independent of grammatical knowledge per se … and indeed of any cognitive or analytical skills associated with such knowledge" (p. 115). She adds, however, that a command of technical or semitechnical terminology may help learners to verbalise the grammar rules of target languages, implying the importance of metalanguage knowledge.

Supporting the statements of Elder (2009) and R. Ellis (2004, 2009), empirical findings reporting a positive correlation between metalinguistic knowledge and metalanguage knowledge have been obtained (e.g., Fortune, 2005; Hu, 2011). Hu (2011), for instance, identified a link between metalinguistic knowledge and metalanguage knowledge. The participants were 76 upper-intermediate Chinese learners of English who had received formal English education (i.e., a great deal of exposure to metalinguistic information) for at least six years. Six target structures, which were complex and believed to be difficult for Chinese learners of English, were chosen and 49 sentences, which included one of these structures, were prepared for a rule verbalization task. Upon reading each sentence, the participants were instructed to explain the uses of underlined target structures, i.e., "why the underlined structures were used" (p. 66). The analysis of the participants' rule verbalizations demonstrated a strong positive correlation between their metalinguistic knowledge and metalanguage knowledge. To be more precise, the participants who were more successful on the verbalization task, i.e., who possessed higher metalinguistic knowledge, turned out to have more metalanguage knowledge. Similar results were observed with learners who received explicit grammar instruction in Elder and Manwaring's (2004) study.

Along the same lines, in the aforementioned study by Ammar and Hassan (2017), the researchers point to the abundance of metalanguage as a possible factor contributing to the positive impact of collaborative dialogue (i.e., OL) on L2 learning. Referring to

previous studies in French that employed zero-error dictation (Nadeau & Fisher, 2014; Wilkinson, 2009, as cited in Ammar & Hassan, 2017), the researchers explain that collaborative dialogues during this type of dictation are "rich with terminological metalanguage" (p. 23) and that the amount of metalanguage that learners received correlates with their progress, i.e., the more metalanguage learners receive, the more progress they make compared to those who receive less. Referring to these findings from previous studies on zero-error dictation, Ammar and Hassan state that the use of metalanguage in their study "might have enhanced the depth of learners' ability to engage in collaborative dialogue, bringing about more metalanguage and eventually more metalinguistic knowledge depth" (p. 23). Their interpretation is compatible with the perspective of R. Ellis (2004), who explains that "metalinguistic knowledge [i.e., of terminology] may assist learners in developing explicit knowledge that has greater precision and accuracy" (p. 261).

In a similar vein, discussing metalanguage knowledge with reference to explicit knowledge, Roehr-Brackin (2015) claims that one of the parameters for explicit learning difficulty is the technicality of metalanguage, i.e., "the relative familiarity and abstractness of metalanguage used in the metalinguistic description" (p. 126). In her view, lower technicality of metalanguage results in lower explicit learning difficulty, which may imply that one has to have at least some level of technical metalanguage knowledge in order to benefit from explicit learning. If this speculation is correct, as WL is expected to facilitate explicit learning, it may be hypothesised that learners with a higher level of metalanguage knowledge benefit from WL more than those with a lower level of metalanguage knowledge. Therefore, another aim of the present thesis was to investigate the relationship between metalanguage knowledge and L2 learning from WL.

3.3 Research Questions and Hypotheses

In view of the previous research discussed above, the present thesis had eight specific goals. First, it intended to contribute to previous research examining the impact of WL on L2 learning. The second goal was to investigate an association between the frequency of T-WLEs (i.e., target construction-related WLEs) and development in the knowledge of the L2 construction. The third goal was to explore an association between the quality of T-WLEs (i.e., the level of awareness demonstrated in the T-WLEs) and L2 learning. The fourth aim was to examine the relationship between the frequency of T-WLEs and quality of T-WLEs.

The fifth aim was to identify aptitude profiles that are likely to interact positively with WL, in order to facilitate the possibility of providing instruction to learners that is matched to their aptitude profiles. Sixth, the present thesis investigated the relationship between learners' metalanguage knowledge and L2 learning from WL. The seventh aim was to examine the frequency of T-WLEs in relation to learners' aptitude and metalanguage knowledge, whereas the investigation of the association between the quality of T-WLEs and learners' individual differences in aptitude and metalanguage knowledge was the eighth aim.

Below are the research questions and corresponding hypotheses addressed in this research endeavour.

Research Question 1: To what extent does WL facilitate L2 learning?

Hypothesis 1: Based on the findings of previous oral languaging and self-explaining studies (e.g., Chi et al, 1994; Swain et al., 2009), it was hypothesized that WL would facilitate L2 learning as it was likely to enhance noticing and metalinguistic reflection of output, resulting in greater learning.

Research Question 2: To what extent does the frequency of T-WLEs relate to development in knowledge of the target construction?

Hypothesis 2: In view of previous research on oral languaging and self-explaining (e.g., Chi et al., 1989; Swain et al., 2009), it was hypothesized that a higher number of T-WLEs would relate to more development in knowledge of the target construction.

Research Question 3: To what extent does the quality of T-WLEs relate to development in knowledge of the target construction?

Hypothesis 3: Similar to Hypothesis 2, in light of previous research on languaging and self-explaining (e.g., Chi et al., 1989; Swain et al., 2009), it was hypothesized that the quality of T-WLEs would be a better predictor of learning than the frequency of T-WLEs (Knouzi et al., 2010; Qi & Lapkin, 2001).

Research Question 4: To what extent are the frequency of T-WLEs and the quality of T-WLEs related?

Hypothesis 4: Again, given the previous research findings on languaging and self-explaining (e.g., Chi et al., 1989; Swain et al., 2009), the frequency and the quality of T-WLEs were hypothesised to correlate closely (Chi et al., 1994; Knouzi et al., 2010; Qi & Lapkin, 2001).

Research Question 5: To what extent does language aptitude moderate the effect of WL on L2 learning?

Hypothesis 5: Considering that WL is expected to involve cognitive processes, such as inferencing, mental model repair, and knowledge integration, language aptitude was expected to moderate the effect of WL on L2 learning. Language aptitude in the present thesis was operationalised as inductive language learning ability and grammatical

sensitivity.

Research Question 6: To what extent does metalanguage knowledge moderate the effect of WL on L2 learning?

Hypothesis 6: Again, considering that WL requires learners to verbalise their thoughts about linguistic issues in writing, familiarity with metalinguistic terminology was hypothesised to enhance the effect of WL on L2 learning.

Research Question 7: To what extent is the frequency of T-WLEs related to individual differences in language aptitude and metalanguage knowledge?

Hypothesis 7: Given that producing T-WLEs is expected to involve noticing the target construction, the frequency of T-WLEs was hypothesised to be related to aptitudes, especially grammatical sensitivity. In addition, as stated, considering that WL is the action of writing about linguistic issues, metalanguage knowledge was hypothesised to be related to the frequency of T-WLEs.

Research Question 8: To what extent is the quality of T-WLEs related to individual differences in language aptitude and metalanguage knowledge?

Hypothesis 8: Again, given that producing higher quality of T-WLEs is considered to involve not just noticing but understanding the target construction, it was hypothesised to correlate with both aptitude constructs (i.e., grammatical sensitivity and inductive language learning ability). Also, as stated above, WL requires learners to write about linguistic issues, so metalanguage knowledge was hypothesised to be related to the quality of T-WLEs as well.

L2 learning was operationalised as acquiring the ability to process as well as produce the target construction more accurately. The next chapter describes in detail the procedures and results of a pilot study (M. Ishikawa, 2018), which contributed to further refinement and adjustment of the research design and methodology employed in the main study of this thesis.

CHAPTER IV

PILOT STUDY

4.1 Introduction

A pilot study for the present thesis was conducted with the aim to carry out a preliminary investigation regarding the effect of WL on L2 learning. In addition, it aimed to identify if learners' proficiency levels would moderate the effect of WL on L2 learning. To this end, the impact of two major independent variables, i.e., an opportunity for WL and learners' proficiency levels, on L2 learning was examined. For expository purposes, the opportunity for WL in the treatment will henceforth be denoted as +WL, while the absence of such an opportunity will be denoted as –WL. The following research questions were addressed.

Research Question 1: To what extent does WL facilitate L2 learning?

Research Question 2: To what extent do learners' proficiency levels moderate the effect of WL on L2 learning?

4.2 Participants

The initial sample consisted of 108 Japanese EFL learners from middle-class socioeconomic backgrounds attending two private universities in Japan. Participants who were placed in the higher-level groups belonged to one of the two required reading courses for first years at one of the universities. They were psychology majors and assigned to those classes based on their Test of English for International Communication (TOEIC) scores taken on entrance to the school about six months before the experiment. Their scores ranged from 480 to 655 (M = 568.3, SD = 62.7). Meanwhile, participants who were assigned to lower-level groups were enrolled in one

of two compulsory first-year English classes that focused on the TOEIC at the other university where the main study was conducted. They were economics majors and were placed in these classes based on their scores on an in-house test administered by the school. Their TOEIC scores ranged from 220 to 395 (M = 311.5, SD = 58.4). (Given the possible inherent differences of these two institutions, this operationalization of L2 proficiency was problematic. It was not feasible, however, to recruit enough participants with different proficiency levels at one institution.)

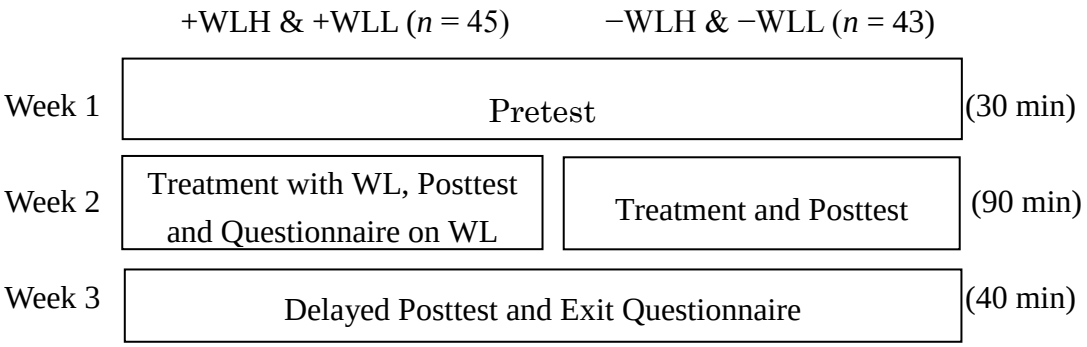
The participants were assigned to four groups according to the experimental treatment they received (+WL or –WL) and their level of proficiency (higher or lower level). Thus, they were assigned to one of the following groups: +WLH or –WLH group for higher-proficiency participants, and +WLL or–WLL group for lower-proficiency participants. Of the 108 participants, 25 were eliminated as they missed one of the posttests and/or scored higher than the threshold of 90%. Thus, the final number of participants was 83 (+WLH (18), –WLH (17), +WLL (25), –WLL (23)). Of these, 41 were male and 42 female, and they were almost equally distributed across the groups. Their ages ranged from 18 to 20 years (M = 19.2, SD = .5). I was the instructor for all these classes.

4.3 Experimental Design and Procedure

As presented in Figure 4.1, below, the experiment was conducted over a period of three weeks using parts of regular 90-minute classes (Weeks 1 and 3) or whole classes (90 minutes each) (Week 2) in the participants' respective classrooms. One week before the experiment, an information sheet about the experiment was distributed. All the participants agreed to participate and submitted signed consent forms (see Appendix A for informed consent documentation). Then, a background questionnaire was

administered (see Appendix B). (This questionnaire was also employed in the main study of this thesis.) In Week 1, a pretest was administered. The participants worked on a set of grammar tests, which consisted of a grammar production test and a recognition test, for 30 minutes. The time decided upon was based on a pilot test conducted before the experiment. All tests were collected at the same time, whether or not the participants had finished or not. This procedure was also adopted for two posttests. In Week 2, the treatment was administered to all groups. A posttest followed. Only the participants in the +WL groups worked on a questionnaire regarding WL. (The –WL participants were instructed to read their course textbooks during this time.) In Week 3, a delayed posttest was given, followed by an exit questionnaire for all participants.

Figure 4.1
Flow of the Procedure for All Groups



4.4 Linguistic Target

The target structure employed in this thesis was the present counterfactual conditional (e.g., “If I had more time, I would travel around the world”). (This target was also used in the main study of this thesis.) This is considered to be a demanding construction for L2 learners for at least two reasons. First, as it consists of two clauses (i.e., a subordinate clause and a main clause), it is syntactically complex. Second, there

are various types of conditional clauses, such as simple conditional and counterfactual conditional, which makes it difficult for learners to discern subtle distinctions in meaning among the different types (Celce-Murcia & Larsen-Freeman, 1999). Concerning the semantic difficulty of conditionals, based on Mental Spaces Theory, Dancygier and Sweetser (2005) explain that an if-clause sets up a mental space, "requesting construal of something … within that space" (p. 18) with a main clause. They further explain that "Much of the diversity of interpretation can be attributed to the fact that the spaces themselves can be quite diverse sorts of entities, related to the linguistic form in a variety of ways" (p. 18).

In the case of the present counterfactual conditional, the major semantic difficulty seems to be attributed to the fact that a mental space set up by an if-clause is described using the past tense, despite the fact that the space refers to the present time. For instance, in the example sentence above, "If I *had* more time, I *would* travel around the world" the word "had" in the if-clause and "would" in the main clause encode the feature (+past). It does not, however, have the semantic function of past time reference in either clause as the sentence refers to the present time. Instead, what the feature (+past) encodes is counter-factuality (see Izumi & Bigelow, 2000, for the distinction between present and past counterfactual conditionals).

In addition, this construction is problematic for Japanese learners of English, possibly because of L1 influence. Unlike English, Japanese is an SOV language and verb suffixes generally function as tense markers. More importantly, counter-factuality is not always subject to particular grammatical treatment, whereas English involves back-shifted tenses (Arita, 2014). As such, Japanese does not always require the past tense for both subordinate and main clauses of a present counterfactual conditional. For

example, the Japanese translation of the if-clause in the above example sentence "If I had more time" is:

Moshi	*watashi ni*	*motto jikan ga*	*ar-eba*(<*a-ru+eba*)/*at-tara*(<*at-ta+tara*)
If	I-NOM	more time-ACC	have-PRES-COND/have-PAST-COND

* NOM: nominative, ACC: accusative, PRES: present, PAST: past, COND: conditional.

As shown above, both the present and past tenses are acceptable in this situation in Japanese (Mizutani, 1989). The morphemes *–ru* and *–ta* in verbs "*a-ru*" (have) and "*at-ta*" (had) are present and past morphemes, respectively. Furthermore, the verbs "*a-ru*" and "*at-ta*" are combined with "*-eba*" and "*-tara*," conditional present and past morphemes, respectively, resulting in "*ar-eba*" (present) and "*at-tara*" (past). Because of the difference between the two languages, Japanese learners are likely to find the present counterfactual conditional difficult both syntactically and semantically, often failing to use the past tense for this construction and/or to understand its reference to the present time.

Thus, probably due to its syntactic and semantic complexity, instances of inaccurate understanding of the construction were observed in my classes. That said, the participants had completed at least six years of English education starting from junior high school in EFL classrooms whose primary focus is on grammar. As such, this construction was supposedly familiar to them. In other words, they were expected to have some explicit knowledge of the target construction. Previous research suggests that form-focused instruction may facilitate L2 learning when learners have partial knowledge of the form (e.g., Izumi & Bigelow, 2000). For these reasons, the construction was deemed to be appropriate for the participants despite its potential difficulty.

4.5 Assessment Tasks and Scoring

The learning of the target construction, i.e., changes in participants' knowledge and use of the present counterfactual due to WL, was measured in terms of participants' scores on the following two grammar assessment tests included in each of the pre- and posttests. (These tests were employed in the main study as well.)

1. A fill-in-the-blank production test;
2. A multiple-choice recognition test.

Three versions of the tests, i.e., Versions A, B and C, were devised for the pre-, post- and delayed posttests. They were different, yet had the same grammatical structure, and were of similar length, differing only in terms of vocabulary. It was confirmed that the vocabulary used in all the tests and the original text of the treatment is on the list of the most frequent 1,000 words in English of the British National Corpus (https://quizlet.com/8935711/british-national-corpus-top-1000-flash-cards/) and/or on the list of vocabulary items that Japanese high school students are expected to learn (Aizawa, S. Ishikawa, & Murata, 2015) in order to ensure that the participants could focus on grammar without being confused due to the difficulty of vocabulary items. Some vocabulary items that are not on these lists and which were not likely to be familiar to the participants, as well as the ones that are on these lists but which were not likely to be familiar to them, were pre-taught, so that they would have enough time to consolidate their vocabulary knowledge. In addition, in order for them to address the grammar production tests without being influenced by their lack of knowledge of past forms, the conjugation of verbs was practised. (The percentage of vocabulary items pre-taught was 1.98%. It was calculated by dividing the number of the items on the sheet by the total number of words in the three versions of the tests and the original text of the treatment.) These vocabulary items were included on a vocabulary sheet distributed one

week before the experiment and the items on the list were taught in part of the previous class prior to the experiment (see Appendix C).

In order to eliminate any possible task effects and teacher-class effects, the three versions of the tests were administered in a split-block design with a counterbalanced order. That is, each class was divided into three groups and one third of each class received one version of their test and the other two-thirds received the other two versions of their pretest. The versions were rotated for the posttest and delayed posttest in the same manner (i.e., ABC, BCA, CAB).

4.5.1 Grammar Production Tests

Grammar production tests were designed to assess the participants' ability to use the target construction, i.e., knowledge of the form-meaning mapping associated with the present counterfactual conditional in a controlled written environment. Each test consisted of 24 fill-in-the-blank items: eight on the target construction (i.e., present counterfactual), six if-distractors on a non-counterfactual (simple conditional) and 10 non-if (non-conditional) distractors, including one each for past, present perfect, comparative and present participle, and two each for passive, reported speech and indirect question (see Appendix D-1 for a sample version). All the verbs used in the eight target items were highly frequent irregular verbs with different past and past participle forms (e.g., *took* and *taken*) in order to make it possible to determine whether participants were using past or past participle forms of the verbs. Each verb was used only once in each test and appeared systematically across all versions. Each item was in the form of a dialogue with an average length of 30 words for both target and distractor items (Version A: 29.1 words, Version B: 29.7 words, Version C: 29.7 words). Instructions were given to fill a blank in each dialogue with a verb supplied next to the

blank in parentheses, changing it to an appropriate form. Below is an example item:

A: Are you walking to the station again? If you _____(take) a bus, it would be much faster.
B: I know, but I love walking.

The initial drafts of all three versions were piloted by two native speakers of English (an American male who teaches English at the same university where the main study was conducted and a British female) to confirm all the items would elicit the expected responses. After reviewing specific items where their answers did not match what was intended and considering their comments regarding the items they found confusing (e.g., multiple answers were possible), the problematic items were revised. Then, the revised three versions were piloted for a second round by the two native speakers, after which the three versions were further revised. When the second revision was complete, a Japanese female who holds a degree in TESOL completed the three revised versions of the tests and I found a perfect match between all her answers and the intended ones. Therefore, in order to ensure that the three versions of the tests were equivalent in terms of difficulty, all versions were administered to a group of 23 learners, who were different from the pilot study participants but comparable in terms of proficiency level and socio-economic background. A series of repeated measures analysis of variance (ANOVA)s were conducted using their scores on the three versions, and the results showed no statistically significant differences across the test versions, $F(2, 44) = .425$, $p = .656$, partial $\eta^2 = .019$. Also, the internal consistency reliability coefficients for the target items in the three versions of the tests were found to be acceptable (Cronbach's alpha: 882 .775, .822). The average scores of the three versions are presented in Table 4.1, below.

Table 4.1
Descriptive Statistics for the Pilot Grammar Production Tests

version	*M (SD)*	95%CI
A	6.35 (5.18)	[4.11, 8.60]
B	6.00 (4.41)	[4.10, 7.91]
C	6.52 (5.62)	[4.10, 8.95]

Note. $N = 23$.

The coding of responses involved assessing the verb forms produced in the if-clauses. Each item was examined to ascertain if the form was marked for the past tense. Participants received two points for each correct past-tense marking, one point for an incorrect but apparent attempt at past marking (e.g., *taked, *tooke) and no points for no attempt or an incorrect answer. First, all the points were added together to give an accurate use score (maximum 16, 2 pts x 8 items). Then, an overall score was calculated according to the use of the six if-distractor items (Pica, 1983). Each time participants used the past form for a present non-counterfactual item, two points were subtracted. These were tallied (maximum 12, 2 pts x 6 items) and subtracted from the accurate use score, thus yielding an overall score (maximum 16, 2 pts x 8 items). Therefore, two scores, i.e., accurate use and overall score, were arrived at for each test. For example, when participants answered all the target items correctly, their accurate use score was 16 (2 pts x 8 items). If they used the past tense for all the if-distractors (examples of overuse), however, 12 points (2 pts x 6 items) were subtracted from their accurate use score and the overall score they received was four. (The average scores presented in Table 4.1 above are based on accurate use scores.)

4.5.2 Recognition Tests

The purpose of the recognition tests was to assess the participants' knowledge of the meaning associated with the present counterfactual conditional. Each test also

consisted of 24 multiple-choice items, of which eight included the target construction (see Appendix D-2 for a sample version). The distractors comprised four if-distractors for the simple conditional, and 12 non-if distractors on the following forms: one item each for present, future, conjunction and relative, two items for comparative, and three items for negative and past.

Each target item consisted of a sentence on the target construction, followed by four options in English. The participants were instructed to choose the one that had the same meaning as the target sentence by circling the letter next to it. Of the four options, one was a correct description of the situation and the rest were incorrect to varying degrees. These descriptions reflected four semantic feature combinations: (1) +present +unreal, (2) +present–unreal, (3) –present +unreal, and (4) –present –unreal. When scoring the target items, two points were awarded for each correct response (+present +unreal), one point for a response with one of the semantic features correct and the other incorrect (+present–unreal, –present +unreal), and zero points if neither feature was correct (–present –unreal), thereby yielding a maximum possible score of 16 (see Révész, Sachs, & Hama, 2014, for a similar procedure). Only items including the target construction were used to assess participants' learning; the rest served as distractors and were not assessed. Below is a sample item.

If I had an international driver's licence, I could drive abroad.
a. I have an international driver's licence.
b. I don't have an international driver's licence.
c. I had an international driver's licence.
d. I didn't have an international driver's licence.

The three versions of the test were also piloted following the same steps as in the grammar production test. First drafts of all three versions were piloted by the same two

native speakers of English (an American male and a British female) to confirm all the items would elicit the expected responses. Their comments were reviewed and items which were found to be problematic and/or ones where they answered differently from what was intended were revised. Then, the three revised versions were piloted for a second round and further minor revisions were made. After the second revision was finished, the same Japanese female who cooperated with me on the grammar production tests completed the three revised versions of the tests. Again, all her answers matched the intended ones. So, the 23 learners who participated in the pilot of the grammar production test also completed these tests in order to ensure the equivalency of the three versions of the tests. Table 4.2, below, presents the means of the three versions of the tests. A series of repeated measures ANOVAs were conducted, and the results showed no statistically significant differences across the tests, $F(2, 44) = 1.191$, $p = .314$, partial $\eta^2 = .051$. The reliability coefficients for the three versions of the tests were .806, .755, .786, suggesting they all had acceptable internal consistency.

Based on the pilot tests as well as a discussion with the teachers who cooperated on the piloting, it was decided to set 30 minutes as a time limit for the two grammar tests (i.e., production test and recognition test). No separate time limit was set for each test. In addition, a cut-off line of 90% was set for the both production and recognition tests. Those who scored over the cut-off line on either or both of the tests on the pretest were removed from the data.

Table 4.2
Descriptive Statistics for the Pilot Recognition Tests

version	*M (SD)*	95%CI
A	9.35 (3.04)	[8.03, 10.66]
B	10.13 (3.07)	[8.81, 11.46]
C	9.39 (2.59)	[8.27, 10.51]

Note. $N = 23$.

4.6 Treatment Task

The treatment task employed in the present thesis was dictogloss, developed by Wajnryb (1990), which is one of the most commonly used language-focused tasks in research on writing (Storch, 2013). There are variations but dictogloss usually requires learners to reconstruct a short passage after they listen to it and then analyse and correct reconstructed texts in comparison with an original text. Through these procedures, learners are expected to "refine their understanding of the language they have used" (Wajnryb, 1990, p. 5). Thus, one of the major features of dictogloss is to enable learners to direct their attention to form while also focusing on meaning in order to reconstruct the original text based on their understanding of it. When learners reconstruct the passage, they usually work in small groups or pairs, which is another major feature of dictogloss. According to Wajnryb, "learner involvement and interaction" (p. 16) is the essence of the reconstruction and comparison stages.

Meanwhile, dictogloss employed in the present thesis was slightly different from that explained above. That is, instead of analysing and correcting their reconstructions by talking with their peers, the participants were instructed to engage in that process individually, i.e., individual written dictogloss. The change to the standard dictogloss procedure was made in order to elicit WL. During the comparison of their reconstructions with the original text, the +WL participants were instructed to engage in written interaction with themselves individually (i.e., WL) and the –WL participants to do the same silently without writing. (This task was also used in the main study.)

The rationale for choosing dictogloss was as follows. First, as dictogloss requires learners to reconstruct a text they listen to, it enabled me to provide a context where the target construction was used. As stated, L2 learning in the present thesis was defined as acquiring the ability to process as well as produce the present counterfactual

construction more accurately. Therefore, ensuring that obligatory contexts were created for its use was important. Furthermore, having control over the content was significant in order to facilitate learners' cognitive comparisons between their output and input. Second, as important as the first one, if not more so, dictogloss enabled me to elicit WL. Previous studies (e.g., Kowal & Swain, 1994, 1997) that employed dictogloss as a treatment task reported that it was effective in eliciting languaging. Referring to her previous studies with Kowal (i.e., Kowal & Swain, 1994, 1997) on metatalk (i.e., OL), Swain (1998) states that dictogloss provided them with the sort of data they hoped the task would elicit, i.e., "talk about the language of the text they [= learners] were reconstructing (metatalk)" (p. 70), further stating "not just any task will elicit metatalk" (p. 79). For these reasons, dictogloss was deemed more appropriate than other tasks, such as essay-writing, which allow freedom in terms of content and form, making it difficult for learners to engage in cognitive comparisons between input and output (Izumi, 2000) and for me to elicit WL.

Before listening, a handout giving instructions in Japanese and spaces with bullet points for participants to take notes was distributed in order to facilitate the participants' note-taking (see Shintani et al., 2014, for a similar procedure) (Appendix E). In order to avoid the participants attempting to write all the words in the original text, they were told not to write whole sentences, but only to take notes of those words that they found important, necessary and/or useful to reconstruct the texts afterwards. Pauses between sentences were also included in such a way that writing the text verbatim was not feasible. The passage contained 14 sentences made up of 116 words, but the first sentence (14 words) was printed on the sheet to make the task less demanding. Thus, the actual length of reconstruction was 13 sentences (102 words), including three sentences on the target form (32 words, 31.3% of total words).

In order to conduct the experiment efficiently, practice sessions for WL and dictogoloss were run beforehand over a period of five weeks. First, as "familiarity with task procedures is important" (Swain, 1998, p. 80), two training sessions were conducted in the first two weeks in order to familiarize the participants with the activity of WL (McNamara, 2017; Swain, 1998). The participants practiced WL in their L1, Japanese. Then, two practice sessions on dictogloss were given over the next two weeks. In the final week, a practice session that combined WL and dictogloss was conducted. Shorter passages (around 60 words) were used for dictogloss exercises and all the participants practised WL when comparing their reconstructions with the original text.

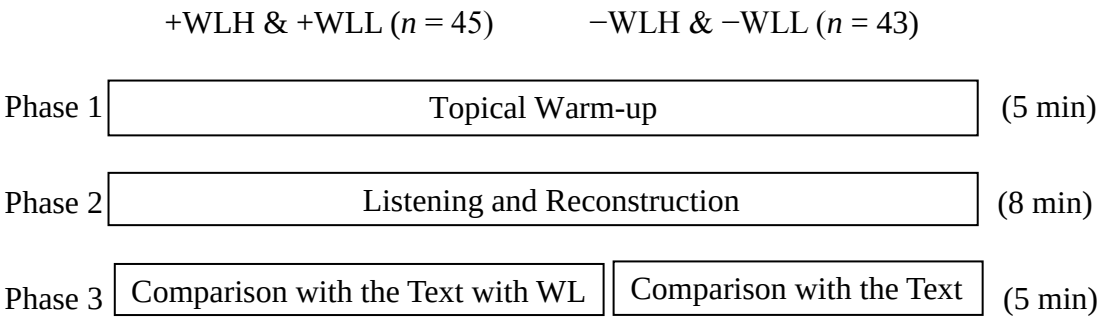
4.7 Treatment and Its Procedure

A three-phase treatment was administered to both the +WL and –WL groups. Figure 4.2 presents an overview of this sequence. First, as a topical warm-up, the participants were shown PowerPoint slides which contained pie charts on the future expectations of graduates from the previous year extracted from their respective universities' websites, and a brief discussion in Japanese ensued as to the future path they would like to pursue (Appendix F). Five minutes were allotted for this phase. Second, a task sheet was distributed (Appendix E). The participants listened twice to the text, recorded by a male native speaker of English, took notes during the second listening and for five minutes reconstructed the text from their notes. A total of eight minutes was allocated for this phase.

Finally, a sheet consisting of instructions in Japanese with the original text below it was issued to all participants. The instructions for the +WL group and –WL group were slightly different. The one for the +WL group asked the participants to examine

the original text carefully and engage in WL on a WL sheet, i.e., write down their thoughts upon examining the original text on the sheet (Appendix G-1), whereas the instructions for the –WL group asked the participants to compare their reconstruction and the original text carefully without any writing (Appendix G-2). It also informed the –WL groups of a subsequent activity (i.e., write an essay), so that –WL participants would be motivated to make a comparison, focusing on constructions that might be useful for their next task.

Figure 4.2
Overview of the Sequence of Treatments



The rationale of not giving extra tasks to the –WL groups was to isolate the effect of WL from other tasks, such as grammar exercises (M. Ishikawa & W. Suzuki, 2016). Also, according to Cumming (1990), learners use metalinguistic thinking as they compose in L2, which seems to be comparable to languaging in learners' minds. Therefore, it sought to separate the effect of WL from possible silent languaging. The participants were instructed to engage in WL in their L1, i.e., Japanese, so that they would be able to do so without being affected by their English proficiency or cognitive complexity. The instructions regarding the use of L1 were given orally and not included in the WL sheet (Appendix G-1). As reported in Sachs and Polio's (2007) study, when learners were instructed to engage in think-alouds in their L2, negative reactivity was

observed. Therefore, although the target language is English, engaging in WL was deemed to be best accomplished in the participants' L1. It was also a decision made by Swain et al. (2009) in their aforementioned study on OL. Explaining the use of the participants' L1 in a Vygotskyan framework, the researchers state that using L1 to mediate the understanding of a concept that is then applied to their understanding of how L2 works is not unreasonable, because language is considered to be a tool to mediate cognitive activity (see also Alegría de la Colina & García Mayo, 2009; Storch & Wigglesworth, 2003, for a similar perspective). Five minutes were allotted for this phase and all sheets were collected at the end of this time. The timings assigned to each of the phases were based on pilot tests conducted beforehand at the university, where the participants of lower proficiency belonged as well as the main study was conducted. Immediately after Phase 3, all participants did a posttest.

4.8 Statistical Analyses

First, descriptive statistics for the tests were calculated. In order to further investigate the research questions, inferential statistical analyses were conducted using the statistical package IBM SPSS 23.0. The significance threshold was set at an alpha level of .05. As the skewness and kurtosis ratios were outside the [– 2, 2] interval for several of the dependent variables, it was decided to run a series of non-parametric tests. In order to answer the research questions, for each pair at both proficiency levels, Mann-Whitney tests were conducted with the +/– WL condition as the between-subjects variable to examine the impact of WL on L2 grammar learning for both production and recognition tests. The dependent variables in the analyses were the pretest-posttest and pretest-delayed posttest gain scores. As for RQ1, the results were examined to identify any possible differences between the +WL and –WL groups at both proficiency levels.

As for RQ2, the same results were analysed to discern any possible differences depending on proficiency levels.

Although *d*-values assume a roughly normal distribution (Plonsky, personal communication) and may not be ideal for non-parametric tests whose assumption is non-normal distribution, there are no guideline values for non-parametric tests. As such, the standards in SLA research of small (*d* = .40), medium (*d* = .70) and large (*d* = 1.00) (Plonsky & Oswald, 2014) were employed when interpreting the effect sizes of the results.

4.9 Results

4.9.1 Grammar Production Tests

Table 4.3 presents descriptive statistics for the results of the grammar production tests across the four groups. Because of the skewed distribution, it was decided to employ medians and interquartile ranges as measures of the central tendency and variation, respectively. Both of the +WL groups improved their accuracy in use and overall scores over the three tests, especially the +WLL group, which tripled its accurate use and overall scores on the posttest compared to those for the pretest (from 4.00 to 12.00 and 2.00 to 6.00 for accurate use and overall, respectively). Although their delayed posttest accurate use score dropped to 10.00, their overall score increased further, to 8.00. Similarly, both the –WL groups improved their scores over the three tests, but the improvements were not as marked as those of the +WL groups. In terms of proficiency levels, there was a noticeable gap in the pretest scores between the higher- and lower-level groups, which narrowed over time.

Table 4.3
Grammar Production Test Scores for the Four Groups

		Pretest			Posttest			Delayed Posttest		
	N	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI
AU										
+WLH	18	12.00	6	[8.00,14.00]	13.50	5	[11.00,15.00]	15.00	4	[13.00,16.00]
–WLH	17	13.00	4	[10.00,14.00]	14.00	8	[8.00,16.00]	14.00	3	[12.00,15.00]
+WLL	25	4.00	6	[2.00, 6.74]	12.00	8	[7.13,13.74]	10.00	7	[8.00, 14.00]
–WLL	23	6.00	12	[4.00, 12.37]	8.00	12	[3.50,13.87]	10.00	10	[4.51, 13.87]
OA										
+WLH	18	8.50	9	[5.00,12.00]	11.00	7	[10.00,14.00]	13.00	7	[9.00,14.00]
–WLH	17	11.00	5	[9.00,13.00]	12.00	9	[6.00,15.00]	12.00	9	[9.50,14.00]
+WLL	25	2.00	6	[.00, 4.74]	6.00	7	[4.00, 9.49]	8.00	7	[5.00, 9.50]
–WLL	23	3.00	10	[2.00, 8.00]	3.00	7	[2.00, 8.49]	4.00	11	[3.63, 6.00]

Note. AU: accurate use, OA: overall, maximum score = 16.

Then, inferential statistics were calculated. First, Mann-Whitney tests were run to determine whether accurate use and overall scores on the pretest differed significantly between the +WL and –WL groups. Separate analyses were conducted for each proficiency level. The results demonstrated no significant differences for accurate use scores (+WLH group vs –WLH group, $z = –1.321$, $p = .133$, $d = .449$; +WLL group vs –WLL group, $z = –.734$, $p = .471$, $d = .212$) or for overall scores (+WLH group vs –WLH group, $z = –1.113$, $p = .248$, $d = .381$; +WLL group vs –WLL group, $z = – .881$, $p = .383$, $d = .252$), suggesting they were comparable at the outset.

Mann-Whitney tests were also run to evaluate whether pretest-posttest and pretest-delayed posttest gain scores differed significantly across the +WL and –WL groups on the two proficiency levels. For the higher-level pair, significant differences with a medium effect size were identified for the pretest-delayed posttest gain scores for accurate use ($z = –2.333$, $p = .020$, $d = .857$) and overall scores ($z = –2.630$, $p = .007$, $d = .981$), but not for the pretest-posttest gain scores (accurate use, $z = –1.132$, $p = .207$, $d = .391$, and overall, $z = –1.034$, $p = .302$, $d = .351$).

For the lower-level pair, significant differences were identified not only for the pretest-delayed posttest gain scores for accurate use (z = –2.221, p = .031, d = .683) and overall scores (z = –2.023, p = .044, d = .581) with small effect sizes, but also for the pretest-posttest gain scores for accurate use (z = –2.772, p = .013, d = .872) and overall scores (z = –2.551, p =.012, d = .793) with medium effect sizes. Taken together, the gain scores for the +WL groups on both levels were found to be significantly higher than for the –WL groups on some tests, suggesting the facilitative impact of WL on L2 grammar learning (RQ1). In addition, in terms of proficiency levels, the +WLL group was found to have benefited more from WL than the +WLH group (RQ2).

4.9.2 Recognition Tests

Table 4.4 summarises the descriptive statistics for the results of the recognition tests across the four groups. The scores of the higher-level groups were already high on the pretest (13.50 and 14.00 for the +WLH and –WLH groups, respectively), but the gap between the higher- and lower-level groups was narrower compared to that of the production test. Although all groups scored higher on the two posttests than they did on the pretest, the increases were small.

Table 4.4
Recognition Test Scores for the Four Groups

		Pretest			Posttest			Delayed Posttest		
	N	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI
+WLH	18	13.50	2	[12.00,14.00]	15.00	2	[14.00,16.00]	15.00	2	[14.00,16.00]
–WLH	17	14.00	4	[11.00,14.00]	15.00	4	[12.50,16.00]	15.00	3	[13.00,16.00]
+WLL	25	9.00	4	[8.00, 11.00]	11.00	6	[8.63, 13.00]	11.00	6	[9.00, 14.00]
–WLL	23	8.00	4	[6.00, 10.00]	9.00	7	[8.00, 12.74]	9.00	3	[7.63, 10.50]

Note. maximum score = 16.

A Mann-Whitney test on the pretest scores revealed no statistically significant differences in the performances of the +WL and –WL groups at either proficiency level

(+WLH group vs –WLH group, z = –.753, p = .461, d = .254; +WLL group vs –WLL group, z = –.732, p = .481, d = .212), indicating they were comparable at the beginning.

In order to answer the research questions, Mann-Whitney tests were conducted to compare the gain scores of the +WL and –WL groups at both proficiency levels. First, the gain scores of the higher-level groups were examined, which revealed no significant differences between either of the gain scores (pretest-posttest, z = –.804, p = .421, d = .274, pretest-delayed posttest, z = –.415, p = .678, d = .140). Similar results were obtained when examining the gain scores of the lower-level groups (pretest-posttest, z = –1.212, p = .231, d = .352; pretest-delayed posttest, z = –.214, p = .839, d = .062). Overall, in contrast to the results of the production tests, both the +WL and –WL groups at each proficiency level demonstrated similar gains over the two time periods, as reflected in the effect sizes, which were all in the range of small (d < .40).

4.10 Discussion

The first research question addressed the impact of WL on L2 grammar learning. The results of the grammar production tests produced evidence suggestive of the positive impact of WL on learning the target construction. Specifically, comparisons of the gain scores of the +WL and –WL groups at each proficiency level revealed statistically significant differences, with small to medium effect sizes. Of these, what is especially noteworthy is that both the +WLH and +WLL groups improved their scores significantly from the pre- to delayed posttests, for both accurate use and overall. That is, both groups benefited in the long term.

These results are likely to derive from the externalisation of thoughts in writing (i.e., WL) which the +WL participants produced, benefiting them through the process of their WL. Namely, the use of WL when comparing the original text and their

reconstructions is likely to have triggered the +WL participants to engage in deeper processing (Craik & Lockhart, 1972), resulting in longer-lasting and stronger memory representations compared to the –WL participants. Perhaps due to this deeper processing, the +WL groups seem to have retained the treatment effect better than the –WL groups, thus contributing to long-term significant gains.

In contrast to the findings for the production tests, no positive impact of WL was identified with respect to the recognition tests. One possible reason is a ceiling effect, at least for higher-level participants. In addition, the different findings for the two tests may be better explained by the Output Hypothesis (Swain, 1995), which claims that output provides opportunities for learners to move from semantic processing to syntactic processing. Although output and WL are not identical (especially, WL in this thesis was conducted in the participants' L1), given that both of them entail learners' language production, it can be argued that the hypothesis applies to WL to some extent. If this speculation is correct, the WL opportunity (i.e., to language about language) might have allowed the +WL participants to shift from semantic processing to the syntactic processing needed for accurate production (Swain, 2000), resulting in their having better performance than the –WL participants on the grammar production tests. In contrast, given that the recognition tests were designed to measure the participants' knowledge of meaning, a mere comparison with the original text and their reconstructions (i.e., to process input) might have been sufficient to carry out the recognition tests, contributing to no significant differences between the two groups.

The second research question addressed the relationship between learners' proficiency levels and the impact of WL on their L2 grammar learning. As mentioned, the threshold of 90% was too high and a ceiling effect was observed in the pretest for the higher-level groups on the recognition tests, making it impossible to compare the

+WLH and +WLL groups. Thus, the discussion here will focus on the results of the grammar production tests. As reported, the +WLL group benefited in both the short and long term (all of their gains were statistically significant, with small to medium effect sizes), whereas the +WLH group only benefited in the long term (their short-term gains were not statistically significant). Considering these results, the +WLL participants most likely benefited more from WL. This is consistent with the findings of Ammar and Hassan (2017) and McNamara (2017), who found that lower-knowledge learners benefited more from OL and oral self-explaining, respectively, and Muñoz et al. (2006), who noted that typed self-explaining (WL) helped lower-level learners. Two reasons might account for this outcome.

First, the target construction was probably not challenging enough for the +WLH participants, and even if it was, it was not so in its operationalisation in the grammar production tests. Therefore, they scored fairly high on the pretest, leaving less room for them to improve their scores as much as the +WLL participants. Namely, a ceiling effect might have contributed to the result, as was the case with the recognition tests. Second, related to the first point, although WL is expected to assist learners by enhancing their metalinguistic reflections (Swain, 2006), as well as by triggering deeper processing (Craik & Lockhart, 1972), the +WLH participants might not have needed such facilitation, at least in the short term, resulting in no significant differences in their pre-post gains, and more favourable results for the +WLL participants who benefited from such facilitation (McNamara, 2017).

It is important to note, however, that the +WLH group also improved their accurate use and overall scores significantly from the pre- to delayed posttests. What is more, although in the range of medium ($d < 1.00$), the effect size for the overall scores, a better indicator of improvement than accurate use scores, was the largest of all those

observed for both levels (d = .981), indicating a positive impact of WL in the long term. Considering these results, WL may benefit higher-level learners as well, depending on the task and target construction.

4.11 Limitations and Conclusion

Although the pilot study has produced some evidence concerning the positive impact of WL on L2 learning and the role of learners' proficiency level as a moderator on its impact, a number of methodological issues surfaced that need to be addressed in the main study. First, both the production and recognition tests were highly structured. The inclusion of more productive open-ended tests would have provided a more useful view of learners' real productive ability. Second, as the pilot study did not include a true control group, which did not receive the treatment, variables other than languaging, such as test repetition effects (Agarwal, Finley, Rose, & Roediger, 2017), might have contributed to the findings. Third, although WL produced by the +WL participants turned out to be a rich source of data, demonstrating the participants' understanding and perceptions of the task, it was not closely explored. Given that the amount and quality of languaging have been reported to influence learning (McNamara, 2017; Swain et al., 2009), the product of WL should have been analysed both quantitatively and qualitatively, and been further analysed in relation to test scores. Especially, the qualitative analyses of languaging, i.e., analyses of the content of WL, such as their focus/types and quality, would have provided more accurate and thorough information regarding the participants' learning that was not obtained by assessment tests.

An additional shortcoming was that the ceiling effect was likely to have contributed to the result. The threshold of 90% was too high, resulting in a ceiling effect for the recognition tests. Although the lower-proficiency participants demonstrated

significantly higher gain scores, the higher-proficiency participants scored quite high already on the pretests, leaving little room for them to improve their scores. Thus, a stricter cut-off line should have been set to exclude those participants who achieved high scores. Further methodological limitations of the study included the questionnaire on WL being administered immediately after the posttest, which might have raised the +WL participants' consciousness towards the target construction, possibly confounding the results. If it had been given after the delayed posttest, the potential confounding effect could have been avoided. In addition, although two scores, i.e., accurate use scores and overall scores, were employed to examine participants' learning, the inclusion of overuse scores would have revealed more accurate information regarding changes in the participants' learning that might have gone undetected.

Finally, although the pilot focused on learners' proficiency level as one of the major independent variables and as a possible factor influencing the impact of WL on learning, the questionnaire on WL produced interesting data. Most of the participants responded positively. There were, however, some negative remarks, such as "I don't like writing much," regardless of their proficiency level. As mentioned earlier, similar comments were reported by Simard et al. (2015), who employed written verbalisation data to examine learners' perceptions of corrective feedback (see also García Mayo & Loidi Labandibar, 2017, for similar comments). The fact that the remarks came from not only lower- but also higher-proficiency participants might be explained by individual differences among them. Thus, the examination of WL in relation to individual differences, such as aptitude, seems worthy of future investigation in order to identify the characteristics of learners who are likely to extract the maximum benefit from WL.

In the next chapter, the discussion turns to the design and methodological procedures of the main study, in which an attempt was made to circumvent these

shortcomings as well as to heed the methodological issues considered in the previous section. Below is a list of the changes made for the main study.

Group:
- the inclusion of a control group

Procedures, Analysis and Others:
- the inclusion of an additional combined practice session of WL and dictogloss
- the inclusion of an analysis of the contents of WLEs
- the inclusion of interviews and case studies
- the timing of the administration of an exit questionnaire (after a delayed posttest instead of an immediate posttest)
- the timing of the distribution of the vocabulary check sheet (two weeks, instead of one week, before the experiment)

Assessment tests:
- the inclusion of essay tests
- a lowered cut-off line of 80% for all the assessment tests
- the partially changed scoring of the grammar production tests
- the inclusion of tests on individual differences in aptitude and metalanguage knowledge

CHAPTER V

METHODS

This chapter describes in detail the procedures that were followed to investigate the effect of WL on learning the present counterfactual conditional, the target construction. First, an overview of the research design, details about the participants, the linguistic target, the measures used to assess L2 learning, aptitude and metalanguage knowledge and the treatment are presented, followed by the coding of WLEs. Finally, questionnaires, interviews, case studies and statistical analyses are explained.

5.1 Overview of the Design

The main study was conducted during a period of five weeks using the participants' 90-minute regular class times in a pretest-treatment-posttest-delayed posttest design. The participants were divided into two treatment groups and one control group, and they engaged in the tasks in their respective classrooms except for Week 2, when a computer room was used for the two treatment groups. (The control group used their own classroom throughout the experiment.)

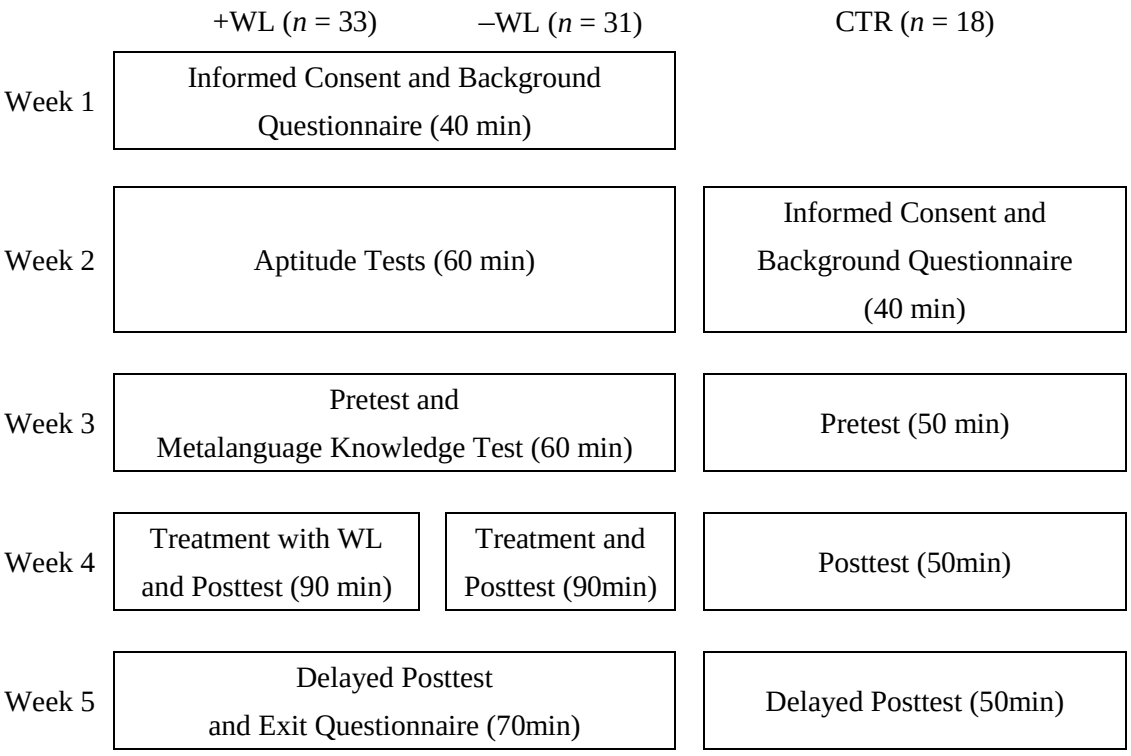
As summarised in Figure 5.1, the two treatment groups differed with respect to the opportunity to engage in WL during the treatment in Week 4, which was the major independent variable. The dependent variable was any change in the results from the pretest to the posttests, which was used to investigate the impact of WL on learning the target construction. In addition, two moderating variables, the results of aptitude tests and a metalanguage knowledge test, were included to identify whether these factors would moderate the effects of WL on L2 development. The control group only participated in the pre- and posttests and did not have the treatment, aptitude tests or

metalanguage knowledge test. The inclusion of this group was deemed important to tease out the effect of WL from potential test repetition effects. A brief overview of the experiment is given in the remainder of this section, and more details regarding each step and the measurement instrument are provided in the sections following.

In Week 1, an explanation regarding the experiment was given with an information sheet to the two treatment groups. All the students agreed to participate and submitted a signed informed consent document (see informed consent documentation in Appendix H). Then, as stated, the background questionnaire used in the pilot was administered (Appendix B). The procedure took about 40 minutes. The same procedure was followed in Week 2, the first week, for the control group. In Week 2, aptitude tests (i.e., the MLAT, LABJ and LLAMA) were administered to the two treatment groups. This phase was completed in about an hour.

In Week 3, the pretests were administered, followed by a metalanguage knowledge test. First, the participants worked on the pretest, consisting of an essay test and two grammar tests (a production test and a recognition test) for 20 minutes and 30 minutes, respectively, 50 minutes in total. Then, the metalanguage knowledge test was given, which took about 10 minutes. In Week 4, the treatment was administered, followed by a posttest, using a 90-minute class for the whole procedure. The treatment involved participants carrying out individual written dictogloss activities. In Week 5, the delayed posttest was administered and an exit questionnaire was given, which took about 70 minutes. As stated above, the control group took only the pre- and posttests during those three weeks.

Figure 5.1
Flow of the Procedure for All Groups



5.2 Participants

The participants were Japanese EFL learners at a private university in Japan where the pilot study was conducted with participants of lower proficiency. They were enrolled on one of four compulsory first-year English classes that prepare students for the TOEIC. Of the four classes, two were the second- and sixth-highest levels of the ten English classes in the pharmacy department and the other two were the second- and third-highest levels of the 13 English classes in the management department. They were placed in these classes based on their scores on an in-house test administered by the school. There is a gap concerning English proficiency between the two departments and the proficiency of pharmacy majors is generally higher than that of management majors. Their TOEIC scores ranged from 240 to 525 (M = 356.7, SD = 77.7). Of the four classes, three (second-highest-level pharmacy and the two management classes) were

assigned to the two treatment groups (either the +WL or –WL group), while the sixth-highest level pharmacy class was designated a control group in order to address possible treatment and test repetition effects. The class was chosen as a control group because the participants in this class were assumed to possess average English proficiency among the four classes judging from their TOEIC scores. Although it would have been ideal to assign all the classes to the three groups in stratified random sampling, this was not feasible because of administrative and practical constraints (see Table 5.1 for the average TOEIC scores of the three groups).

Based on the results of the pretest, aptitude tests, metalanguage knowledge test and TOEIC test, the participants in the three classes (second-highest-level pharmacy and the two management classes) were assigned to either the +WL or the –WL group through stratified random sampling so that the two groups would not differ in terms of their pretest scores or individual variables. The initial number of participants was 133, but 51 participants who missed one of the posttests and/or who scored above the cut-off line of 80% on the pretest were eliminated. Thus, the final number of participants was 82. Of these, 44 were male and 38 were female, and their ages ranged from 18 to 20 years (M =19.0, SD = .5). Table 5.1 summarises the background information for the final pool of participants across the two experimental groups and the control group.

The participants had received between six and a half to 12 years of English instruction prior to the study (M = 7.3, SD = 1.2). There were some participants who reported 10 years or more of learning experience, stating that they started to go to English conversation schools, whose focus is on communication, when they were elementary school students. Most of the participants, however, had around seven years of learning experience, including the six years of English education at junior high and senior high school (three years each), where English is typically taught using grammar-

translation methods by non-native speaker teachers, i.e., Japanese teachers of English, who often teach all the grammar in Japanese. In addition to English, the majority of the participants (*n* = 77) had studied, or were studying, an additional foreign language (i.e., French, German, Chinese, Korean or Hungarian) in courses at the university. Most of the participants had never visited an English-speaking country, with the exception of seven participants who had gone on a one-week school trip to Australia or New Zealand (four participants) and three who had travelled to the United States with their respective families for about a week. Again, I was the instructor of all the classes. A one-way ANOVA was conducted to detect any initial differences across the groups in terms of age, mean length of study or proficiency (measured by TOEIC scores). No significant differences emerged for any of the variables ($F(2, 79) = .982$, $p = .379$ for age, $F(2, 79) = 1.049$, $p = .355$ for mean length of study, and $F(2, 79) = .737$, $p = .482$ for proficiency).

Table 5.1
Background Information for Each Group

Group	*N*	Age *M (SD)*	Gender Male/Female	Mean length of Previous English Study in Years *M (SD)*	TOEIC score *M* (*SD*)
+WL	33	18.9 (.5)	19/14	7.2 (1.3)	353.2 (76.9)
–WL	31	19.1 (.5)	18/13	7.6 (1.2)	359.8 (72.5)
CTR	18	19.1 (.6)	7/11	7.5 (1.2)	335.3 (40.0)

5.3 Linguistic Target

The linguistic target was the same as in the pilot study (see section 4.4).

5.4 Design of the Tests and Scoring

5.4.1 Assessment Tasks and Scoring

As stated, the two grammar tests devised for the pilot study were also employed

in the main study. In addition, in order to address one of the limitations of the pilot study, an essay test was included. Therefore, the following three tests were involved in each of the pre- and posttests in order to assess changes in the participants' declarative and procedural knowledge of the target construction, the present counterfactual conditional, due to the treatment.

1. An essay test (extended production test);
2. A grammar fill-in-the-blank production test (limited production test);
3. A grammar multiple-choice recognition test.

The rationale behind the decision to include these tests was to measure the treatment effects, if any, on different aspects of L2 acquisition, considering the importance of using multiple modes of testing to assess grammatical knowledge (Norris & Ortega, 2003). As indicated in the previous text, L2 learning in this thesis is conceptualised in terms of the distinction between declarative and procedural knowledge of the target construction. In DeKeyser's (2007) view, declarative knowledge that can be obtained via explicit learning processes can be converted into procedural knowledge through practice. Therefore, based on this distinction, an essay test was included to assess participants' knowledge of use of the target construction in written production in a relatively natural context where learners had freedom to choose the content. As stated, the grammar production test was designed to assess the participants' knowledge of the meaning-to-form mapping in a highly-controlled written context, while the recognition test was designed to assess their knowledge of form-to-meaning mapping. Participants could rely on their declarative knowledge in all the tests, given that they were written and there was little time pressure, but the essay test was likely to elicit the most extensive use of procedural knowledge, followed by the grammar production and recognition tests. An essay test was administered first so that

more controlled grammar tests would be less likely to influence the participants' performance on the essay test.

Three versions of the essay tests, i.e., Versions A, B and C, were devised for the pre-, post- and delayed posttests. Each version was combined with the three versions of the two grammar tests (production test and recognition test) and they were administered in a split-block design with a counterbalanced order, following the procedures used in the pilot study. Also following the same procedure, all tests were collected at the same time, whether the participants had finished or not. As a detailed description of the two grammar tests has already been provided above, only the essay tests are explained in detail below, followed by an explanation of three changes made to the administrative procedure for the two grammar tests.

5.4.1.1 Essay Tests

As stated above, the goal of the essay tests was to measure the participants' knowledge of the target construction in a relatively uncontrolled context. Three writing prompts were prepared referring to the Test of English as a Foreign Language (TOEFL)-related websites such as the ETS (https://www.ets.org/s/toefl/pdf/qp_v1_web_a4.pdf) and "TOEFL resources" (https://www.toeflresources.com/index.php?id=sample-toefl-essays). The rationale behind this decision was that former SLA studies whose participants, also Japanese university students, seemed to have no difficulty writing essays based on these prompts (e.g., M. Suzuki, 2008; W. Suzuki, 2009; 2012). It should be pointed out, however, that the current study included participants whose proficiency levels were lower than those in previous studies.

Twenty minutes were allotted because this was deemed sufficient to write meaningful essays based on the pilot tests conducted beforehand. Below is one of the

three prompts. Each prompt was written in Japanese, in order to avoid any copying of sentences on the target construction, but included English words that might have been challenging for the participants, in order to encourage them to write. Each prompt instructed the participants to write at least three things that they would like to do if they were in the situation described in the prompt (see Appendix I for the three versions). The use of dictionaries was not allowed.

もしあなたが過去に戻って誰にでも 1 人だけ会うことができるとしたら、誰に会いますか？　もしその人物に会ったら、何をしますか？　楽しい可能性を色々と考えて、少なくとも3つあなたがその人物とするであろうことを書いてください。「もしも私が過去に戻る (travel back to the past) ことができるとしたら、、、」で文章を始めること。
If you could travel back to the past and meet one person, who would you meet? If you saw that person, what would you do? Think of fun possibilities and write at least three things that you would do with the person. Start your essay with: "If I could travel back to the past…"

Initially, four versions of the essay tests were prepared and piloted on a group of 15 learners, who were different from the main study participants but comparable in terms of proficiency level and socio-economic background. A series of repeated measures ANOVAs was conducted, and the results showed no statistically significant differences across the four test versions, $F(3, 42) = .867$, $p = .466$, partial $\eta^2 = .058$ (see below for scoring). Therefore, in order to identify a combination with the least difference, additional repeated measures analyses for various combinations of three of the four versions were conducted, which produced the results presented in Table 5.2, below. Based on these results, it was decided to keep the combination comprising versions A, B and C ($F(2, 28) = .248$, $p = .782$, partial $\eta^2 = .017$).

Table 5.2
Repeated ANOVA Results for Comparability of Essay Tests

combination	*F*	*p*	partial η^2
ABC	.248	.782	.017
ABD	1.713	.199	.109
ACD	.780	.468	.053
BCD	.981	.387	.066

The coding of the essay tests examined the suppliance of obligatory contexts where use of the target construction was necessary (Pica, 1983). Then, as presented in Table 5.3 below, the sentences used in these contexts were assessed based on the following four component features: (1) the past form of a verb in an if-clause (two points), (2) an auxiliary plus a verb, (3) the past form of the auxiliary, and (4) the root form of the verb in a main clause (one point each, three in total). Thus, each sentence was assessed based on these four criteria, and one or two points was/were awarded for each, five points being the maximum score (see Izumi & Bigelow, 2000; Shintani et al., 2014, for a similar procedure).

Table 5.3
Data Scoring

Clause	Criterion	Content	Points
If-clause	1	Verb past	2
Main clause	2	Auxiliary + verb	1
	3	Auxiliary past	1
	4	Verb root	1
		Total	5

In order to assess improvements in accuracy (or lack thereof), points per context (P/C) were calculated for each essay by dividing each participant's total score by the number of obligatory contexts produced in the participant's essay. Furthermore, the

number of obligatory contexts (OC) was counted as they reflect attempts at the target construction, a change in which might be an indicator of improvement regarding meaning. According to Bardovi-Harlig (2000), employing a form-focused approach (i.e., focusing on acquisition of the target construction) alone would not be successful in capturing the developmental stages of learners. Examining the acquisition of temporal expressions in a second language, she emphasizes that a meaning-oriented approach (i.e., focusing on expression of the target construction) should also be followed for accurate measurement of learners' development (see also Bardovi-Harlig, 1994, for a similar discussion).

Below are some examples of the scoring. (The second example sentence was actually written by one participant. The same sentence is used with modification to show incorrect sentence types often observed in the essays in order to make comparisons easier.)

Sample scoring:

1. If I met Michael Jackson, I would dance with him.
verb past 2pts auxiliary + verb, auxiliary past, verb root = 3pts Total 5pts (full score)

2. *If I meet Michael Jackson, I will dance with him.
0pt auxiliary + verb, verb root = 2pts Total 2pts

3. *If I met Michael Jackson, I would danced with him.
verb past 2pts auxiliary + verb, auxiliary past = 2pts Total 4pts

4. *If I had met Michael Jackson, I would have danced with him.
0pt auxiliary past = 1pt Total 1pt

5. * If I meet Michael Jackson, I dance with him.
0pt 0pt Total 0pt

I first coded all the essays. Then, in order to ensure the reliability of my coding, a second coder, another Japanese colleague who holds a degree in TESOL, coded one third of each pre- and posttest. Prior to the coding, a practice session was conducted using the essays from the pilot tests. As presented in Table 5.4, Cohen's kappa values for the obligatory contexts and points per context were both high, indicating strong agreement above chance. In addition, I coded all the essays four months after the initial coding and intra-coder agreement analyses were conducted. As shown in Table 5.5, Cohen's kappa values were high as well. A cut-off line of 80% was set for P/C.

Table 5.4
Inter-coder Agreement on OC and P/C

	Cohen's kappa
OC	.93
P/C	.88

Note. *N* = 82, OC: obligatory contexts, P/C: points per context.

Table 5.5
Intra-coder Agreement on OC and P/C

	Cohen's kappa
OC	.93
P/C	.85

Note. *N* = 246, OC: obligatory contexts, P/C: points per context.

5.4.1.2 Grammar Tests

As stated, the three versions of grammar tests used in the pilot study were employed in the main study as well. Based on experience gained in the pilot study, three changes were made regarding the administrative procedure. First, the sheet which contained words supposedly unfamiliar to the participants was distributed two weeks, instead of one week, prior to the experiment, so that they would have ample time to consolidate their vocabulary knowledge (Appendix C). The items on the list were taught

in parts of the previous two classes prior to the experiment. No test, however, was conducted to confirm the expected consolidation, which is one of the limitations of the current thesis. Second, in addition to accurate use and overall scores, overuse scores were also examined in the main study in order to identify possible subtle changes the two scores might have failed to detect. Repeating the same procedure followed in the pilot study, first, an accurate use score was calculated by adding all the points correctly answered on the eight target items (maximum 16, 2 pts × 8 items). Then, an overuse score was calculated from use of the six if-distractor items (Pica, 1983). Namely, each time a participant used the past form for a present non-counterfactual (i.e., simple conditional) item, two points were subtracted, which were tallied, yielding an overuse score (maximum 12, 2 pts × 6 items). In other words, if a participant overused the past tense for two of the present non-counterfactual items, their total overuse score was 4 points (2 pts x 2 items). Finally, an overall score was calculated for each participant by subtracting the overuse score from the accurate use score. The maximum possible overall score was 16 (2 pts × 8 items). Thus, three scores, i.e., accurate use (AU), overuse (OU) and overall (OA) scores were arrived at for each test.

Third, as reported, in order to avoid ceiling effects, a cut-off line of 80%, lower than that in the pilot study (90%), was set for the grammar pretests. Another change in terms of the cut-off line was that it was not set for accurate use scores on the grammar production tests in the main study. (A cut-off line was applied for both accurate use and overall scores in the pilot study.) This was decided upon because some participants who scored 16 for accurate use showed examples of overuse for all six if-distractor items, receiving four points as their overall score. Therefore, in order to examine the participants' learning by changes in terms of scores more accurately, these participants were not excluded. Participants who scored at least 13 (81.3%) out of 16 on either or

both grammar pretests (as stated, overall scores only for grammar production tests) were excluded from the data, and only the data of the participants who scored 12 (75.0%) or less were included.

5.5 Individual Difference Measures

As stated above, language aptitude was measured by the LLAMA_F and LABJ (for inductive language learning ability) and an adapted version of the MLAT (for grammatical sensitivity) one week before the pretest (Week 2). As the LLAMA is a computer-based test and access to computers is necessary, all three aptitude tests were conducted in one of the computer rooms at the school. The three tests are introduced below, according to the order in which they were conducted.

5.5.1 The Modern Language Aptitude Test

The test employed to measure the participants' grammatical sensitivity was an adapted version of the MLAT Part IV, which involves recognising the grammatical roles of words in sentences without having to label, describe or explain them. It is one of the most robust subtests to be used in language aptitude testing (Skehan, 1986) and getting a high score on this test is expected to reflect a test-taker's awareness of the syntactical patterning of sentences and of the grammatical functions of individual elements in a given sentence (Carroll, 1981). There are 45 items in this test, each item consisting of two sentences. One word in the "key sentence" (i.e., a term used in the MLAT guide for the first sentence) is underlined and printed in capital letters, followed by a second sentence with five underlined words. Test-takers are instructed to select the letter of the word in the second sentence that plays the same role in that sentence as the underlined word in the key sentence. Below is an example item retrieved from the MLAT website

(http://lltf.net/mlat-sample-items/mlat-part-iv/):

JOHN took a long walk in the woods.

Children *in blue jeans were singing and dancing in the park.*
A B C D E

The above sample is accompanied by the explanation "You would select 'A,' because the key sentence is about 'John' and the second sentence is about 'children."' As stated above and is clear from this explanation, what test-takers are required to do is to choose the most appropriate one of five choices, and no explanation is required regarding their decision.

As the MLAT was originally created for L1 English speakers learning other languages in order to predict their successful foreign language learning (Carroll, 1981), English is used in the test. Following this, in the adapted version used in this study, the participants' L1, Japanese, was used in order to assess the participants' grammatical sensitivity without being affected by their English proficiency. The development and piloting of the test items was an iterative process supported by my colleague, a Japanese male instructor who holds a degree in teaching Japanese. Following the format of the original in English and referring to Stefanou (2014), who also adopted the MLAT Part IV in Greek, I initially prepared a rough draft with 15 items, which was a Japanese translation of Stefanou's English items translated from Greek. It was piloted by the Japanese teacher, who pointed out many problematic issues (e.g., unnatural use of Japanese), probably deriving from the differences between the two languages (Japanese and English).

Therefore, instead of revising the translated items, it was decided to develop new items in Japanese, following the format of the original. Thus, each Japanese key

sentence includes an underlined word, followed by a second sentence in which five words or phrases are underlined (Appendix J). The participants had to identify which word or phrase in the second sentence served the same grammatical role as the underlined word in the key sentence. With the help of the Japanese male instructor and a Japanese female who holds a degree in teaching Japanese, the first Japanese version with 15 items was developed and piloted on a group of 28 learners with similar backgrounds to the participants in the current study. One point was awarded for each correct answer, 15 points being the maximum score. The average turned out to be 13.1 and there was limited variance, their scores ranging from 11 to 15.

As such, in order to have more variance, it was decided to make the test more challenging as well as to include more items. After many consultations with the two Japanese instructors (conducted separately), 25 items were prepared, including the revised original 15 items. They were piloted on another group of 30 learners who were comparable to the participants in the current study. After the pilot, item analyses were conducted to identify mis-performing items, which were removed one by one, rerunning the analysis each time until acceptable interconsistency reliability (over .70) was achieved. The initial Cronbach's alpha was .591, but when five items were removed, the alpha went up to .732. Thus, it was decided to include the remaining 20 items in the final version of this test. (The average score on these 20 items was 16.2, with a range of nine.) One point was awarded for each correct answer and the maximum possible score was 20. No time limit was set, but most of the participants finished it in four minutes, ranging from three to six minutes. The participants who finished the test early were instructed to read their course textbook.

5.5.2 The Language Aptitude Battery for the Japanese

One of the two test instruments employed to measure the participants' inductive language learning ability was the LABJ (Sasaki, 1991, 1996), which was developed specifically for Japanese-speaking people. So far, several SLA studies (e.g., Robinson, 2005; Shintani & R. Ellis, 2015) have used the test with Japanese university students similar to the participants in the current study and reported significant correlations with measures of explicit learning (Robinson, 2005), and with the effect of written feedback (Shintani & R. Ellis, 2015).

According to Sasaki (1991, 1996), the LABJ was modelled on the short version of the MLAT (Carroll & Sapon, 1959), consisting of the last three parts of its long version, Part III: spelling clues, Part IV: words in sentences, Part V: paired associates, which are intended to measure test-takers' phonetic coding ability, grammatical sensitivity and rote learning activity for foreign language materials, respectively. Therefore, the LABJ also consists of three parts, Part 1: paired associates, Part 2: language analysis and Part 3: spelling clues. As shown, unlike Parts 1 and 3, which correspond to the MLAT Parts V and III, respectively, Part 2 does not correspond to the MLAT Part IV for the following reasons.

As already stated, the MLAT was developed for test-takers whose L1 is English. Accordingly, Part IV is designed to measure test-takers' sensitivity to the English grammar system. Therefore, pointing to the difficulty in creating a test equivalent to it in Japanese, due to the major grammatical differences between the two languages (Sasaki, 1991), Sasaki (1991, 1996), the creator of the LABJ, used the translation of the PLAB Part 4 (Pimsleur, 1966) as the LABJ Part 2. The PLAB Part 4 taps into inductive language learning ability, i.e., one of the four constructs identified by Carroll (1981), but not included in any subtests of the MLAT. Following the PLAB Part 4, all glosses and sentences of the unknown language were changed into *katakana*, a Japanese phonetic

alphabet. The LABJ, therefore, is a translation of both the MLAT (Parts 1 and 3) and the PLAB (Part 2).

For the purposes of the current study, only the LABJ Part 2, which consists of 15 multiple-choice questions, was used. First, test-takers are instructed to examine examples of an artificial language and learn about that artificial language for one minute. Then, based on the knowledge they acquire, they are instructed to infer the grammatical rules of the artificial language based on a set of words and sentences presented with their Japanese translations. In each question, a Japanese sentence is presented and the participants are asked to choose the correct translation in the artificial language from four choices provided. (For copyright reasons, no sample of the LABJ can be included in this thesis.) One point is given for each correct answer, 15 points being the maximum score. The test was administered soon after the MLAT. The time limit was six minutes and all participants completed the test at the same time. The whole procedure took about 15 minutes. Although it is a validated test, Cronbach's alpha was calculated to assess its reliability and this was found to be high (.823). The LABJ Part 2 is denoted as the LABJ in the remainder of this thesis.

5.5.3 LLAMA_F

Another test instrument employed to gauge the participants' inductive language learning ability was the LLAMA Language Aptitude test (Meara, 2005). As mentioned earlier, it is a computer-based test, consisting of four subtests, LLAMA_B, a test of vocabulary learning, LLAMA_D, a test of sound recognition that requires previously heard sound sequences to be identified in new sequences, LLAMA_E, a test of sound-symbol associations, and LLAMA_F, a test of grammatical inferencing. Like the LABJ, an artificial language is used in the LLAMA, so it is accessible to people regardless of

their L1. Analysing the reliability of the LLAMA, Granena (2013) compared the test scores of learners with three different L1s (Spanish, Chinese and English) and found no statistically significant differences among the three L1 groups, providing some evidence that the LLAMA is a language-independent test. It has been used in several SLA studies on participants with various L1s, such as Turkish (e.g., Yalçin, 2012; Yalçin and Spada, 2016; Y. Yilmaz, 2013) and Spanish (e.g., Abrahamsson & Hyltenstam, 2008).

As reported earlier, only LLAMA_F, which is "particularly good at identifying outstanding analytical linguists" (Meara, 2005, p. 18), was used for the purposes of the present thesis. Like the LABJ, the sub-test consists of two phases, a learning phase and a testing phase. In the learning phase, test-takers are instructed to learn as much as they can about the artificial language in five minutes. There are 20 small buttons in the main panel, and every time a test-taker clicks one of them, a picture and a sentence that describe it appear on the screen. Test-takers are allowed to take any notes during this phase, which they can then use in the testing phase. When their time is up, test-takers hear a bleep, and the second phase (i.e., testing phase) starts when they click on an arrow button.

In the testing phase, the program displays a picture and two sentences that describe the picture. "One sentence is grammatically correct, while the other contains a major grammar error" (Meara, 2005, p. 17) and test-takers are required to choose one of the sentences based on the grammar rule they learned in the first phase. There are 20 test items and this phase is untimed. When test-takers have finished answering all the questions, their scores, ranging from 0 to 100, are displayed on the panel. According to Granena (2013), who investigated the internal consistency of the LLAMA, i.e., "the degree to which the results on the tests were consistent across items" (p. 117), the Cronbach's alpha coefficient of the overall LLAMA test was in the acceptable range

of .77, that of LLAMA_F, however, was slightly lower at .60. Henceforth, LLAMA_F is referred to as the LLAMA.

In order to administer the test efficiently, I demonstrated the procedure with an overhead projector. After the demonstration, worksheets for participants to take notes in the first phase were distributed (Appendix K) and they started the program individually. On finishing the testing phase, they were instructed to print out their results using a printer in the computer room, and they submitted both their worksheets and printouts of their results. Although the second phase was not timed, most of the participants finished the test in around 10 minutes (7 minutes on average, ranging from 4 to 12 minutes). Once again, participants who finished the test early were instructed to read their course textbooks.

5.5.4 Metalanguage Knowledge Test

As mentioned, although the use of metalinguistic terminology was not required for WL, given that WL requires participants to write down their thoughts concerning their linguistic issues, familiarity with metalinguistic terminology was considered to be important for expressing their thoughts. It was decided, therefore, to include a test of the participants' metalanguage knowledge. Of the tests on metalanguage knowledge created so far (e.g., Alderson, et al, 1997; Berry, 2009), probably one of the earliest and most famous is the one developed by Alderson et al. for L1 English speakers. It consists of two sections. In the first section, consisting of four parts, participants are instructed to read a target sentence in either their L1, English (Parts 1 and 2) or L2, French (Parts 3 and 4) and to identify various parts of speech, such as nouns and verbs.

The test devised by Alderson et al. (1997) has been adapted by many researchers, including Elder (2009), who used an adaptation of the test to investigate the validity of

metalinguistic knowledge tests. The adapted version of her test also consists of two parts. In Part 2, which is based on Section 1 of Alderson et al.'s test, participants were requested to match items from a list of grammatical terms to their corresponding exemplars in an English sentence. Modifying the original version by Alderson et al., Elder included terminology relevant to grammatical structures in the first part of her test. Sachs (2010) is another researcher who has adapted parts of Alderson et al.'s instruments, i.e., Parts 1 and 2, which involved identifying parts of speech in the L1, English. The participants in her study were asked to read a single paragraph and find exemplars for 16 grammatical terms in the paragraph. Like Elder, Sachs included terms that are relevant to stating rules about reflexives, the target construction of her study (e.g., direct object, reflexive pronoun), as well as ones that were not as closely related (e.g., subordinating conjunction, adverbial subordinate clause), all of which are more difficult than the ones used in Alderson et al.'s test.

Referring to the content and formats of the tests used in the studies above, a brief metalanguage knowledge test was developed for the current study (Appendix L). Following Sachs' (2010) instrument, a short passage in English (79 words) was prepared. (The passage was written in simple English in order to be independent of the participants' L2 proficiency.) Unlike her test, however, the grammar terms were given in Japanese, the participants' L1. Moreover, none of them were difficult, all being simple parts of speech (e.g., noun, verb), including those that were likely to be necessary for the WL task of the current study (e.g., verb root, auxiliary), following Elder (2009) and Sachs. Thus, the participants in the current study were instructed to read the passage in English and fill in the blanks next to each item of metalinguistic terminology written in Japanese with appropriate English words in the passage. The rationale of giving the terms in Japanese was to provide a similar condition to that of the treatment task, where

the participants were required to compare an original text with their reconstruction, both written in English, and to write their thoughts in Japanese on the comparison. One point was awarded for each correct answer and the maximum possible score was 17.

An initial version of this test was first piloted on the same two native speakers who cooperated with me on the grammar tests (an American male and a British female). (The Japanese grammar terms were translated into English for them.) They found no problem with the grammar terms, but suggested some revisions to the original passage. Therefore, after minor revisions to the original text, the revised version was piloted on the same Japanese female who holds a degree in TESOL and also cooperated with me in piloting the grammar tests. As no major issues were identified, the revised version was piloted on the same 28 participants who participated in the pilot of the initial version of the MLAT with 15 items. The average was 10.7, ranging from 3 to 16. (None of the participants scored the maximum.) Cronbach's alpha was computed to test internal-consistency reliability and this was in the acceptable range (.806).

5.6 Treatment

As stated, the treatment task employed in the pilot study, individual written dictogloss, was used in the main study as well (see sections 4.6 and 4.7 for details and the procedure). The only difference between the two studies was that an additional practice session was included in the main study in order to conduct the experiment more efficiently than the pilot study. Therefore, practice sessions for WL and dictogloss were given beforehand over a period of six weeks. After two training sessions on WL and two practice sessions on dictogloss, given one per week for four weeks, a practice session that combined WL and dictogloss was conducted in each of the final two weeks. (This combined practice was offered only once in the pilot study.) Like the pilot study, as the

sessions in this experiment were conducted during my regular classes, all the participants in the treatment groups, including the ones who were later assigned to the –WL group, joined the WL practices, which might have had some effect on the results. Similarly, all the participants practised the treatment task (i.e., dictogloss) over four weeks before the experiment, which might have resulted in a task repetition effect (Bygate, 2001, 2018). To be more precise, there is a possibility that the experience might have caused the participants to focus their attention closely on the relevant form-meaning mappings of the treatment task.

5.7 Coding of T-WLEs

In order to address the research questions that investigate the frequency and quality of T-WLEs (target construction-related WLEs) in relation to L2 learning (RQ2, RQ3, RQ4) and individual differences in aptitude and metalanguage knowledge (RQ7, RQ8), all the T-WLEs were coded based on how frequently they were produced (frequency) and what level of awareness, i.e., noticing or understanding, they demonstrated (quality). Below are details of the coding for each category.

5.7.1 Frequency of T-WLEs

In order to examine the relationship between the frequency of T-WLEs and L2 learning, all the WLEs were coded into two categories depending on whether or not they focused on the target construction: target construction-related WLEs (T-WLEs) or non-target construction-related WLEs (NT-WLEs). Any WLEs that were related to the target construction, such as explanations of grammar rules as well as translations of the target sentences, were categorised as T-WLEs, and the rest as NT-WLEs. Below are examples of each type of WLE.

T-WLEs:

1. *kateiho* (counterfactual)

2. "*if" dakara kako* (if so past tense)

3. *moshi eigo ga hanasetara…* (Japanese translation of "if I could speak English…")

NT-WLEs:

4. *fukusukei datta* (oh, plural form)

5. *kono tango no imi wakara nai* (I don't know the meaning of this word.)

6. *kore oomoji ni shinakatta* (I didn't capitalise this letter.)

5.7.2 Quality of T-WLEs

In order to investigate the possible impact of the quality of T-WLEs on L2 learning, all the T-WLEs were further coded depending on the level of awareness which the participants' comments displayed; that is, "noticing," i.e., "being aware of something" (Schmidt, 1993, p. 211), the target construction in this thesis, or "understanding," i.e., a higher level of awareness (Schmidt, 1990, 1994, 1995, 2001). According to Schmidt (2010), "knowledge of rules and metalinguistic awareness of all kinds belong to this higher level of awareness" (p. 5). Therefore, following his distinction, all the T-WLEs were coded into one of three categories. Below are explanations of each category, followed by respective example sentences.

Noticing:

Any T-WLEs that demonstrated the participants' noticing of the target construction were coded as noticing (examples 1, 2 and 3). In addition, T-WLEs that showed incorrect understanding, such as incorrect explanations of the target

construction (example 4.1) and incorrect translations (example 4.2), were also categorised as noticing.

1. *if no bun* (if sentence)

2. *kateiho* (counterfactual)

3. *kako kei* (past tense)

4. Next to the the target sentence: If I were good at computers, I could find a good job.

4.1. *kako no hanashi* (This is about past.)

4.2. *yoi shigoto wo mitsuketa* (I found a good job) *incorrect translation

Understanding:

T-WLEs that showed participants' correct understanding, such as an explanation of the target (examples 5 and 6), the grammar rule (example 7) or full translations (example 8), were coded as understanding.

5. *jijitsu to kotonaru toki kateiho de kakokei*
(counterfactual for something that is not true and use past tense)

6. *kateiho dakara jisei ga hitotsu sagaru*
(counterfactual so tense should be moved backward)

7. if + S + verb past, S + would/could + verb root (correct grammar rule)

8. *moshi eigo ga hanasetara kaigai de hataraku darouni*
(full correct translation of "if I could speak English, I could work abroad")

Partial understanding:

Episodes that seemed to demonstrate more than noticing, but not quite understanding, such as a partial explanation of the grammar rule of the target construction (example 9), reflection/mention of the use of tense (examples 10 and 11) or

a partial translation (example 12), were categorized as "partial understanding."

9. *if dakara kakokei* (if so past tense)
10. Underlining "if" and/or a verb in past tense:

    genzaikei nishiteta (I used present tense.)
11. *kateiho jisei cyuui*

    (I have to be careful about the tense of the counterfactual conditional.)
12. *eigo ga hanasetara* (if I could speak English) *partial translation

I first coded all the WLEs. Then, in order to ensure the reliability of my coding, all the WLEs were coded by a second researcher, the same Japanese female who holds a degree in TESOL and served as a second coder for the essay tests. As shown in Table 5.6, the Cohen's kappa value for the total number of WLEs was moderately high (.87) and those for the frequency and quality of T-WLEs were both over .90, indicating a high level of agreement between the two coders' judgements. In addition, like the essay tests, I coded all the WLEs again to check the intra-coder agreement and found strong agreement above chance, with kappa values ranging between .88 and .94. The results of the analyses are shown in Table 5.7, below.

Table 5.6
Inter-coder Agreement on the Total of WLEs and the Frequency and Quality of T-WLEs

	Cohen's kappa
WLEs	.87
Frequency	.92
Quality	.91

Note. $N = 33$.

Table 5.7
Intra-coder Agreement on the Total of WLEs and the Frequency and Quality of T-WLEs

	Cohen's kappa
WLEs	.94
Frequency	.91
Quality	.88

Note. $N = 33$.

5.8 Questionnaires

In order to obtain information in terms of the participants' background and their perspectives on their experience of the experiment, two questionnaires, i.e., a background questionnaire and an exit questionnaire, were administered before the pretest and after the delayed posttest, respectively. Both questionnaires were administered to the participants in their L1, Japanese, in order to encourage them to respond clearly. Each questionnaire is described in detail below.

5.8.1 Background Questionnaire

In order to obtain information regarding the participants' basic biodata, e.g., age, previous study experience of English and studying abroad, the background questionnaire was given to the participants in all groups: the two treatment groups in Week 1 and the control group in Week 2, which was the first week for the control group (see Appendix B for details). It was not timed. Participants who finished the questionnaire early were instructed to read their course textbooks, and the questionnaire sheets were collected when all the participants had finished. The average time for the participants to finish it was four minutes, ranging from three to seven minutes.

5.8.2 Exit questionnaire

In Week 5, the final week, the exit questionnaire was only given to the participants in the two treatment groups immediately after the delayed posttest. Two versions were prepared, one for the +WL group and the other for the –WL group. The two versions differed only in that the one for the +WL group included extra four questions concerning their experience of WL in order to investigate the participants' perceptions with respect to the experience. As with the background questionnaire, no time limit was set. The participants who finished the questionnaire early were instructed to read their course textbooks and the questionnaires were collected when all the participants had finished. On average, it took the –WL participants four minutes to finish the questionnaire, ranging from three to eight minutes, and seven minutes on average for the +WL participants, ranging from six to 14 minutes. Below are the questions asked (see also Appendix M).

Part I. Reflecting on the experiment: (both groups)

1. What do you think was the purpose of the experiment? (Besides just doing research on language learning, was there something specific you thought I might be studying?)
2. Do you think that you learned something about English during the 3-week experiment? If yes, what?
3. Related to Q2, are there any vocabulary items or grammar rules you learned? If so, please write them below in detail.
4. If there are any vocabulary items or grammar points you focused on, please write them in detail.
5. Please write any other comments you may have regarding this experience. (Your comments may overlap with what you wrote above.)

Part II. Regarding last week's task of writing your thoughts or questions while checking the original text: (+WL group only)

1. What feelings do you have about the experience of writing your thoughts and questions? Why?
2. In the task, you wrote your thoughts while checking the original text. However, do you think there would be a difference if you hadn't written your thoughts while checking? If so, why?
3. Do you think you learned something from the experience of writing while checking? Why or why not?
4. Please write any comments you may have regarding the experience of writing your thoughts while checking the original text.

5.9 Interviews

In order to complement the written data obtained from the questionnaire, two +WL participants from two of the three classes were interviewed in terms of their WL experience immediately after the posttest in Week 4. (In Week 3, I asked for volunteers in each of the three classes participating as the two treatment groups, and one participant from each of the three classes volunteered to be interviewed. One of them, however, missed the class in Week 4.) It was a semi-structured interview, in that all the questions in the exit questionnaire were asked. With the participants' permission, the interviews were recorded and transcribed. Their data were excluded from the analyses as the interviews were expected to raise the interviewees' consciousness toward the target construction or the purpose of the experiment and to confound the results of the delayed posttest.

5.10 Case Studies

In order to achieve a deeper understanding with respect to the potential impact of WL on L2 learning, a case study approach was employed from an SCT perspective. Four case-study participants were chosen and their approach to WL and WLEs was analysed in relation to their tests scores and the questionnaire results.

5.11 Statistical Analyses

The data were analysed with SPSS 23.0. The level of significance was set at .05. First, in order to address each research question, descriptive statistics for the pre-and posttests (RQ1), the frequency of T-WLEs (i.e., target construction-related WLEs) (RQ2), the quality of T-WLEs (RQ3), the correlation between the frequency and quality of T-WLEs (RQ4), aptitude tests (RQ5), the metalanguage knowledge test (RQ6), the correlations between the frequency of T-WLEs and the results of aptitude tests and the metalanguage knowledge test (RQ7) and the correlations between the quality of T-WLEs and the results of aptitude tests and the metalanguage knowledge test (RQ8) were calculated. Skewness and kurtosis ratios were obtained for each test and category of T-WLEs separately.

Then, as for RQ1, RQ5 and RQ6, the normality of distributions was examined using the skewness and kurtosis ratios in order to identify if the results satisfied the assumptions underlying parametric statistical analyses (i.e., if the distributions of the results were normal). In terms of RQ1, some of the ratios were found to be outside the acceptable range of [–2, 2]. Non-parametric inferential statistical analyses were, therefore, conducted to interpret the data, as they do not require data to be normally distributed. Similarly, with respect to RQ5, as one of the ratios turned out to have violated the assumption for parametric tests, non-parametric inferential analyses were

conducted. Although the skewness and kurtosis ratios for RQ6 were in the acceptable range, non-parametric analyses were conducted to be consistent with RQ5.

Similarly, for RQ3, RQ4 and RQ8, as the quality of T-WLEs was rank-order data, non-parametric analyses were administered. Meanwhile, for RQ2 and RQ7, although the frequencies of T-WLEs were interval data and their skewness and kurtosis ratios were in the acceptable range of [–2, 2], i.e., the distribution of the data was normal, non-parametric inferential statistical analyses were conducted in order to be consistent with RQ3, RQ4 and RQ8 (see Appendix N for skewness and kurtosis ratios for all the tests).

Finally, as stated, some of the data were not normally distributed. Thus, medians and interquartile ranges were employed as measures of the central tendency and variation, respectively, because they are "resistant to extreme values" (Larson-Hall, 2010, p. 65). For consistency throughout the thesis, medians were employed for normally-distributed data as well. It should be pointed out, however, that for the results of the aptitude tests (RQ5), means were presented to make the comparisons with the results of other studies easier. In order to be consistent with RQ5, means were also used for the results of the metalanguage knowledge test (RQ6).

For RQ1, which addressed the impact of WL on L2 learning, the distributions were found to deviate from normality, that is, the skewness and kurtosis ratios were outside the [–2, 2] interval for some of the gain scores on the three assessment tests (i.e., essay tests, grammar production tests and recognition tests). It was decided, therefore, to run non-parametric tests. First of all, Mann-Whitney tests were conducted to ensure that there were no significant differences among the three groups at the pretest. Then, a series of Mann-Whitney tests was conducted, combining two of the three groups, to see if there were any significant differences regarding L2 learning measured by the gain scores between the pretest-posttest and pretest-delayed posttest on the three assessment

tests.

RQ2 investigated the relationship between the frequency of T-WLEs and L2 learning (again, measured by the gain scores between the pretest-posttest and pretest-delayed posttest on the three assessments, i.e., essay tests, grammar production tests, and recognition tests). As stated above, although the distribution of frequency counts is interval data and no violations of the assumptions for parametric tests were detected, non-parametric tests, i.e., Spearman's correlation analyses, were conducted instead of Pearson's correlation analyses in order to be consistent with the analyses of the quality of WLEs (RQ3, RQ4 and RQ8),

In terms of the quality of T-WLEs (RQ3), first of all, all the +WL participants were categorised into four groups, numbered from 0 to 3, depending on their levels of awareness measured by their T-WLEs. The participants who produced no T-WLEs received a rating of 0, those who showed noticing in their T-WLEs were given a rating of 1, participants with partial understanding on their T-WLEs were awarded a rating of 2, and those who displayed understanding in their T-WLEs were given a rating of 3. When a participant produced multiple T-WLEs with different levels of quality, the highest level was used as the basis for categorisation. Unlike the frequencies of T-WLEs, the categories of the quality of T-WLEs are ordinal in nature. Thus, nonparametric correlations, Spearman's correlation analyses, were conducted to investigate the correlations between the quality of T-WLEs and the gain scores between the pretest-posttest and pretest-delayed posttest assessments here as well.

In addition, in order to address RQ4, the association between the frequency and quality of T-WLEs was investigated to see the extent to which they were related to each other. In other words, RQ4 sought to find out if the more frequently learners produced T-WLEs, the more likely they were to show higher understanding. As stated above, the

frequency and quality of T-WLEs are interval and rank-order data, respectively. So Spearman's correlation analyses were conducted here as well.

With respect to RQ5, the associations between the scores on the three aptitude tests (i.e., the MLAT, LABJ, LLAMA) and the gain scores between the pretest-posttest and pretest-delayed posttest on the three assessment tests (i.e., essay tests, grammar production tests, recognition tests) were examined in order to see whether the effects of WL depended on the participants' aptitude. Like the procedure followed for RQ1, the distributions of the participants' aptitude test results were examined, which detected a violation in one of the assumptions for parametric tests, i.e., one of the ratios was outside the acceptable range of +/−2. Therefore, a series of Spearman's correlation analyses were also conducted in order to identify the relationships between the participants' aptitude and gain scores between the three sets of pretest-posttest and pretest-delayed posttest assessments here as well.

Similarly, for RQ6, the associations between the scores on the metalanguage knowledge test and the gain scores on the three assessments (essay tests, grammar production tests, recognition tests) were examined in order to see whether the effects of WL depended on the participants' metalanguage knowledge. Following the procedure for the previous research questions, the distribution of the results of the metalanguage knowledge test was examined first, which confirmed the normality of its distribution. However, in order to be consistent with the other correlation analyses (RQ2, RQ3, RQ4 and RQ5), a series of Spearman's correlation analyses were again performed here.

In order to address RQ7, the frequency of T-WLEs was examined in relation to the results of the three aptitude tests and the metalanguage knowledge test. As stated, frequency counts constitute interval data, but the aptitude tests results were not normally distributed. Thus non-parametric Spearman's correlation analyses were conducted.

Similarly, for RQ8, the associations between the quality of T-WLEs and the results of the aptitude tests and the metalanguage knowledge test were investigated. As mentioned above, the quality of T-WLEs is rank-order data and the aptitude tests results violated the assumptions of parametric tests. Therefore, Spearman's correlation analyses were again performed.

Finally, in terms of interpreting the magnitude of effect sizes, Cohen's (1988) guidelines have frequently been employed in many SLA studies to date. They are, however, rather arbitrary and not relative to the field of research (Norouzian & Plonsky, 2017). On this point, Larson-Hall (2010) argues that "the importance of effect sizes … should very much depend on the field itself" (p. 115). As such, effect sizes were interpreted according to the standards in SLA research suggested by Plonsky and Oswald (2014), which are summarised in Table 5.8, below. As in the pilot study, *d* values were interpreted according to their field-specific benchmarks of small (*d* = .40), medium (*d* = .70) and large (*d* = 1.00) for differences between groups. And as for correlation coefficients, the benchmarks of small (*r* = .25), medium (*r* = .40) and large (*r* = .60) were adopted when interpreting the effect sizes of the results. In terms of eta-squared, squared values of *r* were employed to interpret effect sizes. Therefore, the benchmarks of small (η^2 = .06), medium (η^2 = .16) and large (η^2 = .36) were adopted. The same benchmarks were employed to interpret partial eta-squared as well.

Table 5.8
SLA-specific Effect Sizes Standards from Plonsky and Oswald (2014)

	small	medium	large
d values	.40	.70	1.00
r values	.25	.40	.60
eta-squared/partial eta squared	.06	.16	.36

CHAPTER VI
QUANTITATIVE RESULTS

The results of the statistical analyses are reported according to each research question in this chapter. In each section, the results for descriptive statistics are presented first, followed by those for inferential statistics.

6.1 Effects of WL on L2 Learning (RQ1)

As reported above, the normality of distributions was not confirmed in terms of the gain scores of the three tests results (i.e., skewness and kurtosis ratios were outside the [−2, 2] interval for some of the dependent variables). Therefore, non-parametric tests, i.e., Mann-Whitney tests, were employed in order to investigate the impact of WL on learning the target construction. The results are reported by each test below.

6.1.1 Essay Tests

As stated earlier, in order to assess learning based on the essay tests, two scores, i.e., numbers of obligatory contexts (numbers of attempts at the target construction) and points per context (points per each obligatory context), were examined. Table 6.1 presents descriptive statistics for the results of the essay tests of the three groups, which are also graphically illustrated in Figures 6.1 and 6.2. Also as stated, as the distributions were not normal, medians were employed as a measure of the central tendency. Accordingly, interquartile ranges were used as a measure of variation. Medians and interquartile ranges are expected to provide more reliable information than means and standard deviations when data are not normally distributed (Field, 2013).

As for obligatory contexts, where use of the target construction was necessary, all the groups produced four contexts on the pretest. On the posttest, however, only the

+WL group increased their number of obligatory contexts to 5.00 on the posttest. Furthermore, as reflected at the higher end of the 95% confidence interval (i.e., 7.50), some +WL participants produced noticeably higher numbers of obligatory contexts than they did in the pretest. Although their median stayed the same for the delayed posttest, it was still higher compared to that of the pretest. In contrast, the number of obligatory contexts of the –WL group remained the same in both posttests (i.e., 4.00). The control group decreased their scores somewhat, from 4.00 on the pretest and posttest to 3.50 on the delayed posttest.

As stated, points per context were calculated in order to examine possible changes in accuracy by dividing each participant's total score by the number of obligatory contexts. Again, the three groups demonstrated the same median, i.e., .00, for the pretest. Replicating the result for obligatory contexts, only the +WL group increased their points per context over the three time periods, demonstrating medians of .33 and .50 for the posttest and delayed posttest, respectively. In contrast, the medians of the –WL group and the control group did not show any improvement, remaining at .00.

Table 6.1
Descriptive Statistics for the Essay Tests for the Three Groups

		Pretest			Posttest			Delayed Posttest		
	N	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI
OC										
+WL	33	4.00	3	[4.00, 5.00]	5.00	4	[5.00, 7.50]	5.00	4	[4.00, 6.50]
-WL	31	4.00	3	[3.00, 5.00]	4.00	3	[3.00, 5.00]	4.00	2	[3.50, 5.00]
CTR	18	4.00	2	[3.50, 5.00]	4.00	2	[3.00, 5.00]	3.50	2	[3.00, 5.00]
P/C										
+WL	33	.00	.45	[.00, .35]	.33	.63	[.00, .46]	.50	1.25	[.00, 1.00]
-WL	31	.00	.57	[.00, .25]	.00	.80	[.00, .25]	.00	.67	[.00, .50]
CTR	18	.00	.70	[.00, .60]	.00	.50	[.00, .25]	.00	.55	[.00, .40]

Note. OC: number of obligatory contexts, P/C: points per context, maximum score for P/C = 5.

Figure 6.1
Medians of Obligatory Contexts

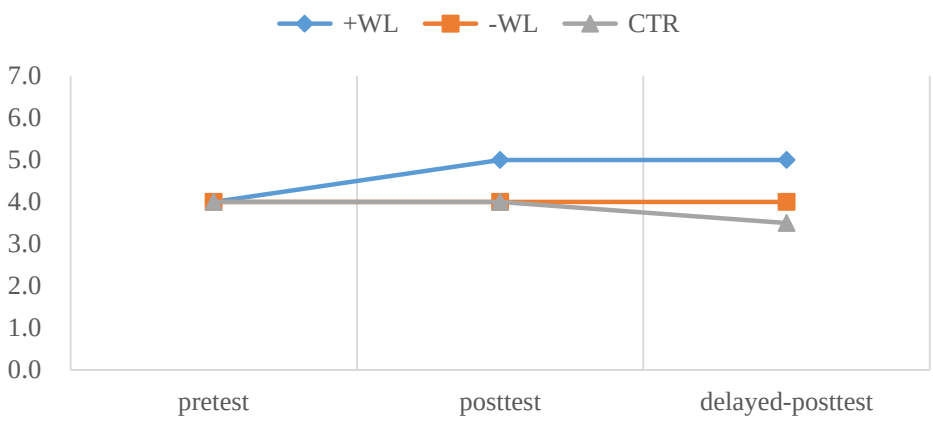
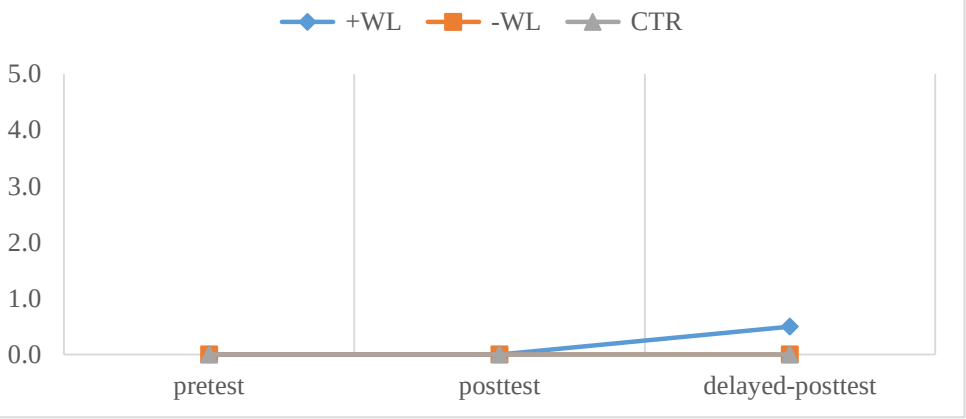


Figure 6.2
Medians of Points per Context



Moving onto inferential statistics, first, a series of Mann-Whitney tests was performed with three combinations of two of the three groups (i.e., +WL group vs –WL group, +WL group vs Control group, –WL group vs Control group) to compare their pretest results in order to ascertain whether there were any statistically significant initial differences among them. As shown in Table 6.2, the results demonstrated that there were no significant differences among the groups at the outset of this experiment.

Table 6.2
Mann-Whitney Tests on the Essay Pretest Scores

		z	*p*	*d*
+WL vs –WL	OC	-.299	.765	.092
	P/C	-.283	.777	.070
+WL vs CTR	OC	-1.167	.243	.303
	P/C	-.436	.662	.122
–WL vs CTR	OC	-.612	.541	.175
	P/C	-.580	.562	.167

Note. OC: number of obligatory contexts, P/C: points per context.

Next, another series of Mann-Whitney tests was conducted, again with three combinations of two of the three groups, in order to identify if there were any significant differences among the pretest-posttest and pretest-delayed posttest gain scores for the numbers of obligatory contexts and points per context (run separately). The results of the analyses appear in Table 6.3 (obligatory contexts) and Table 6.4 (points per context), below.

In terms of obligatory contexts, statistically significant differences emerged between the +WL group and the other two groups. As for the +WL group and the –WL group, statistically significant differences were found between both the pretest-posttest and pretest-delayed posttest gain scores in the range of large and small, respectively. With respect to the +WL group and the control group, a statistically significant difference was only found between the pretest-posttest gain scores with a size of medium, and not between the pretest-delayed posttest gain scores. No significant differences were observed between the –WL group and the control group. These results mean that the +WL group produced a greater number of obligatory contexts at the posttest and delayed posttest than the –WL group, and at the posttest than the control group, using the pretest as the baseline.

Meanwhile, as for points per context, which is an accuracy measure, no

statistically significant differences were found among the groups.

Table 6.3
Mann-Whitney Tests on the Gain Scores of OC

	gain scores	*z*	*p*	*d*
+WL vs −WL	pre-post	-3.881	<.001	1.110***
	pre-delayed post	-2.300	.021	.600*
+WL vs CTR	pre-post	-2.684	.007	.811**
	pre-delayed post	-.793	.427	.223
−WL vs CTR	pre-post	-.873	.382	.251
	pre-delayed post	-1.210	.226	.351

Note. OC: number of obligatory contexts, * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 6.4
Mann-Whitney Tests on the Gain Scores of P/C

	gain scores	*z*	*p*	*d*
+WL vs −WL	pre-post	-.262	.793	.066
	pre-delayed post	-1.413	.158	.359
+WL vs CTR	pre-post	-.709	.478	.200
	pre-delayed post	-1.566	.117	.450
−WL vs CTR	pre-post	-.518	.605	.148
	pre-delayed post	-.450	.653	.129

Note. P/C: points per context.

In sum, as reflected in the results for descriptive statistics, inferential statistics revealed that the +WL group attempted to use the target construction much more than the –WL group (in both the short and long term) and the control group (in the short term). It should be pointed out, however, that no such differences were observed in terms of points per context, i.e., accuracy, among the groups. It is also worth noting that no statistically significant differences were observed between the –WL group and the control group, in spite of the exposure to the text with the target sentences for the –WL group.

6.1.2 Grammar Production Tests

Table 6.5 presents descriptive statistics for the results of the three scores on the grammar production tests, i.e., accurate use (AU) score, overuse (OU) score and overall (OA) score, of the three groups. In addition, Figures 6.3–6.5 illustrate the results graphically. As explained earlier, accurate use scores were obtained by tallying the points of the eight target items, overuse scores by tallying the points of the six if-distractors, and overall scores by subtracting overuse scores from accurate use scores.

As for the +WL group, both their accurate use and overall scores rose over the three time periods. The increases from pre- to posttest of both scores were noticeable, especially the overall score, which tripled from 2.00 to 6.00. Its overuse score stayed the same at 2.00. Meanwhile, all the scores of the –WL group showed rather different changes from those of the +WL group. To be more specific, they lowered their accurate use score from 8.00 on the pre- and posttests to 5.00 on the delayed posttest. Meanwhile, although they increased not only their overall but also overuse scores (i.e., an indication of the increase in overuse) on the posttest, their scores decreased on the delayed posttest, resulting in the same scores as the pretest. Thus, the –WL group demonstrated the lowest medians of the three groups in terms of accurate use and overall scores (5.00 for accurate use score, 2.00 for overall score). It should be pointed out, however, as reflected in the rather large interquartile ranges and wide 95% confidence intervals, there was a noticeable variance in the group.

Finally, compared to the two treatment groups, the control group demonstrated only moderate changes. Their accurate use scores increased over the three testing sessions (from 6.50 on the pretest to 8.00 and 9.00 at the posttest and delayed posttest, respectively). Meanwhile, their overall score showed a decrease from 3.00 on the pretest to 2.00 on the posttest but recovered to the same score at 3.00 on the delayed posttest as

compared to the pretest. It is worth mentioning that their overuse score on the delayed posttest increased slightly from 2.00 on the pretest and posttest to 3.00, making it the only group that showed an increase in overuse on the delayed posttest.

Table 6.5
Descriptive Statistics for the Grammar Production Tests for the Three Groups

		Pretest			Posttest			Delayed Posttest		
	N	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI
AU										
+WL	33	6.00	8	[3.00, 8.00]	10.00	9	[8.00, 14.00]	12.00	10	[8.50, 15.00]
−WL	31	8.00	9	[5.00, 10.00]	8.00	11	[4.03, 12.00]	5.00	13	[3.00, 14.00]
CTR	18	6.50	5	[5.00, 9.00]	8.00	10	[2.00, 11.00]	9.00	11	[.00, 10.50]
OU										
+WL	33	2.00	4	[.00, 4.00]	2.00	6	[.00, 5.00]	2.00	4	[.00, 3.00]
−WL	31	2.00	4	[.00, 4.00]	4.00	6	[.00, 4.00]	2.00	4	[.00, 4.00]
CTR	18	2.00	5	[.00, 4.00]	2.00	6	[.00, 5.00]	3.00	7	[.00, 5.00]
OA										
+WL	33	2.00	6	[.00, 5.00]	6.00	10	[4.00, 10.00]	8.00	13	[3.00, 13.50]
−WL	31	2.00	8	[.50, 5.00]	4.00	12	[2.00, 7.00]	2.00	12	[.00, 8.00]
CTR	18	3.00	6	[1.00, 5.50]	2.00	6	[1.00, 6.00]	3.00	7	[.00, 5.00]

Note. AU: accurate use, OU: overuse, OA: overall, maximum score for AU & OA =16, OU = 12.

Figure 6.3
Medians of Accurate Use Scores

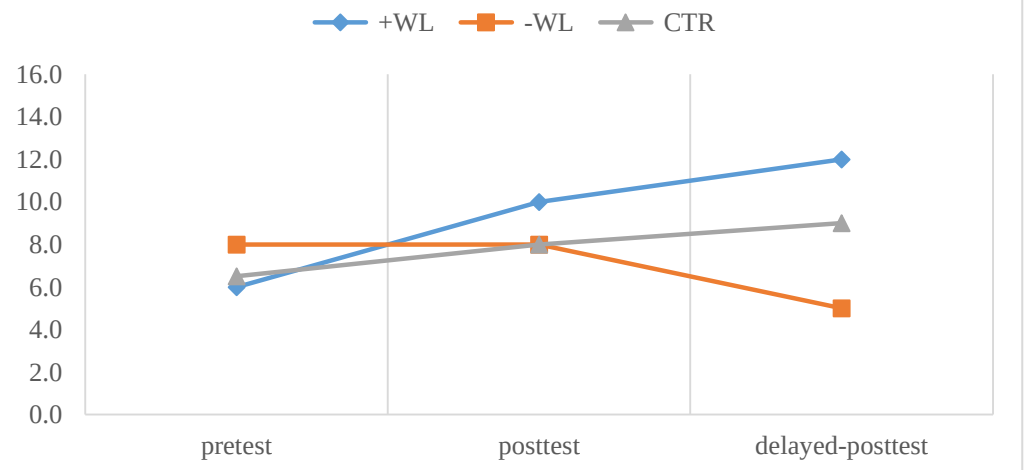


Figure 6.4
Medians of Overuse Scores

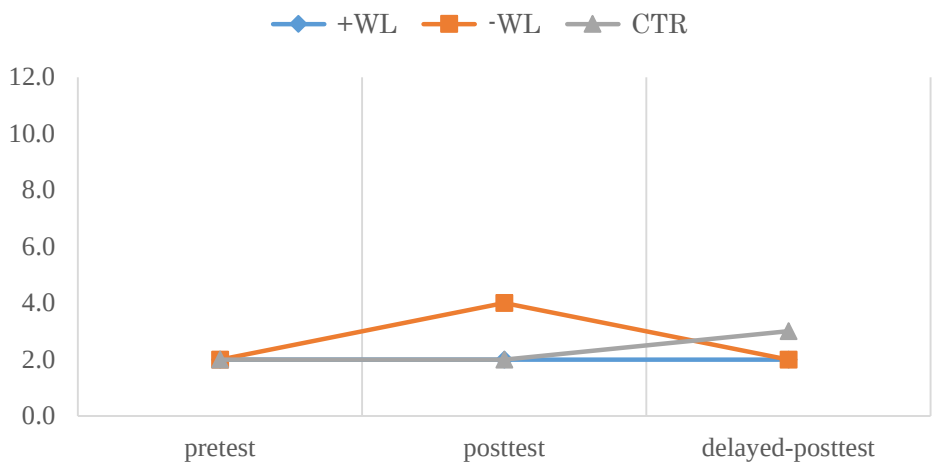
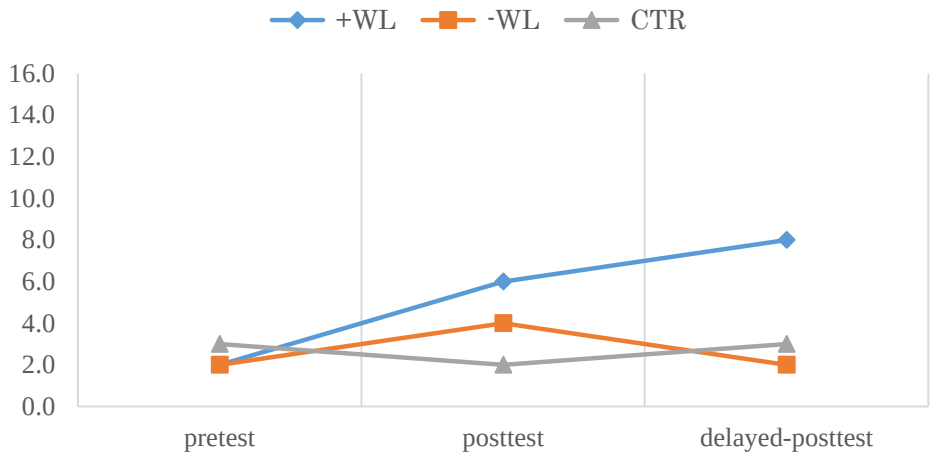


Figure 6.5
Medians of Overall Scores



As with the essay tests, a series of Mann-Whitney tests was first performed to ascertain that there were no significant differences among the three groups at the outset of this experiment. They were run separately for each score (i.e., accurate use score, overuse score, overall score). As presented in Table 6.6, no statistically significant differences were observed among the three groups for any of the scores.

Table 6.6
Mann-Whitney Tests on the Grammar Production Pretest Scores

		z	*p*	*d*
+WL vs −WL	AU	-.616	.538	.154
	OU	-.014	.989	.004
	OA	-.312	.755	.078
+WL vs CTR	AU	-.288	.773	.080
	OU	-.399	.690	.112
	OA	-.308	.758	.086
−WL vs CTR	AU	-.396	.692	.116
	OU	-.398	.691	.114
	OA	-.146	.884	.042

Note. AU: accurate use, OU: overuse, OA: overall.

Then, in order to determine whether the three groups differed in terms of the changes in their posttests scores, another series of Mann-Whitney tests was performed with two of the three groups with the gains of three scores (run separately).

As for the accurate use gain scores, as presented in Table 6.7, the analyses demonstrated statistically significant differences between the +WL group and the control group in both the short and the long term with small effect sizes. Meanwhile, no significant differences were found between either the +WL group and the –WL group or the –WL group and the control group. These results mean that the +WL group achieved higher gain scores than the control group in both the short and the long term, but not the –WL group. With respect to overuse gain scores, as Table 6.8 demonstrates, no significant differences were found among the three groups.

As for overall gain scores, as Table 6.9 demonstrates, significant differences in the range of small were again identified between the +WL group and the control group regarding both the pretest-posttest and pretest-delayed posttest gain scores. What is noteworthy is that a significant difference emerged between the +WL group and the –

WL group regarding the long-term gain score with a small effect size. No significant differences were observed between the –WL group and the control group.

Table 6.7
Mann-Whitney Tests on the Gain Scores of AU

	gain scores	*z*	*p*	*d*
+WL vs −WL	pre-post	-1.242	.214	.314
	pre-delayed post	-1.734	.083	.445
+WL vs CTR	pre-post	-2.038	.042	.594*
	pre-delayed post	-2.053	.040	.599*
−WL vs CTR	pre-post	-1.444	.149	.421
	pre-delayed post	-.865	.387	.250

Note. AU: accurate use, *$p < .05$.

Table 6.8
Mann-Whitney Tests on the Gain Scores of OU

	gain scores	*z*	*p*	*d*
+WL vs. −WL	pre-post	-.076	.939	.002
	pre-delayed post	-.288	.773	.072
+WL vs. CTR	pre-post	-.574	.566	.161
	pre-delayed post	-.618	.508	.187
−WL vs. CTR	pre-post	-.665	.536	.177
	pre-delayed post	-.535	.593	.152

Note. OU: overuse.

Table 6.9
Mann-Whitney Tests on the Gain Scores of OA

	gain scores	*z*	*p*	*d*
+WL vs. −WL	pre-post	-1.713	.087	.438
	pre-delayed post	-2.008	.045	.519*
+WL vs. CTR	pre-post	-1.988	.047	.579*
	pre-delayed post	-2.324	.020	.687*
−WL vs. CTR	pre-post	-.679	.497	.195
	pre-delayed post	-.687	.492	.197

Note. OA: overall, *$p < .05$.

Taken together, the +WL participants achieved significantly higher gain scores than the –WL group in the long term with respect to overall scores, indicating the facilitative impact of WL on L2 learning. Moreover, the +WL group outperformed the control group in both the short and the long term for both accurate use and overall scores, excluding the possibility of test repetition effects. It should be noted, however, that no significant differences were observed regarding overuse scores. Also, no statistically significant differences were identified between the –WL group and the control group here, either.

6.1.3 Recognition Tests

Table 6.10 presents descriptive statistics for the results of the recognition tests of the three groups. The results are also presented graphically in Figure 6.6. Compared to the grammar production tests, the medians of all the groups were already rather high on the pretest (10.00 for the two treatment groups and 9.00 for the control group) and with much smaller interquartile ranges than those of the grammar production tests, ranging from 2 to 4. Furthermore, the two treatment groups increased their scores on the posttest and delayed posttest in similar trajectories, increasing their scores over the three time periods, somewhat markedly from pretest to posttest, but only slightly from posttest to delayed posttest. Although the control group also achieved gains on the posttest, their score remained the same on the delayed posttest as on the posttest.

Table 6.10
Descriptive Statistics for the Recognition Tests for the Three Groups

		Pretest			Posttest			Delayed Posttest		
	N	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI	*Mdn*	*IQR*	95% CI
+WL	33	10.00	3	[9.00, 10.00]	13.00	7	[9.00, 15.00]	14.00	7	[10.50, 16.00]
−WL	31	10.00	4	[8.00, 11.00]	12.00	8	[9.00, 15.00]	13.00	8	[9.00, 16.00]
CTR	18	9.00	2	[8.50, 10.00]	11.00	2	[10.00, 11.00]	11.00	3	[9.00, 12.00]

Note: maximum score = 16.

Figure 6.6
Medians of Recognition Tests

Following the procedures used for the essay and grammar production tests, a series of Mann-Whitney tests were conducted to ensure that there were no significant differences among the three groups at the outset of this experiment. As presented in Table 6.11, below, no statistically significant differences were detected among them.

Table 6.11
Mann-Whitney Tests on the Recognition Pretest Scores

	z	*p*	*d*
+WL vs −WL	-.503	.615	.126
+WL vs CTR	-.631	.528	.179
−WL vs CTR	-.754	.451	.217

Therefore, in order to determine whether the three groups differed in terms of their improvements on the posttests scores, as with the essay tests and grammar production tests, additional Mann-Whitney tests were conducted, combining two of the three groups. As shown in Table 6.12, the results demonstrated significant differences not only between the +WL group and the control group, but also between the –WL group and the control group this time. To be more precise, for the –WL group and the control group, statistically significant differences were identified concerning both short-

and long-term gain scores in the range of medium and small, respectively. Although the results were similar to the +WL group and the control group, the *d*-values were larger than those for the –WL group and the control group, both being in the range of medium. No significant differences were identified between the two treatment groups despite the difference in terms of opportunities to engage in WL.

Overall, in contrast to the results of the essay and grammar production tests, both the +WL and –WL groups demonstrated similar gain scores over the two time periods, resulting in no statistically significant differences between the two groups. Meanwhile, significant differences were observed between the two treatment groups and the control group, suggesting some impact of the treatment on the results. That said, the larger effect sizes of the +WL group indicate that they improved their scores considerably more than the –WL group.

Table 6.12
Mann-Whitney Tests on the Gain Scores of the Recognition Tests

	gain scores	*z*	*p*	*d*
+WL vs −WL	pre-post	-1.585	.113	.404
	pre-delayed post	-1.037	.300	.262
+WL vs CTR	pre-post	-3.017	.003	.918**
	pre-delayed post	-3.093	.006	.829**
−WL vs CTR	pre-post	-2.489	.016	.735*
	pre-delayed post	-2.507	.037	.624*

Note. *$p < .05$, **$p < .01$.

6.1.4 Summary

To summarise, the result that the +WL group outperformed the control group with respect to all the test scores seems to prove the facilitative impact of WL on learning above and beyond test repetition effects. Moreover, the result that the +WL group outperformed the –WL group on two of the three assessment tests (i.e., essay tests and

grammar production tests) seems to prove the facilitative impact of WL on learning. It is worth noting, however, that the +WL group only outperformed the –WL group on the essay tests and grammar production tests, which required the participants to produce language (productive skills), but not on the recognition test, which required the learners to process input (receptive skills).

6.2 Frequency of T-WLEs and L2 Learning (RQ2)

The total number of WLEs was 222 (*Mdn* = 6.00, *IQR* = 7), of which 49 (22.1%) were categorised as T-WLEs (*Mdn* = 1.00, *IQR* = 1) and the rest (173, 77.9%) as NT-WLEs (*Mdn* = 5.00, *IQR* = 6). As stated earlier, only T-WLEs were examined further in this thesis. Spearman's correlation analyses were run in order to investigate the relationships between the frequency of T-WLEs and learning of the target construction measured by the gain scores on the three assessments (i.e., essay tests, grammar production tests and recognition tests).

The three tables below present the results of the correlation analyses for the essay tests (Table 6.13), grammar production tests (Table 6.14), and recognition tests (Table 6.15). As shown in Table 6.13, on the one hand, none of the correlations between the frequency of T-WLEs and essay test gains turned out to be statistically significant. On the other hand, as demonstrated in Tables 6.14 and 6.15, a single statistically significant correlation was identified for each grammar test. For the grammar production test, a significant correlation with a small magnitude was identified with a short-term overall gain score (OAG1) (Table 6.14). Similarly, as for the recognition tests, a statistically significant correlation was also observed with short-term gain (G1), in the range of medium (Table 6.15).

To sum up, statistically significant correlations between the frequency of T-WLEs

and L2 learning were identified for the grammar production tests and recognition tests. No such correlation, however, was detected for the essay tests. These results may suggest that focusing on the target construction and writing about it frequently contributed to learning, as for the two grammar tests, at least in the short term. In contrast, with respect to the essay tests, the frequency with which the participants focused on the target construction did not predict their gains in either the short or long term.

Table 6.13
Correlations between Frequency of T-WLEs and Essay Test Gains

	OCG1	PCG1	OCG2	PCG2
r	.113	-.155	.237	.073
95% CI	[-.236, .437]	[-.483, .191]	[-.090, .528]	[-.312, .432]
p	.532	.390	.185	.688

Note. OC: obligatory context, PC: points per context, G1: pre-post gain, G2: pre-delayed post gain.

Table 6.14
Correlations between Frequency of T-WLEs and Grammar Production Test Gains

	AUG1	OUG1	OAG1	AUG2	OUG2	OAG2
r	.089	-.221	.357*	.109	-.222	.328
95% CI	[-.273, .438]	[-.508, .150]	[.026, .636]	[-.258, .441]	[-.535, .130]	[-.052, .654]
p	.623	.217	.041	.548	.215	.062

Note. AU: accurate use, OU: overuse, OA: overall, G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$.

Table 6.15
Correlations between Frequency of T-WLEs and Recognition Test Gains

	G1	G2
r	.405*	.315
95% CI	[.059, .675]	[-.045, .585]
p	.019	.075

Note. G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$.

6.3 Quality of T-WLEs and L2 Learning (RQ3)

As stated in the previous section, the +WL participants were divided into four groups, first depending on the presence or absence of T-WLEs, then on the level of awareness regarding the target construction as demonstrated in their T-WLEs. Of the 33 +WL participants, five were categorised as 0 (no T-WLEs), 14 as 1 (noticing), eight as 2 (partial understanding) and six as 3 (understanding). Next, as the categories were rank-order data, they were submitted to Spearman's correlation analyses to investigate a possible association between the quality of T-WLEs and L2 learning (measured by the gain scores on the three assessment tests). Below are the results of the correlation analyses between the quality of T-WLEs and the gain scores on the essay tests (Table 6.16), grammar production tests (Table 6.17) and recognition tests (Table 6.18).

As shown in Table 6.16, no statistically significant correlations were identified between the quality of T-WLEs and gain scores on the essay tests. In contrast, the analyses detected statistically significant correlations for the two grammar tests. As presented in Table 6.17, analyses with the grammar production tests demonstrated statistically significant correlations concerning the pretest-posttest as well as pretest-delayed posttest overall gain scores (OAG1 and OAG2), both in the range of medium. No statistically significant correlations were identified in terms of accurate use or overuse gain scores. Finally, as for the recognition tests, as presented in Table 6.18, a statistically significant correlation with a medium effect size was again observed, but it was with a long-term gain score (G2) this time.

To summarise, for the grammar production tests, significant correlations emerged with overall gain scores in both the short and long term with medium-sized effects, respectively. Similarly, as for the recognition tests, a significant correlation was also identified with a medium-sized effect, but only with the long-term gain score. In

contrast to the two grammar tests, the essay tests did not demonstrate any statistically significant correlations with the quality of T-WLEs. These results indicate that the higher the level of the participants' awareness became, the greater the gain scores they achieved, especially in the long term, on the two grammar tests, but not the essay tests.

Table 6.16
Correlations between Quality of T-WLEs and Essay Test Gains

	OCG1	PCG1	OCG2	PCG2
r	-.228	-.209	.188	.090
95% CI	[-.540, .110]	[-.553, .162]	[-.125, .487]	[-.245, .424]
p	.203	.242	.295	.618

Note. OC: obligatory context, PC: points per context, G1: pre-post gain, G2: pre-delayed post gain.

Table 6.17
Correlations between Quality of T-WLEs and Grammar Production Test Gains

	AUG1	OUG1	OAG1	AUG2	OUG2	OAG2
r	.224	-.190	.409*	.333	-.253	.504**
95% CI	[-.194, .399]	[-.614, .013]	[.086, .645]	[-.094, .518]	[-.628, .031]	[-.141, .735]
p	.209	.290	.018	.058	.156	.003

Note. AU: accurate use, OU: overuse, OA: overall, G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$, **$p < .01$.

Table 6.18
Correlations between Quality of T-WLEs and Recognition Test Gains

	G1	G2
r	.295	.400*
95% CI	[-.047, .583]	[.132, .641]
p	.096	.021

Note. G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$.

6. 4 Correlation between Frequency and Quality of T-WLEs (RQ4)

RQ4 asked the extent to which the frequency and quality of T-WLEs were related. In order to investigate this relationship, a Spearman's correlation analysis was

conducted, which produced the result presented in Table 6.19, below. As expected, the two variables demonstrated a statistically significant correlation. What is noteworthy is that the size of the correlation turned out to be in the range of large, showing the strongest correlation of the analyses conducted regarding the frequency (RQ2) and quality (RQ3) of T-WLEs.

Table 6.19
Correlation between Frequency and Quality of T-WLEs

		Frequency
Quality	*r*	.654***
	95% CI	[.360, .875]
	p	<.001

Note. ***$p < .001$.

In summary, the result suggests that those participants who produced higher numbers of T-WLEs showed higher levels of awareness and/or understanding of the construction. In other words, the more frequently the participants focused on the target construction and produced T-WLEs, the higher the level of awareness and/or understanding they achieved.

6. 5 Aptitude and WL (RQ5)

RQ5 addressed the association between the participants' language aptitude, measured by three aptitude tests (i.e., the MLAT, LABJ and LLAMA), and L2 learning. First of all, descriptive statistics were calculated for the performance of all participants in the two treatment groups on the three measures of aptitude, which are summarised in Table 6.20, below.

With respect to the MLAT, the average was 15.84 with a range of 11. Two participants obtained a perfect score (i.e., 20), while one participant scored nine, which was the lowest score. These results are similar to those of the pilot conducted with

learners comparable to the participants in the main study (as reported, the average was 16.2 with a range of nine). Cronbach's alpha was .758, slightly higher than that of the pilot test (.732).

As for the LABJ, the averages in the study by Shintani and R. Ellis (2015), who also employed LABJ Part 2 with Japanese university students, ranged from 13.57 to 14.23 out of 15 across four treatment groups with much smaller standard deviations than the one observed in the current study (3.83), ranging from .819 to 1.44. Compared to their results, the average produced here was lower (9.78) and with more variance, with scores ranging from 3 to 15.

Finally, with respect to the LLAMA, according to its manual (Meara, 2005), "scores should be interpreted as follows:

0–15 a very poor score, probably due to guessing

20–45 an average score; most people score within this range

50–65 a good score

75–100 an outstandingly good score. Few people manage to score in this range." (p. 18)

Given this, the average found in the present thesis, 59.06, may be interpreted as "a good score." It should, however, be pointed out that the participants' scores varied greatly, ranging from 10 to 100, as can be seen from the rather large standard deviations (24.28). In the aforementioned validation study to assess the reliability of the LLAMA by Granena (2013), the average score on the LLAMA was 56.67 (SD = 24.09) (p. 117).

Then, a series of independent samples t-tests were run on the scores of the three aptitude tests in order to detect any initial group differences. None of the t-tests revealed any statistically significant differences (the MLAT: $t(62) = –.382$, $p = .704$, the LABJ: $t(62) = –.440$, $p = .662$, and the LLAMA: $t(62) = –.502$, $p = .617$).

Table 6.20
Descriptive Statistics for the Aptitude Tests

	N	*M*	*SD*	95%CI
MLAT (max. score: 20)	64	15.84	2.50	[15.22, 16.45]
+WL	33	15.73	2.50	[14.91, 16.58]
–WL	31	15.97	2.54	[15.06, 16.77]
LABJ (max. score: 15)	64	9.78	3.83	[8.81, 10.67]
+WL	33	9.58	3.98	[8.30, 10.85]
–WL	31	9.98	3.72	[8.65, 11.26]
LLAMA (max. score:100)	64	59.06	24.28	[52.82, 64.53]
+WL	33	57.58	25.62	[49.09, 66.36]
–WL	31	60.65	23.09	[52.26, 68.71]

6.5.1 Correlations among the Three Aptitude Tests' Results

Prior to the investigation of the relationship between the results of the three aptitude tests (MLAT, LABJ and LLAMA) and L2 learning, the correlations among the results of the three aptitude tests were investigated. As reported earlier, the normality of the distributions was not confirmed (i.e., one of the skewness and kurtosis ratios was outside the [–2, 2] interval) in terms of the results of the three aptitude tests (see Appendix M). Therefore, non-parametric tests, i.e., Spearman's correlation analyses, were run in order to investigate the possible associations among the participants' scores on the three aptitude tests.

As shown in Table 6.21, statistically significant correlations in the size of small to medium were identified among all the tests. Of the three, the correlation between the LLAMA and the LABJ turned out to be strongest, with the largest effect in the range of medium (r = .498). That said, given that the LLAMA and the LABJ are designed to measure the same construct (i.e., inductive language learning ability), the observed correlation was not as strong as expected. The result suggests that what the two tests measured correlated closely, but was not identical.

Meanwhile, although not as strong as the correlation found between the LLAMA and the LABJ, statistically significant correlations were found between the MLAT, which is designed to measure learners' grammatical sensitivity, and the two tests with effect sizes in the range of small (for the LLAMA) and medium (for the LABJ). The results seem to confirm that these two constructs (i.e., inductive language learning ability and grammatical sensitivity) are also closely related, indicating the validity of Skehan (1998) collapsing the two constructs into one, i.e., language analytic ability.

Table 6.21
Correlations among the Aptitude Tests

		MLAT	LABJ
	r	.425***	
LABJ	95% CI	[.198, .625]	
	p	<.001	
	r	.339**	.498***
LLAMA	95% CI	[.133, .530]	[.267, .678]
	p	.006	<.001

Note. $N = 64$, $^{**}p < .01$, $^{***}p < .001$.

6.5.2 Correlations with Three Assessment Tests

In order to investigate the relationship between the participants' aptitude (inductive language learning ability, grammatical sensitivity) and the effect of WL on their L2 learning, Spearman's correlation analyses were run between the results of the aptitude tests (MLAT, LABJ, and LLAMA) and the gain scores on the three assessment tests (essay tests, grammar production tests, recognition tests) (run separately). The results are reported by test, below.

6.5.2.1 Correlations with Essay Tests

As shown in Table 6.22, the only significant correlation identified with the essay

test gain scores was for the –WL participants between grammatical sensitivity (MLAT) and the short-term gain score on obligatory contexts (OCG1), which was negative and in the range of medium. No statistically significant correlations were identified with either inductive language learning ability (LLAMA and LABJ) or grammatical sensitivity (MLAT) for the +WL participants.

Table 6.22
Correlations between Aptitude and the Essay Test Gains

		OCG1	PCG1	OCG2	PCG2
+WL (*n* = 33)					
MLAT	*r*	.120	-.197	-.020	.070
	95% CI	[-.265, .458]	[-.570, .187]	[-.380, .326]	[-.364, .450]
	p	.506	.273	.913	.699
LABJ	*r*	-.170	.111	-.021	-.021
	95% CI	[-.456, .174]	[-.284, .488]	[-.300, .261]	[-.379, .319]
	p	.343	.540	.907	.907
LLAMA	*r*	-.094	-.103	.151	-.054
	95% CI	[-.450, .290]	[-.460, .316]	[-.259, .467]	[-.411, .326]
	p	.602	.568	.401	.765
–WL (*n* = 31)					
MLAT	*r*	-.420*	.004	.372	.320
	95% CI	[-.656, -.103]	[-.346, .348]	[-.467, .306]	[-.252, .476]
	p	.019	.982	.632	.568
LABJ	*r*	-.227	.212	.006	-.074
	95% CI	[-.510, .152]	[-.169, .549]	[-.259, .372]	[-.476, .320]
	p	.218	.252	.973	.692
LLAMA	*r*	-.197	.254	-.336	.184
	95% CI	[-.525, .180]	[-.084 .555]	[-.646, .021]	[-.196, .543]
	p	.288	.168	.065	.322

Note. OC: obligatory contexts, PC: points per context, G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$.

6.5.2.2 Correlations with Grammar Production Tests

In contrast to the results of the essay tests, as Table 6.23 demonstrates, the

analyses with the grammar production tests detected small to medium statistically significant correlations for both groups. It should be pointed out, however, that most of them were for the –WL group.

Table 6.23
Correlations between Aptitude and the Grammar Production Test Gains

	AUG1	OUG1	OAG1	AUG2	OUG2	OAG2
+WL (*n* = 33)						
MLAT						
r	.054	.011	.150	.297	-.172	.416*
95% CI	[-.349, .421]	[-.366, .400]	[-.191, .458]	[-.068, .606]	[-.500, .177]	[.106, .681]
p	.767	.950	.405	.093	.339	.016
LABJ						
r	.215	.083	.303	.254	-.125	.306
95% CI	[-.172, .523]	[-.289, .446]	[-.034, .574]	[-.103, .547]	[-.469, .241]	[-.052, .565]
p	.228	.646	.086	.154	.490	.083
LLAMA						
r	-.234	-.116	-.097	-.113	-.145	-.030
95% CI	[-.550, .107]	[-.442, .255]	[-.414, .257]	[-.469, .273]	[-.463, .221]	[-.370, .327]
p	.190	.520	.590	.532	.422	.868
–WL (*n* = 31)						
MLAT						
r	.071	-.015	.121	.366*	.187	.383*
95% CI	[-.290, .395]	[-.370, .337]	[-.216, .417]	[.-.028, .663]	[-.185, .487]	[.055, .653]
p	.705	.935	.515	.043	.313	.034
LABJ						
r	.384*	.301	.256	.356*	.362*	.321
95% CI	[.038, .678]	[-.111, .647]	[-.099, .566]	[-.045, .681]	[-.048, .670]	[-.115, .700]
p	.033	.100	.165	.049	.046	.078
LLAMA						
r	.349	.196	.280	.481**	.345	.455*
95% CI	[-.016, .617]	[-.148, .494]	[-.080, .586]	[.108, .764]	[-.074, .657]	[.076, .756]
p	.054	.290	.127	.006	.058	.010

Note. AU: accurate use, OU: overuse, OA: overall, G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$, **$p < .01$.

For the +WL group, there was only one statistically significant correlation between the long-term overall gain score (OAG2) and grammatical sensitivity (MLAT) with a medium effect size. No significant correlations were identified regarding inductive language learning ability (LLAMA or LABJ).

Meanwhile, as for the –WL group, the analyses revealed seven significant correlations, notably with long-term gains. More specifically, all the scores on the three aptitude tests and the long-term accurate use and overall gain scores (i.e., AUG2, OAG2), except for one (LABJ and OAG2), revealed statistically significant correlations ranging from small to medium effect sizes. Moreover, a significant correlation was observed between the LABJ and the long-term overuse gain score (OUG2). What is noteworthy is that the correlations between the long-term overall gain score (OAG2), the most reliable indicator of learning of the three scores, and two of the aptitude test results (LLAMA, MLAT) were statistically significant. In addition to these long-term correlations, the one between the short-term accurate use gain score (AUG1) and the LABJ was found to be statistically significant as well in the range of small.

It is worth noting that all the statistically significant correlations observed for both the +WL and –WL groups were with long-term gain scores except for the one found for the –WL group, suggesting that the participants, irrespective of WL experience, generally needed higher aptitude to achieve gains in the long term.

6.5.2.3 Correlations with Recognition Tests

As presented in Table 6.24, statistically significant correlations were identified only for the –WL group with respect to the recognition tests. To be more precise, the –WL group revealed three statistically significant correlations, between grammatical sensitivity (MLAT) and the short-term gain score (G1) as well as between inductive

language learning ability (LABJ and LLAMA) and the long-term gain score (G2), all with medium effect sizes.

Table 6.24
Correlations between Aptitude and the Recognition Test Gains

		G1	G2
+WL (*n* = 33)			
MLAT	*r*	.255	.204
	95% CI	[-.344, .295]	[-.468, .125]
	p	.152	.256
LABJ	*r*	.107	.085
	95% CI	[-.279, .495]	[-.316, .425]
	p	.552	.639
LLAMA	*r*	-.017	-.156
	95% CI	[-.108, .562]	[-.164, .483]
	p	.926	.385
–WL (*n* = 31)			
MLAT	*r*	.571**	.251
	95% CI	[-.078, .548]	[.238, .724]
	p	.001	.174
LABJ	*r*	.048	.455*
	95% CI	[-.317, .421]	[.121, .711]
	p	.796	.010
LLAMA	*r*	.264	.516**
	95% CI	[-.081, .554]	[.241, .715]
	p	.151	.003

Note. G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$, **$p < .01$.

6.5.3 Summary

In summary, the correlation analyses between language aptitude and L2 learning detected stronger and a higher number of significant correlations for the –WL group than the +WL group. That said, like the results for the correlation analyses concerning the frequency and quality of T-WLEs, the results differed considerably from test to test.

As for the essay test, the correlation analyses produced similar results for the two

groups. Namely, the only significant correlation was identified for the –WL group between grammatical sensitivity (MLAT) and the short-term gain in obligatory contexts (OCG1), suggesting that higher aptitude had almost no impact on achieving gains on the essay tests for the participants, regardless of their WL condition.

In contrast, with respect to the grammar production tests, more significant correlations emerged, but they were mainly for the –WL group. That is, although one significant correlation with medium effect size was found for the +WL group, between grammatical sensitivity and long-term overall gain score (OAG2), seven significant correlations were observed for the –WL group, two of them being medium effect sizes, while the rest being in the range of small. Also, it is worth mentioning that six of them were with long-term gain scores. As for the recognition tests, three statistically significant medium-sized correlations were observed only for the –WL group.

These results indicate that, in general, language aptitude played a facilitative role for the –WL group to a greater extent compared to the +WL group. Put differently, participants with higher aptitude did better under the –WL condition, i.e., when they did not engage in WL, whereas aptitude had little influence on the extent of learner gains under the +WL condition.

6.6 Metalanguage Knowledge and WL (RQ6)

Following the steps taken for the aptitude tests, descriptive statistics were first calculated concerning the metalanguage knowledge test. As reported in Table 6.25, below, the average was 11.55 out of 17, which is a little higher than that of the pilot test (10.7) conducted with learners similar to the participants in the current study. It should be noted, however, that the participants' scores varied rather noticeably, ranging from 3 to 17 (range: 14), which is reflected in the fairly large standard deviation (3.75).

Cronbach's alpha was .821, again a little higher than that of the pilot test (.806).

Table 6.25
Descriptive Statistics for the Metalanguage Knowledge Test

	N	*M*	*SD*	95% CI
	64	11.55	3.75	[10.61, 12.45]
+WL	33	11.52	3.64	[10.27, 12.67]
–WL	31	11.58	3.93	[10.23, 12.87]

Note. maximum score = 17.

6.6.1 Correlations with Three Assessment Tests

First of all, an independent samples *t*-test was administered, to make sure that the metalanguage knowledge test scores of the two groups were comparable, which revealed no statistically significant difference, $t(62) = -.069$, $p = .945$. Then, in order to address RQ6, the relationship between the participants' metalanguage knowledge and the effect of WL on their L2 learning was examined. As reported earlier, the normality of the distribution was confirmed in terms of the results of the metalanguage knowledge test. In order to be consistent with the other correlational analyses, however, non-parametric tests, i.e., Spearman's correlation analyses, were again run between the results of the metalanguage knowledge test and the gain scores on the three assessment tests (i.e., essay test, grammar production test, production test) (run separately). The results appear in Tables 6.26–6.28, below.

As shown in Table 6.26, for the essay tests, no statistically significant correlations were observed for either the +WL group or the –WL group. Meanwhile, as Table 6.27 demonstrates, in terms of the grammar production tests, significant correlations were identified for both groups. It is important to note, however, that a higher number of significant correlations was detected for the –WL group with generally larger effect sizes. In fact, the –WL group demonstrated four significant correlations with accurate

use and overall gain scores in both the short and long term, all medium size. Meanwhile, the +WL group showed significant correlations only with two long-term gains (i.e., overuse and overall with medium and small effect sizes, respectively). Finally, as presented in Table 6.28, with respect to the recognition tests, similar results were observed for both groups. That is, a statistically significant correlation was identified for the +WL group (with the long-term gain score) and the –WL group (with the short-term gain score), both with medium effect sizes.

Table 6.26
Correlations between Metalanguage Knowledge and the Essay Test Gains

	OCG1	PCG1	OCG2	PCG2
+WL (*n* = 33)				
r	.097	-.057	-.021	.342
95% CI	[-.293, .453]	[-.474, .327]	[-.312, .295]	[.032, .597]
p	.591	.753	.907	.052
–WL (*n* = 31)				
r	-.204	.193	-.041	.280
95% CI	[-.567, .207]	[-.185, .546]	[-.427, .360]	[-.078, .575]
p	.271	.297	.828	.128

Note. OC: obligatory contexts, PC: points per context, G1: pre-post gain G2: pre-delayed post gain.

Table 6.27
Correlations between Metalanguage Knowledge and the Grammar Production Test Gains

	AUG1	OUG1	OAG1	AUG2	OUG2	OAG2
+WL (*n* = 33)						
r	-.126	-.273	.083	.116	-.407*	.381*
95% CI	[-.491, .258]	[-.592, .102]	[-.276, .418]	[-.253, .444]	[-.676, -.072]	[-.033, .656]
p	.485	.124	.646	.522	.019	.029
–WL (*n* = 31)						
r	.427*	-.012	.429*	.474**	.177	.470**
95% CI	[.032, .747]	[-.311, .316]	[.092, .687]	[-.074, .755]	[-.200, .536]	[.151, .703]
p	.017	.949	.016	.007	.340	.008

Note. AU: accurate use, OU: overuse, OA: overall, G1: pre-post gain, G2: pre-delayed post gain, *$p < .05$, **$p < .01$.

Table 6.28
Correlations between Metalanguage Knowledge and the Recognition Test Gains

	G1	G2
+WL (*n* = 33)		
r	.250	.564**
95% CI	[-.150, .607]	[.278, .761]
p	.160	.001
–WL (*n* = 31)		
r	.529**	.287
95% CI	[.244, .730]	[-.077 .640]
p	.002	.117

Note. G1: pre-post gain, G2: pre-delayed post gain, **$p < .01$.

6.6.2 Summary

To summarise, like the previous analyses regarding aptitude, the results obtained here varied from test to test. The essay tests revealed somewhat different results from the two grammar tests, showing no statistically significant correlations for either of the groups, indicating that higher metalanguage knowledge had no significant influence on the extent of gains on the essay tests with or without WL. Meanwhile, with respect to the recognition tests, the two groups produced similar results, showing a single statistically significant correlation each in the range of medium, for short term (–WL group) and long term (+WL group). In contrast to the two tests, the most noticeable difference was again observed in the grammar production tests, with a higher number of, and generally stronger, statistically significant correlations for the –WL group.

These results indicate that metalanguage knowledge played a greater role for the –WL group than for the +WL group in achieving gains on the grammar production tests. In other words, the participants with higher metalanguage knowledge performed better under the –WL condition, whereas knowledge of metalanguage had less impact on participants' gains when they engaged in WL, echoing the results of the correlation

analyses for language aptitude.

6. 7 Frequency of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge (RQ7)

RQ7 investigated the frequency of T-WLEs in relation to learners' individual differences in language aptitude and metalanguage knowledge in order to identify if producing T-WLEs to a higher frequency is related to higher aptitude and/or metalanguage knowledge. Prior to the investigation, the relationship between aptitude and metalanguage knowledge was examined. As stated earlier, the frequency of T-WLEs is interval data with normal distribution. However, distributions deviated from normality for the aptitude test results. Spearman's correlations analyses were, therefore, run again between the results of the metalanguage knowledge test and three aptitude tests.

As presented in Table 6.29, statistically significant correlations were observed between the results of the metalanguage knowledge test and those of all the aptitude tests. What is noteworthy is that the correlation with the MLAT (a measure of grammatical sensitivity) was identified to be the strongest, being in the range of large. Meanwhile, the LABJ and LLAMA demonstrated similar correlations, both in the range of small, which is not surprising given that the two tests are designed to measure the same construct, inductive language learning ability.

Table 6.29
Correlations between Metalanguage Knowledge and Aptitude

	MLAT	LABJ	LLAMA
r	.637***	.373**	.350**
95% CI	[.459, .779]	[.125, .576]	[.113, .545]
p	<.001	.002	.005

Note. $N = 64$, $^{**}p < .01$, $^{***}p < .001$.

Then, in order to investigate the relationship between the frequency of T-WLEs and learners' aptitude and metalanguage knowledge, Spearman's correlation analyses were again performed. As shown in Table 6.30, two statistically significant correlations emerged with the MLAT and the metalanguage knowledge test in the range of medium and small, respectively. Meanwhile, the LABJ and LLAMA, measures of inductive language learning ability, showed no significant correlations.

Table 6.30
Correlations between Frequency of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge

	MLAT	LABJ	LLAMA	Metalanguage Knowledge
r	.464**	.319	.152	.356*
95% CI	[.140, .718]	[-.051, .617]	[-.202, .462]	[-.018, .661]
p	.007	.070	.397	.042

Note. $N = 33$, $*p < .05$, $**p < .01$.

These results mean that learners with higher grammatical sensitivity and metalanguage knowledge produced a higher number of T-WLEs, as hypothesized. They also suggest that inductive language learning ability was not related to the number of T-WLEs participants produced. Put differently, metalanguage knowledge and grammatical sensitivity predicted the learners' frequency of T-WLEs, but not inductive language learning ability.

6.8 Quality of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge (RQ8)

RQ8 examined how the quality of T-WLEs was related to learners' language aptitude and metalanguage knowledge in order to identify if producing T-WLEs with higher quality might be due to higher aptitude and/or metalanguage knowledge. As

mentioned earlier, the quality of T-WLEs is ordinal in nature. In addition, the distribution of the aptitude tests results was skewed. Therefore, Spearman's correlations analyses were again conducted between the quality of T-WLEs and the results of the aptitude tests and the metalanguage knowledge test.

As shown in Table 6.31, again, a statistically significant correlation with a medium effect size was detected with the MLAT. It should be pointed out, however, that it was the only significant correlation identified. The correlation with metalanguage knowledge did not turn out to be significant this time. Also, although the quality of T-WLEs was hypothesized to be related to inductive language learning ability, neither the LABJ nor LLAMA (two measures of this ability) showed a significant correlation.

Table 6.31
Correlations between Quality of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge

	MLAT	LABJ	LLAMA	Metalanguage Knowledge
r	.475**	.260	.025	.324
95% CI	[.176, .711]	[-.138, .592]	[-.366, .366]	[-.067, .656]
p	.005	.145	.891	.066

Note. $N = 33$, $^{**}p < .01$.

These results mean that the participants who had higher grammatical sensitivity were likely to produce higher quality T-WLEs, but differences in inductive language learning ability or metalanguage knowledge did not influence the quality of T-WLEs.

6.9 Exit Questionnaire Results

In order to obtain retrospective information about the participants' subjective perspective concerning this experiment as well as to complement the data obtained from the test results, an exit questionnaire was administered immediately after the delayed posttest to the participants in the two treatment groups. As briefly explained earlier, it

consisted of two parts; the first part was for the participants in both groups and contained five questions regarding the experiment in order to identify their perceptions of the experiment in different treatment conditions (i.e., with vs without WL). The second part, consisting of four questions, was only for the +WL participants in order to investigate their subjective perspective in terms of their experience of WL.

As for Part I, in order to make not only cross-group but also within-group comparisons possible, a percentage was calculated for each response category for each group by dividing the number of participants who provided a particular response by the number of the participants in the group. As for Part II, a percentage was calculated in the same way but only for a within-group comparison. The participants were allowed to give multiple answers for each question in both parts. The percentages in the tables, therefore, do not necessarily add up to 100% in each column. All the responses, regardless of numbers, are reported below. This phase was not timed and the participants who finished the questionnaire early were instructed to read their course textbooks. The results are presented by question below.

6.9.1 Part I: Reflecting on the Experiment (Both Groups)

6.9.1.1 The Purpose of the Experiment

The first question in this part sought to obtain information about the participants' perceptions of the purpose of the experiment. The question was worded as follows: "What do you think was the purpose of the experiment? (Besides just doing research on language learning, was there something specific you thought I might be studying?)" As shown in Table 6.32, the three most commonly given responses in both groups were to examine the participants' "English grammar knowledge," "English ability," and "essay writing ability," besides "don't know." It should be pointed out, however, that a higher

percentage of +WL participants responded with not only “grammar” (21.2%) but also “the target construction” (9.1%) (30.3% in total) compared to the –WL participants (16.1% for grammar, 3.2% for the target construction, 19.3 % in total). Another difference is that none of the +WL participants left the space blank and wrote something even if it was “I don’t know,” whereas two –WL participants did so. It is important to note that one participant in the +WL group wrote “the importance of writing?” in addition to “English grammar knowledge.” She did not make it clear, however, what she meant by “writing” (i.e., if she meant essay writing or writing about her thoughts (i.e., WL)).

Table 6.32
Purpose of the Experiment

	+WL (*n* = 33)	–WL (*n* = 31)
The purpose of the experiment was to examine...		
- English grammar knowledge	7 (21.2%)	5 (16.1%)
- English ability	6 (18.2%)	8 (25.8%)
- essay writing ability	6 (18.2%)	6 (19.4%)
- target construction	3 (9.1%)	1 (3.2%)
- attitude, motivation	2 (6.1%)	4 (12.9%)
- aptitude	1 (3.0%)	1 (3.2%)
- repetition effect	1 (3.0%)	1 (3.2%)
- concentration	1 (3.0%)	0 (0.0%)
- creativity/imagination	1 (3.0%)	1 (3.2%)
- importance of writing	1 (3.0%)	0 (0.0%)
don’t know	5 (15.2%)	5 (16.1%)
no answer	0 (0.0%)	2 (6.5%)

6.9.1.2 Perceptions of Learning during the Experiment

In order to obtain information regarding the participants’ perceptions of their learning in general, if any, throughout this experiment, the next question asked: “Do you think that you learned something about English during the 3-week experiment? If yes,

what?"

As summarized in Table 6.33, the most popular response in both groups was "essay writing." Around one third of the participants in each group, nine (27.3%) for the +WL group and 11 (35.5%) for the –WL group, responded that they acquired essay writing ability/skills over the three weeks. As stated earlier, the participants were in English courses whose main focus was the TOEIC test, which consists of only multiple-choice questions. Therefore, they were not used to writing in English, let alone English essays, and many of them found writing essays challenging. Some of them, however, also found it interesting, as reported below under Question 5.

In addition, similar numbers of the participants (five for the +WL group and four for the –WL group) answered "grammar." However, there was a rather striking difference concerning the target construction. That is, a much higher percentage of the +WL participants (24.2%) reported that they learnt the target construction than the –WL participants (9.7%). In contrast, twice as many –WL participants (six) as the +WL participants (three) answered "none." It should be pointed out, however, that the target construction in the current study (i.e., the present counterfactual conditional) is usually taught in junior high school and the participants were supposed to be familiar with it (i.e., to have at least some declarative knowledge of it, even if they did not have procedural knowledge of it). Therefore, the participants of higher proficiency were not likely to have perceived this construction as well as other forms covered in the experiment new to them. Supporting this speculation, one of the participants whose response was coded as "none" wrote "I don't think I learned anything new, but I found it good practice."

Table 6.33
Perceptions of Learning

	+WL (*n* = 33)	–WL (*n* = 31)
I learned…		
- essay writing	9 (27.3%)	11(35.5%)
- target construction	8 (24.2%)	3 (9.7%)
- grammar	5 (15.2%)	4 (12.9%)
- vocabulary	3 (9.1%)	1 (3.2%)
- none	3 (9.1%)	6 (19.4%)
- tense	2 (6.1%)	1 (3.2%)
- the importance of vocabulary	2 (6.1%)	0 (0%)
- the level of my English ability	1 (3.0%)	3 (9.7%)
- the difficulty of English	1 (3.0%)	2 (6.5%)
no answer	3 (9.1%)	4 (12.9%)

6.9.1.3 Perceptions of Learning in More Detail

In order to obtain more detailed information regarding the participants' perceptions of learning, the next question asked: "Related to Q2, are there any vocabulary items or grammar rules you learned? If so, please write them below in detail." As summarized in Table 6.34, again a higher percentage of the +WL participants (33.3%) responded "the target construction" than their –WL counterparts (16.1%). Moreover, it is worth noting that four +WL participants (12.1%) wrote the grammar rule of the target construction, such as "If + subject + verb past, subject + aux past + verb root" while only one participant (3.2%) did so in the –WL group. Therefore, if they are combined, almost half of the +WL participants (45.4%) referred to the target construction, whereas only 19.3% of the –WL participants did so, suggesting a higher ratio of awareness of the target construction. It should be pointed out, however, that the grammar rule given by one of the four +WL participants was partially incorrect. (She wrote, "use auxiliary past and verb past for a main clause.") Although three participants in each group responded "tense," it is not clear if they meant using the past tense for the

target construction or the tense in general by that term. (The term "vocabulary items" was included in the question in order not to raise the participants' consciousness only to grammar, but none of them mentioned any vocabulary items.)

Table 6.34
Perceptions of Learning in More Detail

	+WL (*n* = 33)	–WL (*n* = 31)
What I learned in this experiment was...		
- target construction	11 (33.3%)	5 (16.1%)
- rule of the target construction	4 (12.1%)	1 (3.2%)
- tense	3 (9.1%)	3 (9.7%)
- none	3 (9.1%)	5 (16.1%)
- past	2 (6.1%)	0 (0.0%)
- past perfect	1 (3.0%)	0 (0.0%)
- passive/voice	1 (3.0%)	2 (6.5%)
- partial negative	1 (3.0%)	0 (0.0%)
- comparative	0 (0.0%)	1 (3.2%)
- conjunction	0 (0.0%)	1 (3.2%)
no answer	9 (27.3%)	13 (41.9%)

6.9.1.4 Focus of the Participants during the Experiment

In order to elucidate differences in the participants' cognitive processes, if any, between the two groups, the next question asked: "If there are any vocabulary items or grammar points you focused on, please write them in detail." As shown in Table 6.35, below, echoing the difference between the two groups regarding the perceptions of learning in terms of the target construction, a higher percentage (63.6%) of the +WL participants reported that they focused on the target construction than the –WL participants (38.7%), implying that a higher ratio of the +WL participants not only felt that they learned the target construction but also focused on it. In addition, a little over one fifth of the +WL participants (21.2%) gave no answer, whereas almost half of the –

WL participants (48.4%) did so, suggesting that a higher percentage of the +WL participants attempted to focus on some form, even if it was not the target construction. (Although the question included "vocabulary items," again, none of the participants mentioned any vocabulary items.)

Table 6.35
Focus of the Participants

	+WL (*n* = 33)	–WL (*n* = 31)
I focused on…		
- target construction	21 (63.6%)	12 (38.7%)
- tense	2 (6.1%)	2 (6.5%)
- past tense	1 (3.0%)	0 (0.0%)
- present perfect	1 (3.0%)	1 (3.2%)
- past perfect	1 (3.0%)	0 (0.0%)
- comparative	1 (3.0%)	1 (3.2%)
- passive	1 (3.0%)	1 (3.2%)
- none	1 (3.0%)	2 (6.5%)
no answer	7 (21.2%)	15 (48.4%)

6.9.1.5 Free Comments on the Experiment

The last question in this part asked: "Please write any other comments you may have regarding this experience (Your comments may overlap with what you wrote above.)," in order to elicit any information the participants might have regarding the experience that the previous four questions did not elicit. As reported in Table 6.36, reflecting the results of the previous questions, the most common response of the +WL group was regarding grammar. A little over one third of the participants (36.4%) commented that they felt it necessary to study or review grammar. Prior to the experiment, the participants were told that they did not have to study for this experiment. After the delayed posttest, I asked them if they had studied and none of them said they had. It is worth noting that two of them specifically referred to the target

construction, one of them saying that he realized his understanding of the construction was not solid. In contrast, half the number of the –WL participants (19.4%) wrote about grammar and none of them mentioned the target construction specifically.

Table 6.36
Free Comments on the Experiment

	+WL (*n* = 33)	–WL (*n* = 31)
I realized that I need to study/review grammar.	12 (36.4%)	6 (19.4%)
Essay writing was difficult	7 (21.2%)	6 (19.4%)
It was a good review/exercise.	4 (12.1%)	4 (12.9%)
I realized my vocabulary is limited	3 (9.1%)	2 (6.5%)
I enjoyed writing essays/about myself.	3 (9.1%)	1 (3.2%)
The LLAMA was interesting	3 (9.1%)	1 (3.2%)
I think my essay writing skills improved.	2 (6.1%)	2 (6.5%)
The tests were hard because I could not use my dictionary.	2 (6.1%)	3 (9.7%)
The reconstruction task was difficult.	2 (6.1%)	1 (3.2%)
The three tests were similar and I got used to them.	1 (3.0%)	1 (3.2%)
no answer	2 (6.1%)	7 (22.5%)

6.9.2 Part II: Reflecting on WL (Only for the +WL group)

This part was only given to the +WL group in order to elicit their perceptions regarding their WL experience. The answers are reported by question below.

6.9.2.1 Perceptions of the Activity of WL in General

In order to elucidate how the participants felt about having engaged in WL activity during the treatment task, the first question in this section asked: "What feelings do you have about the experience of writing your thoughts and questions? Why?" As presented in Table 6.37, the majority of the comments (81.8%) demonstrated the participants' positive views toward the experience. Of these, one of the most popular responses was about the depth of thinking involved (eight, 24.2%). In addition, there were nine responses regarding noticing, of which six were about noticing the

differences between the original text and their reconstructions, and three about noticing their lack of understanding (18.2% and 9.1%, respectively, 27.3% in total).

Meanwhile, almost one fifth of the replies (18.2%) were on the difficulty of putting their thoughts in writing. It should be pointed out, however, that as the participants were allowed to give multiple answers, there were some participants who referred not only to the difficulty of describing their thoughts on paper but also to the usefulness of writing them, such as "I found it hard to write what I was thinking, but I feel it helped me to notice my mistakes." One found it embarrassing to express his inner thoughts in writing, saying "I felt shy." Finally, in spite of the WL practice conducted before the experiment, four +WL participants commented that writing their thoughts was "refreshing" as it was something "they had never done before," suggesting the WL practice was not enough. It is worth noting that all the participants responded to this question without skipping it, i.e., there were no responses which were coded as "no answer."

Table 6.37
Perceptions of WL

Response category	Number of answers
I found WL …	
useful because it helped me to…	27 (81.8%)
- think more deeply	8 (24.2%)
- organise my thoughts	6 (18.2%)
- notice my mistakes	6 (18.2%)
- review and reflect	4 (12.1%)
- notice my lack of understanding	3 (9.1%)
difficult	6 (18.2%)
refreshing	4 (12.1%)
embarrassing	1 (3.0%)
no answer	0 (0.0%)

Note. $N = 33$.

6.9.2.2 Perceptions of the Impact of WL

The next question concerned the +WL participants' perceptions of learning in relation to WL. It was worded as follows: "In the task, you wrote your thoughts while checking the original text. However, do you think there would be a difference if you hadn't written your thoughts while checking? Why?" As presented in Table 6.38, almost 90% of the participants (29) answered in the affirmative to this question. (The details of their responses are shown in Table 6.39.) Meanwhile, two participants responded "I don't know," another participant "no," saying "I don't need to write because I can think in my head without writing," and a fourth participant gave no response.

Table 6.38
Perceptions of the Differences between with or without WL

Response category	Number of participants
Yes	29 (87.9%)
Don't know	2 (6.1%)
No	1 (3.0%)
No answer	1 (3.0%)

Note. $N = 33$.

When the responses of the 29 participants were examined closely, they were divided into five main categories (Table 6.39). (The percentages in parentheses are calculated based on positive responses (29).) The most common response was "WL helped me to notice my mistakes" (10, 34.5%). Some of the participants also wrote that they probably paid more attention to the comparison of the original text with their reconstructed text in order to write their thoughts. The second common response concerned memory and was coded as "WL helped me to remember what I was thinking" (nine, 31.0%). Although nine responses were coded in this category, it is worth noting that there were two types of comments. That is, five of the nine

respondents commented that they felt WL was helpful because what they wrote was likely to stay in their brain thanks to the action of writing, while the other four found WL helpful because they could go back to their writing (i.e., the product of their writing) later when they wanted to remember what they were thinking.

The third most common responses (eight, 27.6%) concerned the participants' thinking processes, i.e., depth of thinking, which were coded under "WL helped me to think deeply." One of the respondents wrote that her thinking might have been superficial without writing, whereas she felt that she could think more deeply with WL. Similarly, the fourth most common response referred to thinking processes, which was coded under "WL helped me to organize my thoughts." Seven respondents (24.1%) mentioned that writing their thoughts helped them to organize their thinking internally and visually as they could read their writing, referring to the function of external memory. What is interesting is that four participants (13.8%) also commented on thinking. They wrote, however, that writing their thoughts forced them to think, saying that they were not likely to have thought as seriously without the requirement to do so.

Table 6.39
Reasons for Positive Perceptions of WL

Response category	Number of answers
Yes, because WL helped me to…	
- notice my mistakes	10 (34.5%)
- remember what I was thinking	9 (31.0%)
- think deeply	8 (27.6%)
- organize my thoughts	7 (24.1%)
- think	4 (13.8%)

Note. $N = 29$.

6.9.2.3 Perceptions of Usefulness of WL

The third question in this section asked: “Do you think you learned something from the experience of writing while checking? Why or why not?” As presented in Table 6.40, again almost 90% of the participants answered yes to this question. It should be pointed out, however, that these 29 participants overlapped with the 29 participants who answered yes to Q2. One of the participants who answered “I don’t know” to the previous question answered “no” this time, saying that she could not describe her thoughts well, while the other wrote “I don’t know” again. As for the two participants who gave no answer and who answered “no” to Q2, they left the space blank, giving no answer.

Table 6.40
Perceptions of Learning due to WL

Response category	Number of participants
Yes	29 (87.9%)
Don’t know	1 (3.0%)
No	1 (3.0%)
No answer	2 (6.1%)

Note. $N = 33$.

The details of the reasons given by the 29 participants were divided into four main categories, as presented in Table 6.41, below. (Again, the percentages are calculated based on positive responses (29).) The reasons are closely related to the results presented above, in Table 6.39. Most of the participants reported similar ideas given in answer to the previous question here as well. Although not a majority, there were two responses (one each, 3.4%) which are worth mentioning listed at the bottom of Table 6.41. Namely, one participant wrote that she found writing helpful, stating that “I did not write what I knew, but I wrote to understand.” (Her T-WLEs and those of the two participants who gave no answer to this question are examined closely in section

7.2.) Meanwhile, the other participant responded, “writing my thoughts might not directly lead to learning, but I feel it will help me to learn things more easily than without writing.”

Table 6.41
Perceptions of Learning due to WL in More Detail

Response category	Number of answers
I feel I learned by WL because it helped me…	
- to deepen my thoughts	11 (37.9%)
- to organize my thoughts	9 (31.0%)
- to notice my weaknesses	7 (24.1%)
- to remember my linguistic issues	7 (24.1%)
- to understand	1 (3.4%)
I feel WL will help me to learn easily	1 (3.4%)

Note. *N* = 29.

6.9.2.4 Free Comments regarding WL

Similar to the last question in Part I, the last question in this part was asked in order to elicit any information the participants might have with respect to the experience of WL that the previous three questions failed to elicit. It was worded as follow: “Please write any comments you may have regarding the experience of writing your thoughts while checking the original text.” As presented in Table 6.42, below, the majority (69.7%) of the responses were on the importance of WL, with various reasons given. In addition, the other responses were mainly positive. However, the participant who responded that engaging in WL was embarrassing to the first question reported the same but with an additional comment, “I don’t like to write my thoughts very much as it is embarrassing” (He responded “I don’t know” to both Q2 and Q3.). Three participants gave no answer.

Table 6.42
Free Comments on WL

Response category	Number of answers
Writing my thoughts is important because…	23 (69.7%)
- it makes my problems clear.	5 (15.2%)
- I can look back at my writing later.	4 (12.1%)
- it helps me remember them.	3 (9.1%)
- I can organize them in the process.	3 (9.1%)
- I can deepen my understanding.	3 (9.1%)
- it is easier than just to think in my head.	3 (9.1%)
- it enables me to learn efficiently.	2 (6.1%)
It was a good experience.	3 (9.1%)
It was refreshing and/or I enjoyed it.	3 (9.1%)
It was difficult.	3 (9.1%)
I would like to try writing with other subjects.	2 (6.1%)
It was embarrassing and I don't like WL much.	1 (3.0%)
No answer	3 (9.1%)

Note. $N = 33$.

CHAPTER VII

QUALITATIVE RESULTS

As stated earlier, two interviews were conducted to complement the written data from the exit questionnaire. Moreover, in order to gain further insights into the potential impact of WL on L2 learning, a case study approach was employed. Framed within an SCT perspective, this analysis attempted to elucidate the participants' approach towards WL. In this chapter, therefore, the results of the two interviews are reported, followed by those of the four case studies.

7.1 Interview Results

As reported, two +WL participants from two of the three participating classes were interviewed regarding their WL experience immediately after the posttest (during lunchtime) in Week 4 in order to complement the written data obtained from the exit questionnaire and to obtain some insights into their perceptions regarding their experience of WL. The two interviewees, Mari and Yoko (pseudonyms), were both motivated female learners, as can be imagined in that they volunteered to be interviewed, sacrificing part of their lunchtime. Their levels of English proficiency, however, were different. Mari was from the second-highest pharmacy class, and she was one of the higher-proficiency participants in the +WL group, judging from her TOEIC score (525) and pretest scores. The other participant, Yoko, was from the second-highest level management class. As stated, the proficiency levels of the pharmacy students were generally higher than those in the management department, and Yoko's proficiency level was rather lower than Mari's, also judging from her TOEIC score (315) and pretest scores, but about the average of the +WL group. As mentioned earlier, the interviews

were semi-structured (all the questions in the exit questionnaire were asked) and the participants were requested to elaborate their answers when necessary. As stated, the interviews were recorded with their permission and the transcripts were analysed. Each interview lasted for about 10 minutes. The details are introduced part by part.

7.1.1 Part I: Reflecting on the Experiment

The answers of the two interviewees are summarized in Table 7.1. As can be seen, their responses generally overlapped with those of the +WL participants in the exit questionnaire. As for the purpose of this experiment, neither of them got it right, stating to examine "essay writing ability" (Mari) and more generally "English ability" (Yoko). Regarding the next question on their perceptions of learning, both of them replied essay writing. In addition, they commented:

Mari: Essay writing was new to me, but it was a nice surprise to learn that I can write my ideas in English better than I anticipated.

Yoko: I found writing essays challenging because I had not written essays in English before.

Probably because of her higher proficiency, Mari's attitude was more positive than Yoko's, as reflected in her reply "I enjoyed writing essays" in answer to the last question on free comments (Q5). What is noteworthy is that both of them replied that they focused on the target construction in response to Q4. Asked about why they focused on the target construction, both of them replied that they noticed many sentences on the target construction in the original text, the pretests and posttests.

When their WL sheets were examined, both Mari and Yoko were found to have

produced T-WLEs. Mari produced three T-WLEs, one each for three target sentences. Underlining "had" in the first target sentence, "If I had better grades," she wrote *kateiho dakara kako* "counterfactual conditional so past tense" for the first target sentence. The next T-WLE was written under the second target sentence, *mata kateiho dakara jijitsu jyanai* "again counterfactual conditional, so not true in reality." For the third target sentence, she wrote *kateiho dakara jisei zurasundatta* "I should have moved tense backward because it is counterfactual" (She used the present tense for the sentence.).

Yoko produced two T-WLEs, one on the first target sentence and the other on the second one. As for her first T-WLE, like Mari, she underlined the words "if" as well as "had" in the first target sentence, saying *moshimo dakara kako* "if so past tense." The other one was a partial translation of the second target sentence. Yoko's reconstruction was not as successful as Mari's and only the first target sentence was reconstructed, but unsuccessfully, which might have made it difficult for her to compare the original text with her reconstruction. As stated, Mari was a higher-proficiency learner than Yoko, which is likely to account for the difference in the quality of their reconstructions. Moreover, the difference in their proficiency levels, and possibly ZPD, may also explain the outcome that Mari's T-WLEs reached the level of understanding, whereas it was partial understanding for Yoko.

Table 7.1
Results of the Two Interviews (Part I)

	Mari	Yoko
Q1. The purpose was to examine...	essay writing ability	English ability
Q2. I learned…	essay writing	essay writing
Q3. What I learned was...	I got used to writing essays.	the target construction
Q4. I focused on…	the target construction	the target construction, past tense
Q5. Free comments	I enjoyed writing essays.	I felt I worked harder than usual.

7.1.2 Part II. Reflecting on WL

The answers to the questions in Part II appear in Table 7.2. In contrast to the responses to Part I, which were not very different from the exit questionnaire results of the +WL participants, this part elicited responses unique to each interviewee, enabling me to obtain insights into their perceptions of WL as well as to reflect my teaching. Of all the replies, what was especially noteworthy was Mari's comment in response to the first question regarding her experience of WL.

Mari: I usually do something similar to what I did in the experiment.

Researcher: Is that so? Please tell me more.

M. Well, my high school maths teacher always encouraged us to write the reasoning of our thinking. Even in the case of multiple-choice questions for our homework, he instructed us to write reasons for our choices.

R: That's very interesting!

M: Oh, really? I've never thought much about it. Some students did not like it because the requirement made it difficult for them to copy their friends' answers [giggling], while others simply found it troublesome.

R: How did you feel?

M: I found it helpful to organise my thoughts and to confirm my understanding when I myself was not sure without writing it down.

R: I see. Do you still write down your thoughts?

M: Yes, sometimes. Not always, but I still write down my thoughts and questions when I encounter difficult problems, mainly in the courses in my pharmacy department. Also, what I find helpful is that I can look back at what I wrote earlier. I guess I can deepen my thoughts and notice my mistakes more easily that way.

Her high school maths teacher's (a Japanese male) background is not clear, but what he encouraged his students to do, and presumably is still doing, is precisely self-explaining in the written mode. Because of the experience, although not for language learning, Mari seems to have been using written self-explaining as a learning tool, benefitting from the process and product as a mediational artefact. In the end, she suggested that I encourage my students to use WL regularly, not just for the experiment, because she believed that WL was beneficial.

Similarly, interesting findings also emerged from Yoko's replies. Unlike Mari, WL was new to her, so she found it difficult to write her thoughts on linguistic issues. At the same time, however, she felt it was helpful. Asked about how she perceived WL, she replied as follows:

Yoko: I think it is easier to notice my mistakes when I write. Also, I think I can remember what I write easily. I always write English vocabulary items and Chinese characters when I need to remember them.

Researcher: Is that so?

Y: Actually, I use writing in my daily life. As you know, I'm in the track and field club, and I always write my next goal, such as running under a certain time limit or joining a big competition, on a card and put it on the wall of my room. Every time I read the card, I feel motivated. Also, the card reminds me what I have to do to achieve my goal.

Yoko's comment regarding writing English vocabulary and Chinese characters to help remember them seems to suggest that she is using writing as a tool for language learning (Manchón, 2011). What is more, judging from her next comment with respect to the product of writing, she is likely to consider that being able to read the product of

writing repeatedly is the strength of writing, echoing those researchers who have pointed to the heuristic nature of writing (e.g., Emig, 1977; Luria, 1999; Manchón, 2011). At the end of the interview, she said that she might try using WL again, which may suggest that now she is aware of the facilitative effect of externalising her thoughts by writing, i.e., WL.

Table 7.2
Results of the Two Interviews (Part II)

	Mari	Yoko
Q1. I found WL…	beneficial, I usually do similar things	difficult but helpful
Q2. Perceptions of WL	I can think deeply and look back.	I can notice my mistakes.
Q3. What I learned from WL	Writing is useful although I knew it.	I can remember what I write.
Q4. Free comments on WL	WL should be employed regularly.	I may try WL again.

7.2 Case Studies

As mentioned earlier, the four case studies were conducted from an SCT perspective. The four case study participants, Yuta, Rieko, Hiroshi and Takuya (all pseudonyms), were selected for the following reasons. First, two of them, Yuta and Rieko, demonstrated a positive attitude towards WL, but the quality of their T-WLEs and the level of their learning differed (see Tables 7.3–7.5 below for the pre- and posttests results and Table 7.6 for the frequency and quality of WLEs of the four participants). Second, the other two, Hiroshi and Takuya, were two of the participants who left negative comments regarding WL in the exit questionnaire for different reasons. Third, three of them, Rieko, Hiroshi and Takuya, were comparable in terms of their English proficiency judging from their pretest scores, TOEIC scores and regular class performance (i.e., lower-intermediate proficiency learners). Probably because of the differences in their attitudes towards WL, however, their WLEs differed both

qualitatively and quantitatively. The three participants were in the third-highest management class, whereas Yuta was from the second-highest pharmacy class.

Table 7.3
Essay Test Scores of the Case Study Participants

	Pretest		Posttest		Delayed posttest	
	OC	P/C	OC	P/C	OC	P/C
Yuta	6	.67	5	2.00	7	1.86
Rieko	3	.00	4	.00	4	1.00
Hiroshi	6	.33	9	.22	6	.67
Takuya	1	.00	3	.00	1	.00

Note. OC: number of obligatory contexts, P/C: points per context, maximum score for P/C = 5.

Table 7.4
Grammar Production Test Scores of the Case Study Participants

	Pretest			Posttest			Delayed Posttest		
	AU	OU	OA	AU	OU	OA	AU	OU	OA
Yuta	11	0	11	16	0	16	15	0	15
Rieko	0	2	-2	10	4	6	7	2	5
Hiroshi	0	0	0	3	8	-5	4	6	-2
Takuya	4	4	0	8	4	4	2	4	-2

Note. AU: accurate use, OU: overuse, OA: overall, maximum score for AU & OA = 16, OU = 12.

Table 7.5
Recognition Test Scores of the Case Study Participants

	Pretest	Posttest	Delayed Posttest
Yuta	10	14	16
Rieko	9	10	15
Hiroshi	8	11	12
Takuya	7	6	11

Note. maximum score = 16.

Table 7.6
Frequency and Quality of T-WLEs of the Case Study Participants

	Frequency	Quality
Yuta	3	3
Rieko	3	2
Hiroshi	1	1
Takuya	0	0

7.2.1 Participant 1: Yuta

Yuta was the participant who commented that he felt WL was beneficial because it enabled him to review what he wrote. In addition, he replied that WL helped him to organise his thoughts regarding linguistic issues. He produced five WLEs, of which, as presented in Table 7.6, above, three were categorised as T-WLEs. In the first T-WLE, he wrote "Oh, past tense," probably noticing that the present counterfactual conditional was used, which was categorised as "noticing." The next one, "past because of if," written next to the second target sentence, was coded as "partial understanding." The final one, the correct grammar rule of the target construction, written below the third target sentence, was coded as "understanding" (i.e., categorised as 3, the highest level).

These T-WLEs seem to demonstrate the process through which he improved his quality of noticing, i.e., by writing what he noticed and consolidating his understanding. Put differently, the T-WLEs may reflect a transformative process of internalization, facilitated by WL as a mediational tool. Supporting this speculation, as presented in Tables 7.3–7.5, above, an examination of his pretest and posttest scores showed improvement in his test scores, especially on the essay tests. Namely, he achieved gains not only in terms of obligatory contexts, but also in points per context (i.e., accuracy).

Meanwhile, as stated above, of Yuta's five WLEs, two were not on the target construction, i.e., NT-WLEs. Both of them concerned the failure to put plural –s on the word "option" (*sentakushi* in Japanese). His first NT-WLE was on the word in the sentence "I need to think about my *options*." Circling the "s" of the word, he wrote *ire wasure* "I forgot to put (an 's')." His second NT-WLE was also on the word in the sentence "None of these *options* work." Circling the "s" again, he wrote *fukusu no sentakushi dakara s hitsuyou* "plural options, so an 's' is necessary." In Japanese, singular and plural forms are not always differentiated. To be more precise, the word

sentakushi can be used for both singular and plural options, as in *hitotsu no sentakushi* and *fukusu no sentakushi*, respectively, which makes it hard for Japanese learners to use plural-s on plural nouns.

Although these WLEs were not on the target construction, they seem to demonstrate Yuta's attention to form rather than meaning. Like Mari, Yuta's reconstruction was fairly successful, which is likely to have enabled him to pay attention to a non-salient form (i.e., plural-s). Moreover, the two NT-WLEs seem to demonstrate a process whereby he consolidated his understanding.

7.2.2 Participant 2: Rieko

As stated earlier, Rieko's attitude toward WL was extremely positive. In fact, she was the participant who commented "I did not write what I knew, but I wrote in order to understand," regarding WL in the questionnaire (Part II, Q3), which seems to be a perfect example of the "writing-to-learn" perspective proposed by Manchón (2011). An examination of her WL sheet revealed that she was one of the written high-languagers, producing 13 WLEs, of which three were categorised as T-WLEs. (The medians of WLEs and T-WLEs were 6.00 and 1.00, respectively.) Of the three, for the first two, she circled "If," the first word of each target sentence and wrote its Japanese translation *moshimo* for the first sentence, whereas *kateiho?* "counterfactual?" for the second target sentence, which were both coded as "noticing." As stated above, she was a lower-proficiency learner and as the second T-WLE demonstrates, she seems to have been unsure about the structure of the sentence. Meanwhile, her third T-WLE was a correct partial translation of the third target sentence, coded as "partial understanding" (labelled as 2 on a scale of 3). Given the three T-WLEs and her comment in the exit

questionnaire, engaging in WL and producing WLEs is likely to have heightened her awareness towards the target construction, developing her understanding of it.

Compared to Yuta's T-WLEs, Rieko's T-WLEs seem to demonstrate that her understanding is still insufficient (understanding as opposed to partial understanding). Nonetheless, they seem to indicate "learning in progress" (Swain, 1995). Supporting this speculation, as shown in the tables above, she improved her posttest scores on all three tests. As for the essay tests, although she only showed a slight increase in terms of obligatory contexts, she achieved a considerable gain regarding points per context, using the past tense for all the if-clauses of the target sentences she produced. Similarly, she improved her score drastically on the recognition test, achieving almost a perfect score on the delayed posttest. Although her improvement was not as drastic concerning the grammar production tests, she still showed the greatest gain scores of the four case study participants in terms of her overall score.

Meanwhile, the 10 WLEs which were not on the target construction, i.e., 10 NT-WLEs, concerned translations of non-target sentences. They were likely to be written to consolidate her understanding of the original text, i.e., "in order to understand" as she wrote on the questionnaire. This speculation is compatible with the view of Cumming's (1990), who examined the think-alouds of participants (university students) in his aforementioned study with the aim of identifying their thinking processes while they were writing in L2. Pointing to the fact that some of the participants translated words when they faced linguistic issues, Cumming explains that "translations … served a *compensatory* purpose—to counter problems in expressing ideas in the L2" (p. 502). Although the situation was not exactly the same, Rieko might have also used translations to counter her problems, i.e., in understanding the original text, in her case.

7.2.3 Participant 3: Hiroshi

As reported, Hiroshi showed some reservations towards WL, commenting that he did not find the need to write because he could think in his head without writing. Despite his comment, an examination of his WL sheet showed that he still produced three WLEs, of which one was on the target construction, i.e., T-WLE. Next to the first target sentence he wrote "if sentence," which was categorised as "noticing." (The other two NT-WLEs were about the two vocabulary items, i.e., "besides" and "none." The first one said "I don't know the meaning of this word." The second one said "I remember seeing this word…," probably intending to say that he did not know the meaning of it, either.) As presented in Tables 7.3–7.5 above, he was a lower scorer at the outset of the experiment. When his pre- and posttest scores were examined, it was found that he improved his scores on the essay tests and recognition tests, but not on the grammar production tests, which showed an increase in overuse.

His T-WLE and scores on the pre- and posttests seem to support Ericsson and Simon's (1993) claim regarding the level of verbal protocols. That is, in their perspective, only Level 3 verbalization, which requires learners to explain what they are thinking, affects the way they think, resulting in reactivity, while Level 1 or Level 2 verbalizations, that do not ask them to explain, i.e., the type of T-WLE which Hiroshi made, has no reactivity, either positive or negative. It is not clear if the quality of his T-WLE (i.e., noticing, labelled as 1) reflected his true inner level of understanding or if he produced it even though he had deeper understanding, thinking it unnecessary to write in detail but still feeling he had to write something. Judging from his posttest scores, however, the former seems to be more plausible. Although he claimed to be able to think in his head, the test results suggest that this was not as successful as he claimed, at least in terms of the grammar production tests. Put differently, his performance on the

grammar production tests seems to indicate the facilitative role of WL in acquiring explicit knowledge.

Furthermore, Hiroshi's WLEs and exit questionnaire comments indicate two issues that were supposedly improved points, based on the experience of the pilot study. First, despite the increased practice sessions on WL, they were probably still not sufficient to let all the participants fully understand its concept. Second, as his second NT-WLE ("I remember seeing the word…") indicates, the word "none" was on the vocabulary sheet distributed two weeks before the experiment. (In the pilot study, the sheet was given out one week prior to the experiment.) As stated, some time was spent going through the vocabulary items in two classes to familiarise the participants with the items beforehand. Nonetheless, his comment on the exit questionnaire and WLEs indicate neither the WL practice nor the vocabulary training was enough.

7.2.4 Participant 4: Takuya

Like Hiroshi, Takuya was one of the participants who made negative remarks about WL on the questionnaire. Unlike Hiroshi, however, his reason derived from an affective factor, that is, he was the participant who reported that he felt shy about writing his inner thoughts, commenting "it was embarrassing." An examination of his WL sheet revealed that he produced no T-WLEs (i.e., coded as zero for both frequency and quality), but two NT-WLEs, *bumpo nante wakaranai* "I don't understand grammar," and *konnna koto hazukashii* "this is embarrassing." Although it is beyond the scope of this thesis, learners' affective factors, such as attitudes, beliefs and goals, have been identified to influence the impact of corrective feedback (Storch & Wigglesworth, 2010; Wigglesworth & Storch, 2012). Examining the effect of feedback in relation to learners' individual differences, Wigglesworth and Storch reported that the affective

factors of the participants not only influenced how they processed the feedback, but also "their willingness to accept the feedback" (p. 329), which could be applied to WL. Given this, these two NT-WLEs may be interpreted as "his unwillingness to accept WL" or even his rejection of it. As expected from his attitude and the lack of T-WLEs, he did not demonstrate much improvement on the posttests (see Tables 7.3–7.5 above). More specifically, like Hiroshi, he improved his recognition test scores from 7 on the pretest to 11 on the delayed posttest, but showed no gain in terms of the two production tests. Conversely, he showed a slight decrease regarding the overall score on the delayed posttest because of his persistent overuse.

Reviewing studies conducted on private speech with children (Saville-Trokie, 1988) and adults (Ohta, 2001), Lantolf (2006) identified that adults produce private speech less frequently compared to children, probably because they are more self-aware (p. 98), which was part of the motivation for the current study, that is, to offer presumably less face-threatening facilitation than speaking to adult self-conscious learners on the assumption that WL is less intimidating for them as it is generally a private/individual process. However, knowing that I, his instructor, would check his WL, Takuya might not have felt it was a private process. This interpretation is consistent with the argument from Hausmann and Chi (2002). They reported that participants generated significantly fewer typed self-explanations (i.e., WL) compared to previous studies on oral self-explaining, probably because they were aware that the researchers would check their typed self-explanations later (i.e., avoidance of errors). Based on this finding, Hausmann and Chi argue that the characteristics of written modality, i.e., the nature of permanent records, are likely to have discouraged the participants to write freely, resulting in far fewer self-explanations than oral self-explanations in their earlier study (Chi et al., 1994). Given this, although producing

objects to reflect on, i.e., external permanent records, is expected to be one of the positive features of WL (W. Suzuki, 2012), the result obtained here seems to suggest that it can also be a negative factor depending on learners' individual differences.

7.2.5 Summary

To summarise the findings, the four case study participants approached WL differently, which is likely to have influenced the frequency and/or quality of their T-WLEs and their learning outcomes. The result that the most noticeable difference was observed regarding the grammar production tests seems to suggest that WL is effective for tasks that can be addressed by mainly drawing on explicit knowledge, as hypothesised. Meanwhile, with respect to the recognition tests, reflecting there being no statistically significant difference observed between the +WL group and the –WL group, the four participants improved their scores regardless of the frequency or quality of their T-WLEs. That said, Yuta and Rieko achieved higher gains than Hiroshi and Takuya.

The three participants, Rieko, Hiroshi and Takuya, started out with similar levels of understanding regarding the target construction. Their different attitudes towards WL, however, brought about different outcomes for them. As for Rieko, she produced notably higher number of WLEs than the other participants in the +WL group, including three T-WLEs, which seems to be the embodiment of Manchón's (2011) writing-to-learn perspective. It is worth mentioning, however, that both Yuta and Rieko produced three T-WLEs, thereby being equal in terms of the frequency of T-WLEs, but not of quality. As reported, while Yuta reached the highest level of understanding, it was partial understanding for Rieko. As stated in section 7.1.1, similar results were found for Mari and Yuko, whose T-WLEs were categorised as understanding and partial understanding, respectively. (It should be pointed out, however, that Mari produced

three T-WLEs, whereas Yoko produced two.) These differences may indicate that the effect of WL is contingent on proficiency level and ZPD.

Meanwhile, Hiroshi and Takuya showed negative attitudes toward WL for different reasons. Not recognising the potentially facilitative impact of WL as a mediational tool, Hiroshi claimed he could think in his head without WL. His scores on the grammar production tests, however, seem to suggest that externalising thoughts in writing, i.e., WL, is more effective than thinking in his head, at least in terms of the grammar production tests. In Takuya's case, he found WL embarrassing, which might also explain the smaller number of obligatory contexts (i.e., one for the pretest and delayed posttest, but three for the posttest) as the essay tests required the participants to write about their personal ideas. In addition, given his other NT-WLE, "I don't understand grammar," which does not show much eagerness to learn, his attitude may be better explained by Lantolf (2006), who states that "not all students enrolled in adult university language classes are motivated to learn the L2—their goal might be to fulfil a language requirement—they are less likely to engage in language-focused private speech" (p. 98).

Although it is beyond the scope of this thesis, as the case studies suggest, the impact of WL on learning is likely to increase, as was the case with Yuta and Rieko, or decrease, as was observed with Hiroshi and Takuya, depending on learners' attitudes and/or beliefs regarding WL. In other words, the case studies seem to prove the idea that a learner is an agent "who perceives, analyses, rejects or accepts solutions offered, makes decisions and so on" (Swain, 2006, pp. 100–101).

CHAPTER VIII

DISCUSSION

In this chapter, the results of the empirical study reported above are discussed in relation to the research questions and hypotheses posed in Chapter 3. Some additional findings are also considered.

8. 1 Effects of WL (RQ1)

The first research question asked the extent to which WL can facilitate L2 learning. The gain scores on the three assessment tests (i.e., essay test, grammar production test, recognition test) of the three groups (i.e., +WL group, –WL group and control group) were compared and analysed, which revealed generally better performance of the +WL group than the other two groups (i.e., –WL group and control group). More specifically, the +WL group achieved greater gains than the control group on all the assessment tests with small to medium effect sizes, thereby excluding the possibility that the improvement in the +WL group could be attributed to test-retest effects. Furthermore, when the two treatment groups were compared, in spite of the same amount of exposure to the target construction, it was found that the +WL participants showed superior gains compared to their –WL counterparts on the two production tests, i.e., essay tests and grammar production tests, also with small to medium effect sizes. No statistically significant difference between the two groups, however, was observed for the recognition tests.

These results indicate that WL had a positive impact on L2 learning, as evidenced in the participants' performance on the two production tests. It is important to note that the findings obtained here are consistent with those of the pilot for this main study (M.

Ishikawa, 2018) and previous studies on WL (e.g., M. Ishikawa & W. Suzuki, 2016). Like this one, those studies identified facilitative results only on production tests, but not on recognition tests. These findings are discussed in more detail below. As the analyses of the three groups confirmed that the results of the current study were not due to test repetition effects, the control group is excluded from the remainder of the discussion.

8.1.1 +WL Group vs –WL Group

As reported, the +WL group significantly outperformed the –WL group on the two production tests (i.e., essay tests and grammar production tests). To be more precise, the +WL participants achieved greater gains on the grammar production tests, with significant differences found for their overall scores (the total of accurate use scores minus overuse scores) with a small effect size in the long term between the two groups. On the essay tests, significant differences were found in terms of obligatory contexts, an indicator of the number of attempts at the target construction and a possible indicator of the development in participants' knowledge of the meaning associated with the target construction, in both the short and long term, with large and small effect sizes, respectively. No such difference was identified, however, in terms of points per context, an accuracy measure.

These favourable results for the +WL group over the –WL group seem to indicate that the extra language use which the +WL participants produced benefited them through both the process and product, as hypothesised. In terms of the process, three factors, (1) enhanced noticing, (2) deeper processing, and (3) generation effect, are likely to have contributed to the superior performance of the +WL participants over their –WL counterparts, as hypothesised. First, supporting the noticing function of the

Output Hypothesis (Swain, 2005), having engaged in WL, i.e., producing language when comparing the original text and their reconstructions, is likely to have enhanced the +WL participants' noticing of the target construction. Given that output is expected to function as an "attention-getting device" (Swain & Lapkin, 1995, p. 373), WL might have contributed to enhanced noticing and greater gains of the +WL participants as a consequence. The results also support Schmidt's (2001) Noticing Hypothesis, which claims noticing facilitates learning.

Some evidence to prove this conjecture was found in the exit questionnaire. That is, as reported earlier in section 6.9.1.4, in response to the question that asked what the participants focused on during the experiment (Part I, Q4), 63.6% of the +WL participants (21 out of 33) responded that they focused on "the present counterfactual conditional," i.e., the target construction, whereas the ratio of the –WL participants who responded similarly was much lower at 38.7% (12 out of 31), indicating that the +WL participants enhanced their noticing. Additional supporting evidence was found in the second part of the questionnaire, which was only given to the +WL group. Responding to the question that asked the +WL participants how they perceived their experience of WL (Part II, Q1), nine participants (27.3%) replied that they found WL helpful because it helped them to notice their linguistic issues. More specifically, of the nine participants, six (18.2%) wrote that WL helped them notice their mistakes, while three (9.1%) stated that WL helped them notice their lack of understanding. These responses seem to be in parallel to the notions of noticing a gap (i.e., a mistake, a difference between their reconstruction and the original text) and a hole (i.e., a lack of knowledge) (Swain, 1998), which are expected to trigger learning.

The results of this thesis also seem to be in line with the findings of reactivity research (e.g., Bowles & Leow, 2005; Yanguags & Lado, 2012). Bowles and Leow

(2005), for instance, reported that all the participants who were instructed to verbalize justifications (OL), i.e., Level 3 verbalisation according to Ericsson and Simon's (1993) typology, generally demonstrated some awareness regarding the target structure. Meanwhile, only the participants with higher proficiency did so under the non-verbalizing condition, suggesting the facilitative impact of verbalization on enhancing awareness towards the target construction. Similarly, interpreting the positive reactivity of think-alouds, Yanguas and Lado (2012) state that the very act of thinking-aloud was likely to have heightened their participants' awareness of syntactic structures. Regardless of the difference in modality, these studies seem to support the argument that WL can enhance noticing. Given these findings, WL may have enhanced the participants' awareness regarding the target construction. As mentioned earlier, although output and WL are not the same constructs, the present thesis hypothesised that WL, i.e., additional language production on linguistic issues, would enhance the noticing function of output (Swain, 2005) even further. The results obtained here seem to support the prediction of this thesis. Also, the fact that the results are consistent with those of OL seem to prove Swain's (2006) statement that the benefits of languaging apply to both speaking and writing.

Second, as discussed when interpreting the results of the pilot study, the process of WL is also likely to have induced the +WL participants' deeper processing, i.e., a greater degree of analysis (Craik & Lockhart, 1972, p. 675), yielding longer-lasting and stronger memory representations compared to the –WL participants and further resulting in a favourable result for the +WL participants. With respect to the depth of processing, highlighting the importance of output, Swain (2000) also states that "output pushes learners to process language more deeply—with more mental effort—than does input" (p. 99). As stated above, it should be kept in mind that output and WL are not

identical, but given that both of them entail learners' language production, Swain's statement is likely to apply to WL, at least to some extent. Supposing this speculation is right, the opportunity of WL, i.e., to language about language, might have pushed "learners to process language more deeply—with more mental effort—*than does no opportunity of WL*," resulting in deeper processing. This interpretation seems to account for the superior performance of the +WL group compared to the –WL group observed in the long term regarding their overall scores, although with a small effect size. Furthermore, the questionnaire results seem to support the interpretation concerning deeper processing. That is, in response to the same question on the perceptions of WL stated above (Part II, Q1), eight participants (24.2%) wrote that they found WL useful because it helped them to think more deeply, which was the second most popular response next to the two types of noticing (i.e., noticing a hole or a gap) combined (nine responses, 27.3%). Some of them even commented that their thinking would have been shallower or superficial without WL, which appears to provide evidence that language, WL in this thesis, can function as a cognitive tool (Vygotsky, 1986).

Third, the result is also likely to be attributable to the generation effect proposed by Slamecka and Graf (1978), who claim that learners recall what they generate better than what they simply read at a later point in time, as well as the self-explanation effect proposed by Chi (2000). The comments found on the exit questionnaire seem to provide additional evidence to support this speculation. That is, in response to the second question which asked the +WL participants if they thought there would have been a difference if they hadn't written their thoughts while checking (Part II, Q2), the majority of them (i.e., 87.9%, 29 out of 33) responded in the affirmative. Asked about the reasons for their responses, nine participants (31.0 %) replied that WL helped them to remember what they were thinking. As reported, one of them also wrote "I can remember if I

write," referring to the benefit of writing (generating language). Also, as mentioned, Yoko, one of the interviewees, made a similar comment, referring to her habit of writing English vocabulary items and Chinese characters when she needed to remember them, which is likely to be explained by the generation effect. Similarly, to the next question regarding their experience of WL (Part II, Q3), seven participants (24.1%) responded that WL helped them to remember their linguistic issues. It is believed that languaging can occur inside learners' minds like the metalinguistic thinking that learners use when they compose in L2 (Cumming, 1990). The results obtained here, however, may imply the superiority of externalising thoughts by producing language, i.e., WL, in this thesis, over silent languaging. Taken together, the process of WL appears to have benefited the +WL participants by enhancing their noticing, triggering deeper processing and bringing about the generation effect.

In addition to the process, the products of WL may be claimed to have profited the +WL participants as a source of further reflection on their linguistic issues (Swain, 2006). Although the treatment in the current study only lasted for five minutes and did not offer any additional chances to utilise the products of WL, evidence to support the claim was again found on the exit questionnaire and the interviews. For example, in response to the last question that asked for free comments regarding WL (Part II, Q4), four participants (12.1 %) commented that they felt WL was beneficial because they could review the products of writing later. It should be pointed out, however, given that the treatment was only for five minutes, these replies might have been stemmed from their previous personal experience regarding writing. Similarly, Mari, the other interviewee and experienced written languager, commented that she found WL helpful because she could go back to her earlier thoughts and build on them whenever necessary. These comments seem to support another hypothesis of this thesis, that is, the

product of WL will enhance the reflective function of output as claimed by the Output Hypothesis (Swain, 2005).

Moreover, given that the facilitative functions of output "may be stronger for written production due to the more generous time constraints and permanent record of writing" (J. Williams, 2012, p. 323), the observed favourable results for the +WL participants might have been attributable to the heuristic nature of writing. As stated, a slower pace and a permanent record are two unique characteristics of writing. Because of these characteristics, the process of writing enables learners to deepen their thoughts without imposing as much time pressure as speaking, allowing them to reflect on the product of their writing, i.e., a permanent record/external memory (Hertel, 1993; W. Suzuki, 2012). In the aforementioned study, which compared typed and oral self-explaining, Muñoz et al. (2006) identified that less-skilled readers used inferencing (i.e., a skill that is supposed to contribute to learning to a greater extent than others such as rereading) comparably to high-skilled readers in typing, whereas this was not the case when they engaged in oral self-explaining. Interpreting this outcome, they pointed to the "more reflective response mode" of typing as a possible contributor. Along the same lines, as noted earlier, Luria (1982, 1999) states that "the conscious act of writing" (1982, p. 167) enables learners not only to develop their existing thoughts but even to revert to their earlier thoughts, calling it a "slower repeated mediating process of analysis and synthesis" (1999, p. 103) compared to speaking.

As introduced in section 7.2.1, Yuta, one of the four case study participants who referred to one benefit of WL being able to review its products, produced three T-WLEs that seem to support Luria's (1999) statement. That is, to repeat his T-WLEs, reading the first target sentence, he wrote "Oh, past tense," probably noticing that the counterfactual was used although he used a simple conditional (i.e., an example of noticing a gap).

Then, reading the second target sentence, he is likely to have noticed the counterfactual and the use of the past tense again, writing "past because of if." Although it does not clearly show his inner thoughts, it might be taken as a self-explanation to himself, i.e., the process of his analysis. Finally, noticing the last target sentence, he produced a third T-WLE, the grammar rule of the present counterfactual conditional, presumably synthesising his understanding of the target construction. If this interpretation is correct, despite the short time of the treatment, he might have benefited from not just the process of WL, but also from its product, reviewing his earlier T-WLEs as he commented on the questionnaire.

It is important to note, however, that there is a possibility that the –WL participants might not have engaged in silent languaging at all, despite the instruction to do so (i.e., to compare their reconstructions with the original text without writing). If that was the case, the favourable results for the +WL group might not be due to WL as opposed to silent languaging, but to WL as opposed to no languaging. Supporting this speculation, four +WL participants replied that they found WL helpful because it helped them to "think," stating that they might not have thought seriously without the requirement to write their thoughts. If that was the case, WL might guarantee that learners engage in some kind of thinking and/or reflection. Moreover, there is another possibility that the participants in both groups (at least some of them in each group) might have engaged in OL through whispering to themselves, such as private speech (e.g., Ohta, 2000; Yoshida, 2008). Given this possibility, the result that the +WL participants outperformed their –WL counterparts could be due to double languaging, (i.e., both WL and OL) by the +WL participants as opposed to single languaging (i.e., OL) by the –WL participants. It is impossible to be sure which was the case, but providing an opportunity of WL seems to guarantee that learners engage in at least one

type of languaging (i.e., WL), pushing them to think and/or reflect as a consequence.

Although further research is needed to elucidate exactly what contributed to the favourable outcome for the +WL group, the results obtained in this thesis seem to indicate that providing opportunities for WL can be facilitative for L2 learning, at least for the production tests (i.e., essay tests and grammar production tests), which demonstrated statistically significant differences between the +WL group and the –WL group. In contrast, no statistically significant differences were observed regarding the recognition tests, which will be discussed below.

8.1.2 Production (Essay & Grammar Production) Tests vs Recognition Tests

As reported above, the +WL group significantly outperformed the –WL group on the two production tests (i.e., essay tests and grammar production tests). No such difference, however, was found on the recognition tests, echoing the findings of previous WL studies (e.g., M. Ishikawa, 2018; M. Ishikawa & W. Suzuki, 2016). Given that the three assessment tasks in this thesis were employed in order to measure the treatment effects on different aspects of L2 acquisition, this discrepancy may come as no surprise. That is, the recognition tests were included to assess learners' receptive knowledge of the form-to-meaning mapping associated with the target construction, whereas the two production tests were employed to gauge their ability to produce language, i.e., their productive knowledge of the meaning-to-form mapping linked to the target construction. With this in mind, at least two reasons may account for the outcome (i.e., more favourable results for the +WL group only for the production tests, but not for the recognition tests).

First, the nature of the treatment and assessment tasks might have contributed to the outcome. That is, on the one hand, the treatment condition that the –WL group

experienced (i.e., comparing the original text and their reconstructions, and trying to recognize differences) and the recognition tests that assessed the participants' receptive knowledge seem to share similar processing (i.e., to process input). Therefore, this similarity might have prepared the –WL participants better for the recognition tests than for the two production tests. On the other hand, the act of writing is what WL and production tests share (i.e., to produce language). Accordingly, having engaged in WL might have prepared the +WL participants better for the production tests. This conjecture is compatible with the notion of the transfer of appropriate processing, whose principal tenet is that people can better remember what they have learned "if the cognitive processes that are active during learning are similar to those that are active during retrieval" (Lightbown, 2008, p. 27). It should be pointed out, however, that there were no significant differences between the two groups on the recognition tests, i.e., the –WL group did not outperform the +WL group. As such, this notion does not seem to explain the outcome completely.

Second, once again, as stated to explain the similar outcome of the pilot study, the results may be explained by the Output Hypothesis (Swain, 2005), which claims that output provides opportunities for learners to shift from semantic processing to syntactic processing. Given this, the two production tests are likely to have required the participants to have a more accurate understanding of the target construction because learners cannot "fake" (Swain, 1995, p. 127) their understanding in producing language. Thus, the favourable results of the +WL group over the –WL group regarding the two production tests seem to suggest that their WL experience allowed them "to move from the semantic, open-ended, strategic processing prevalent in comprehension to the complete grammatical processing needed for accurate production" (Swain, 2000, p. 99), as the Output Hypothesis claims. In contrast, considering that the recognition tests were

designed to measure the participants' knowledge of meaning, a mere comparison with the original text and their reconstructions (i.e., to process input) without WL (i.e., only with potentially silent languaging) might have been sufficient to address the recognition tests, contributing to no significant difference between the two groups.

It should be pointed out, however, that there was also a difference regarding the results of the two production tests, to which this discussion will now turn.

8.1.3 Essay Tests vs Grammar Production Tests

Although statistically significant differences were only observed between the +WL group and the –WL group with respect to the two production tests (i.e., essay tests and grammar production tests), the findings between these two tests also differed. As reported, with respect to the grammar production tests, the +WL group outperformed the –WL group significantly on overall scores (i.e., the total of accurate use score minus overuse score), the most reliable indicator of learning of the three scores, in the long term with a small effect size. On the essay tests, however, significant differences were found only in terms of obligatory contexts, an indicator of the number of attempts at the target construction in both the short and long term with large and small effect sizes, respectively, while no such difference was identified in terms of points per context, an indicator of accuracy. Accounting for the outcome, two possible explanations emerged.

First, as suggested regarding the difference between the production tests and the recognition tests, although both essay tests and grammar production tests required participants to produce language (i.e., to write), the fundamental differences of the nature of the two tests are likely to have contributed to the differing outcomes. To be more precise, the grammar production tests were form-focused and what the learners were required to do was simply to put the appropriate form of a verb supplied in

parentheses next to each blank. Given the nature of the task, during the grammar production tests, the participants were able to draw on both their declarative and procedural knowledge. That said, it is likely that they could have done the tests by mainly relying on their declarative knowledge. In contrast, the essay tests were open-ended and necessitated participants to think about and pay attention to not just grammar, but also content, organization and vocabulary items. Therefore, in addition to declarative knowledge of the target construction, the essay tests are likely to have required the ability to apply it in a more natural setting, i.e., procedural knowledge, to a greater extent than the grammar production tests. Accordingly, the difference is likely to have made the essay tests more challenging than the grammar production tests. Supporting this interpretation, in response to the last question on the exit questionnaire which asked for free comments regarding the experiment to all the participants in the two treatment groups (Part I, Q5), around 20% of the participants in each group commented that they found essay writing difficult.

Given these findings, it seems plausible that the cognitively more demanding characteristics of the essay tests are likely to have exceeded the capacity of the participants' attentional resources (VanPatten, 1990). The depletion of their attentional resources, therefore, might have made it more difficult for the +WL participants to draw on their declarative knowledge, which they seemed to have acquired and/or developed through the experience of WL, rather than in the grammar production tests, resulting in a significant difference only in terms of obligatory contexts, but not in terms of accuracy (i.e., points per context). Similar results were found in the aforementioned Brooks et al.'s (2010) OL study, whose two participants performed significantly better on a written limited production test, which was similar to the grammar production test in this thesis, than on an oral production test (stimulated recall), which is likely to

correspond to the essay test here, in that both are open-ended production tests despite the difference in modality. Interpreting the different outcomes, Brooks et al. also point to the difference in the nature of the two tests as a possible cause. (The gap regarding the two scores, points per context and obligatory contexts, is discussed in more detail in the next section.)

Second, related to the difference in the nature of the two tests, the results could be attributed to a difference between the participants' amounts of declarative versus procedural knowledge. According to DeKeyser (2007), learners initially acquire declarative knowledge. Then, they advance to the next stage, i.e., "the stage of acting on this knowledge, turning it into behaviour … or, in more technical terms, turning declarative knowledge into procedural knowledge" (p. 98). As such, the participants with declarative knowledge of the target construction, but without procedural knowledge or whose knowledge was not fully proceduralised, might have performed well on the grammar production tests, but not on the essay tests which required greater reliance on procedural knowledge. If this speculation is correct, the results appear to indicate the facilitative nature of WL in terms of the acquisition of declarative knowledge but not, and not surprisingly, of procedural knowledge. As stated earlier, the treatment in the present thesis only lasted for five minutes. Considering that a large amount of practice is necessary to proceduralise (and further automatize) one's knowledge (DeKeyser, 2007), the five-minute treatment could not have been enough for proceduralisation, let alone automatization, of one's knowledge. In addition, given that "time pressure makes the use of explicit knowledge harder" (DeKeyser, 2003, p. 326), the time limit of 20 minutes for essay writing, although deemed appropriate, might have made it harder for the participants, especially for the ones of lower-proficiency, to draw on their explicit knowledge.

As reported above, however, the findings regarding the two scores of the essay tests were inconsistent as well, which is the focus of the next section.

8.1.4 Obligatory Contexts (Meaning) vs Points per Context (Form)

As mentioned, the essay tests produced different results with respect to the two measures. On the one hand, the +WL group demonstrated statistically significant gains compared to the –WL group in terms of obligatory contexts, i.e., a possible indicator of improvement with respect to meaning as they reflect attempts at the target construction. On the other hand, no significant difference was observed between the two groups regarding points per context, an indicator of development in terms of the accurate use of the target construction from a form-oriented approach, implying that WL had a significant impact on meaning, but probably not on form/accuracy.

In terms of the favourable results of the +WL group regarding the results for the obligatory contexts, two explanations seem to be available. First, the significantly higher number of obligatory contexts may be attributable to the +WL participants' higher rate of focus on the target construction. As stated earlier in section 8.1.1, having engaged in WL is likely to have facilitated the participants to notice the target construction. This speculation is in line with the aforementioned questionnaire results which observed that a much higher ratio (63.6%) of the +WL participants reported that they focused on the target construction than the –WL participants (38.7%) (Part I, Q4). Reflecting this difference in the participants' focus on the target construction, in response to the question on their perceptions of learning, on the questionnaire (Part I, Q3), 11 of the +WL participants (33.3%) wrote the target construction, in addition to four participants (12.1%) who wrote the grammar rule (45.4% in total), whereas the number of –WL participants who pointed to the target construction was five (16.1%),

together with one (3.2%) who wrote the grammar rule (19.3% in total). If the speculation that WL facilitates learners to direct their attention to the target construction is correct, the potential effect of WL on learning is of great significance, in that attention is regarded as a prerequisite for L2 learning to occur (Leow, 2015; Schmidt, 1990).

Second, as stated earlier, the obligatory contexts were investigated as they might indicate the participants' improvement with respect to the acquisition of the meaning associated with the target construction. In the aforementioned study, which examined the acquisition of temporal expressions in second language from a meaning-oriented approach (as opposed to a form-oriented approach), Bardovi-Harlig (2000) states that learners are expected to acquire temporal expressions through three main stages of development, i.e., pragmatics, lexical and morphological. Regarding this point, she explains that "learners develop a functional, and often rich, means of temporal expression before the acquisition of verbal morphology" (p. 88). Referring to the central role played by time adverbials and lexical expressions in the second stage of temporal development, she adds that "the use of lexical devices to mark temporal expression is the defining characteristic of the stage" (pp. 88–89). In light of her explanation, participants' writing such as "If I travel back to the past, I meet Michael Jackson," which was not given any point in terms of accuracy (i.e., points per context was zero) but counted as one obligatory context as an indicator of an attempt at the target construction, may indicate that they were in the lexical stage of development, preceding the morphological stage. This interpretation seems to run in parallel with the distinction of meaning and form, respectively. That is, participants are likely to have developed their learning in terms of meaning, but not yet of form.

Meanwhile, with respect to the results for points per context, where neither the +WL group nor the –WL group demonstrated much improvement in terms of accuracy,

two interpretations seem to be possible. First, as stated already, given that the treatment of this thesis was only for five minutes, it must have been difficult for the +WL group to improve their accuracy significantly more than the –WL group, only with the facilitation of WL. Second, the result is likely to be due to the insufficient exposure to the target construction during the treatment, which is one of the shortcomings of this thesis. As explained earlier, the original text contained only three sentences on the target construction (31.3% of total words), which might have made it less likely for the participants to improve their accuracy regarding the target construction.

Finally, it is worth mentioning that in Yangas and Lado's (2012) reactivity study, they also reported different results depending on the types of measures used for the writing task. They found, however, positive reactivity in terms of accuracy, but not fluency. These conflicting findings may be explained by the differences in terms of task design and participants. That is, in contrast to the 5-minute WL activity, the participants in Yangas and Lado's study engaged in a think-aloud (i.e., OL) for 25 minutes while they worked on a story writing task. Probably even more influential was the fact that, unlike the participants in the current study who were lower-intermediate EFL learners with very little exposure to English, the participants in their study were heritage language learners, i.e., native speakers of the target language. They are, therefore, likely to have possessed fairly high level of procedural, and possibly declarative, knowledge to complete the task more accurately with the facilitation of 25 minutes of OL. (At the same time, given the characteristics of the participants, a ceiling effect might have contributed to the result in terms of fluency.) Despite these differences, the two studies seem to indicate two important points. First, both studies produced some evidence to prove that languaging, either OL or WL, benefits learners. Second, the findings of the

two studies indicate the importance of employing multiple measures in investigating treatment effects (Norris & Ortega, 2003).

8.1.5 Summary

Although previous WL studies have reported the positive impact of WL on L2 learning, no clear link has been identified (but see M. Ishikawa, 2018). In such a context, the results obtained in the present thesis have produced some evidence regarding the facilitative effect of WL on L2 learning, echoing the findings of previous studies on OL (e.g., Swain et al., 2009) and self-explaining (e.g., Chi et al., 1994) that identified a positive impact on learning. Also, the findings observed here lend support to Swain's (2006) statement that languaging includes both speaking and writing.

The analyses of the pre- and posttests scores, T-WLEs and questionnaire results indicate that the process of WL facilitated the +WL participants' learning, presumably enhancing noticing, in accordance with Swain's Output Hypothesis (2005) and Schmidt's (1990) Noticing Hypothesis. In addition, as hypothesised, WL is likely to have induced deeper processing (Craik & Lockhart, 1972), as reflected in the significant difference regarding the overall scores on the grammar production tests in the long term. Moreover, the generation effect (Slamecka & Graf, 1978) is likely to be another factor contributing to the favourable results for the +WL group.

In terms of the product of WL, although the current study did not provide the participants with sufficient opportunity to make use of their product of WL, some of the +WL participants still referred to the benefits of the product of WL as a source for further reflection on the exit questionnaire. Given this, in spite of the limited time, the product of WL might have facilitated their L2 learning to some degree, as hypothesised. It should be pointed out, however, that there is another possibility, i.e., these comments

might have been based on their previous experience. Whatever the case, these comments seem to imply that the participants were already aware of the heuristic nature of writing, benefiting from its process and product, utilising it as "a vehicle for learning" (J. Williams, 2012).

In addition, judging from the results of the three assessment tests (essay tests, grammar production tests, recognition tests), WL is likely to be effective for production tests, especially grammar production tests, which are likely to be addressed by relying mainly on explicit knowledge of the target construction. WL was hypothesised to help learners engage in an explicit learning process and the results obtained here seem to support the hypothesis. In terms of the essay tests, the +WL participants were found to have outperformed their –WL counterparts regarding their obligatory context scores, which suggests that WL helped the participants to direct their attention to the target construction, resulting in higher numbers of attempts. Meanwhile, no such difference was identified in terms of points per context scores, an accuracy measure. Given that engaging in WL for a short time is not likely to affect the development and/or acquisition of procedural knowledge, the result may not be conclusive.

Finally, echoing the findings of previous studies (e.g., M. Ishikawa, 2018; M. Ishikawa & W. Suzuki, 2016), no significant differences were identified regarding the recognition tests, which suggests that WL may not be as effective for tasks which involve input processing, i.e., an implicit process.

8. 2 Frequency of T-WLEs and L2 Learning (RQ2)

The second research question investigated the frequency of T-WLEs in relation to L2 learning (i.e., the gain scores on the three assessment tests, essay tests, grammar production tests, recognition tests). Similar to the results observed for the first research

question, the results of the correlation analyses varied considerably from test to test. Therefore, they are discussed by test below.

8.2.1 Correlations with Grammar Production Tests

In order to investigate the possible correlation between the frequency of T-WLEs and L2 learning, the gains in the three scores (i.e., accurate use score, overuse score and overall score) on the grammar production tests were examined. Once again, accurate use scores were calculated by simply totalling the points for correct answers on the target items, including overuse scores, whereas overuse scores were the totals where the participants used the past tense for simple-conditional distractor items. Finally, overall scores were obtained by subtracting the total of overuse scores from accurate use scores. Therefore, overall scores are a more accurate and reliable indicator of the participants' learning of the target construction than accurate use scores.

As reported, a Spearman's correlation analysis revealed a statistically significant correlation with the short-term gain in the overall score (OAG1) with a small effect size. As stated earlier, the grammar production tests were designed to assess learners' productive knowledge of the meaning-to-form mapping in a highly controlled written context by measuring mainly the participants' explicit knowledge and/or accurate understanding of the target construction. Judging from the result, focusing more on the target linguistic feature by writing about it with a higher frequency seems to have enabled the participants to deepen their understanding, resulting in the acquisition of explicit knowledge, at least with respect to the short-term overall gain score. This conjecture is consistent with the claim of Schmidt's (1990) Noticing Hypothesis, which assumes that "more noticing leads to more learning" (1994, p. 129). It is also consistent with the findings of self-explaining studies (e.g., Chi et al., 1994; Muñoz et al., 2006),

which have demonstrated a correlation between the number of self-explanations and the amount of learning.

In contrast to the overall scores, no statistically significant correlations were identified with either accurate use or overuse scores. The following two accounts may explain the outcome with respect to accurate use scores. First, there was a tendency of overuse of the target construction, i.e., to use the past tense for all the if-clauses, both counterfactual and simple conditional, by some participants. Therefore, as accurate use scores were the total points for correct answers on the target items, including overuse scores, this tendency might have masked possibly significant correlations. Second, the result may be due to a ceiling effect. As stated, in order to identify changes in terms of overuse and overall scores, the participants whose pretest accurate use score was 16, the maximum possible, were not excluded from the data when they demonstrated overuse of the target construction and their overall score did not exceed the cut-off line of 12 (75.0%). Thus, some of the participants started out with a full score for accurate use, which could have blurred potential significant correlations between accurate use gain scores and the frequency of T-WLEs. Put differently, these accounts seem to suggest that the overall score is a more sensitive measure than accurate use or overuse scores.

8.2.2 Correlations with Recognition Tests

In terms of the recognition tests, which examined the participants' knowledge of meaning associated with the present counterfactual conditional, a statistically significant correlation in the range of small was identified with the short-term gain score. A possible interpretation of this finding is that, like the grammar production tests, focusing on the target construction (and writing about it more frequently) might have had a

sufficiently positive impact on the participants to help in the recognition tests, which involve receptive knowledge of the form-to-meaning mapping, at least in the short term.

8.2.3 Correlations with Essay Tests

As reported already, no statistically significant correlations were identified between the frequency of T-WLEs and L2 learning with respect to the essay tests. These results mean, unlike the two grammar tests (production tests, recognition tests), that those +WL participants who produced T-WLEs with higher frequency did not necessarily achieve higher gains on the essay tests. As it seemed reasonable to expect some kind of correlation as observed in terms of the two grammar tests, especially the grammar production tests, the results were rather unexpected.

Having said that, as stated in section 8.1.3, these results are likely to be conducive to the more demanding nature of the essay tests than the grammar production tests. Namely, although both are production tests, compared to the highly-controlled grammar production tests which could have been completed by largely relying on declarative knowledge, the essay tests were open-ended and designed to measure the participants' productive knowledge of how to use the target construction in a more natural context. Thus, the participants had to think about and plan their answers (i.e., meaning), as well as pay attention to forms and vocabulary items to write essays within the time limit of 20 minutes, which is likely to have required more extensive use of the participants' procedural knowledge. As such, even when they focused on the target construction, resulting in the production of more T-WLEs, they might have drawn less on their explicit declarative knowledge than in the grammar production tests.

8.2.4 Summary

To summarise, as was the case with the results for RQ1, the relationship between the frequency of T-WLEs and L2 learning varied across the three assessment tests. The two grammar tests demonstrated one statistically significant correlation each with the frequency of T-WLEs in the short term. The results indicate that producing T-WLEs with higher frequency was beneficial for the two tests, which could have been carried out by drawing on mainly declarative knowledge, at least in the short term. In contrast, no link between the frequency of T-WLEs and L2 learning was identified regarding the essay tests, suggesting that just producing T-WLEs more frequently was unlikely to have been sufficient to carry out the essay task with more syntactic accuracy, probably because of the more demanding nature of the essay tests than the two grammar tests. In other words, WL seems to be effective in facilitating the acquisition of explicit knowledge, as hypothesised. Meanwhile, WL is not likely to be as effective for tasks which require procedural knowledge, at least in the operationalization of WL in the present thesis (only a 5-minute treatment).

8.3 Quality of T-WLEs and L2 Learning (RQ3)

The third research question investigated the relationship between the quality of T-WLEs and L2 learning. As already mentioned, the quality of T-WLEs was operationalised by categorising the participants from zero to three based on their level of awareness as reflected in their T-WLEs. The quality of T-WLEs was analysed in relation to L2 learning measured by the gain scores on the three assessment tests (i.e., essay tests, grammar production tests, recognition tests), separately. Being consistent with the results observed for the previous research questions, the three tests revealed considerably different results from each other. They are considered, therefore, by test in the following sections.

8.3.1 Correlations with Grammar Production Tests

Following the procedures for the correlation analyses for the frequency of T-WLEs, the quality of T-WLEs was analysed in relation to the gains in the three scores (i.e., accurate use score, overuse score and overall score) on the grammar production tests. Echoing the results of the correlation analyses of the frequency of T-WLEs, Spearman's correlation analyses detected statistically significant correlations for the overall gain scores. It is important to note, however, that significant correlations were observed with both short- and long-term overall gain scores (OAG1, OAG2), whereas it was only with the short-term gain of the overall score (OAG1) for the frequency of T-WLEs. What is more, the size of the correlations turned out to be larger, in the range of medium, whereas it was small for the frequency of T-WLEs.

As noted already, the grammar production tests were included as a measure of learners' knowledge of the target construction in a highly-controlled written environment, which is likely to have required mainly explicit knowledge and/or accurate understanding of the target construction. Therefore, the stronger and higher number of statistically significant correlations between the gain scores and the quality of T-WLES, as compared to the frequency of T-WLEs, means that the participants needed to have a higher level of noticing and/or understanding, as demonstrated by the grammar rule and explicit explanation of the target construction, rather than a higher frequency of focus on the target construction, in order to achieve higher gain scores on the grammar production tests, especially in the long term.

Once again, this reasoning seems to support Schmidt's (1990) Noticing Hypothesis, which claims that a higher level of awareness, i.e., understanding, promotes deeper learning than a lower level of awareness, i.e., noticing. In addition, the results resemble those of Qi and Lapkin (2001), who reported that a higher level of noticing as

measured by participants' think-alouds (i.e., OL) resulted in greater learning than a lower level of noticing, i.e., what they termed "substantive" and "perfunctory" noticing, respectively. The results also suggest that the findings of OL can be applied to WL as well, supporting Swain's (2006) claim.

Finally, echoing the outcomes of the analyses of the frequency of T-WLEs, no statistically significant correlations were identified in terms of the accurate use or overuse scores for the quality of T-WLEs, either. The two accounts mentioned earlier may also explain these inconsistent findings, at least in terms of accurate use scores, i.e., a tendency to overuse the target construction and a ceiling effect. Also, as stated earlier, the results suggest that overall score is a better predictor of learning than the other scores.

8.3.2 Correlations with Recognition Tests

Similar to the findings of the analyses of the frequency of T-WLEs, the correlation analyses of their quality and gains on the recognition tests detected one statistically significant correlation, also in the range of small. It should be pointed out, however, that a significant correlation was identified with the long-term gain score for quality, whereas it was with the short-term gain score for frequency. That is to say, the frequency with which learners focused on the target construction predicted their gains in the short term, while their level of awareness and/or understanding was a better predictor in the long term. This is not surprising since, according to Craik and Lockhart (1972), the depth of processing with which learners encode information influences the strength of resulting long-term memory representations. As such, deeper processing is assumed to result in more durable and stronger memory representations than shallow processing.

Although the difference was observed for time, compared to the grammar production tests, the frequency and quality of T-WLEs predicted learners' gains similarly in terms of the recognition tests. These inconsistent findings may again be explained by the difference in the nature of the two tests. Namely, unlike the grammar production tests, just focusing on the target construction (and writing about it more frequently) might have had sufficient positive impact on the participants to help in the recognition tests. Along the same lines, according to Cook (1991), "the ability to *decode* language, that is, the ability to understand the meaning conveyed by a particular sentence, is not the same as *code breaking*, that is, discovering the linguistic systems which carry that meaning" (p. 375). Given this, to borrow Cook's terms, "decoding" might have been sufficient to address the recognition tests, which were designed to gauge the participants' receptive knowledge (form-to-meaning mapping) as reflected in the similar results regarding the correlation between frequency and quality in relation to L2 learning. Meanwhile, "code breaking," i.e., their productive knowledge of meaning-to-form mapping, seems to have been necessary to carry out the grammar production tests, which are likely to require explicit processing.

8.3.3 Correlations with Essay Tests

Replicating the results of the correlation analyses with frequency, the quality of T-WLEs demonstrated no statistically significant correlations with gains on the essay tests. These findings imply that not only writing more frequently about the target construction, i.e., producing more T-WLEs, but also showing more noticing and/or understanding, i.e., producing higher quality T-WLEs, does not have a significant positive impact on L2 learning when it comes to essay writing. Given that previous OL research suggests that the quality of OL affects the amount of learning, with a higher

quality of OL benefiting learners more (e.g., Knouzi et al., 2010; Swain et al., 2009), the results were rather unexpected.

That said, as discussed in previous sections (8.1.3 and 8.2.3), these results are likely to be attributable to the more demanding nature of the essay tests than the two grammar tests. That is, on the one hand, the grammar tests are likely to have been completed by relying only on declarative knowledge, demonstrating some significant correlations with the quality of T-WLEs as a consequence. On the other hand, in order to address the essay tests, both declarative and procedural knowledge were likely to have been necessary because the participants needed to consider morphosyntactic issues to a greater degree, in addition to lexical and semantic factors. The result that no significant correlations were found in terms of the essay tests appears to suggest that WL may not play a facilitative role in the case of tasks where procedural knowledge is required. It should however be pointed out that, as already mentioned, the WL in this thesis was only for five minutes, which was not likely to have facilitated development in procedural knowledge.

8.3.4 Summary

In summary, echoing the results for RQ2, the correlations between the quality of T-WLEs and L2 learning differed from test to test. In terms of the two grammar tests (especially the grammar production tests), it was found that showing a higher level of awareness and/or understanding of the target construction predicted greater gains, as reflected in the stronger and greater number of correlations than just focusing more on the target linguistic feature by writing about it with a higher frequency. The results indicate a more facilitative impact for the quality of WLEs on L2 learning than their frequency, in terms of the grammar production tests, supporting previous findings on

OL (e.g., Qi & Lapkin, 2001). In contrast to the two grammar tests, no significant correlations were detected regarding the essay tests, probably because of their more demanding nature in terms of semantic and syntactic processing.

These results suggest that the facilitative effect of WL on learning varies depending on the nature of tests/tasks. More specifically, WL appears to be more effective for tasks that are likely to be carried out by relying mainly on explicit declarative knowledge than those which involve both declarative and procedural knowledge. As stated several times already, however, given that the treatment in this thesis only lasted for five minutes, different outcomes may be expected when sufficient time is allotted to WL.

8.4 Correlation between Frequency and Quality of T-WLEs (RQ4)

8.4.1 Frequency and Quality of T-WLEs

As reported, the analysis detected a statistically significant correlation between the frequency and quality of T-WLEs, with a large effect size. It means that the more frequently the participants focused on and wrote about the target construction (i.e., produced higher numbers of T-WLEs), the higher awareness and/or understanding of the construction they demonstrated/achieved. Judging from this result, it may be argued that engaging in WL and writing about the target construction more frequently might have enabled the +WL participants to become more aware of and/or improve their understanding of the target construction during the treatment, resulting in the production of higher quality T-WLEs.

The T-WLEs produced by the four case study participants reported above seem to be a good example to support the above claim that the more T-WLEs learners produce, the greater learning they achieve (see Table 7.6 in section 7.2). That is, as discussed,

Yuta, who produced three T-WLEs, started out with a T-WLE which was categorised as noticing ("Oh, past tense"), demonstrating partial understanding ("past because of if") in his second T-WLE and showing full understanding by writing the grammar rule of the target construction in his third T-WLE. Similarly, the first two T-WLEs of Rieko, another participant who produced three T-WLEs, showed noticing. Although she did not attain full understanding, her third T-WLE, a partial translation of the target sentence, which was categorised as partial understanding, showed some improvement. Meanwhile, Hiroshi, who claimed that he could think in his head without writing, produced one T-WLE, which was categorised as noticing. (Takuya did not produce any T-WLEs, thereby categorised as zero in terms of both frequency and quality.)

A similar relationship was found in Swain et al.'s (2009) aforementioned study, where language units (LUs) (i.e., OL) on the target concept produced by high-languagers (i.e., participants who produced greater numbers of LUs) were found to be qualitatively different from those produced by low- or middle-langaugers (i.e., participants who produced fewer or average numbers of LUs). To be more precise, the high-languagers were found to have produced more inferencing and self-assessment LUs, i.e., the types of LUs that are expected to facilitate learning more than other types, such as paraphrasing and rereading, which require only surface-level processing and are not as likely to facilitate learning (Chi et al., 1994; McNamara, 2004). The differences in the quality of LUs seem to be comparable to the level of the quality of the T-WLEs in the current thesis, with a higher quality of T-WLEs (i.e., noticing with reason) corresponding to inferencing and self-assessment, and a lower quality of T-WLEs (i.e., noticing) corresponding to paraphrasing and rereading. If this interpretation is correct, echoing the findings of Swain et al., it may be claimed that the higher the frequency with which learners engage in languaging and produce language (LUs or WLEs), the

higher quality their noticing and/or understanding becomes, resulting in greater learning, regardless of its modality. Put differently, not just OL but also WL may be claimed to facilitate learning to a greater degree when learners produce more language, again supporting Swain's (2006) claim that languaging includes both speaking and writing.

8.4.2 Summary

To summarise, a significant link between the frequency and quality of T-WLEs was identified, indicating that learners achieve a higher level of understanding by engaging in WL and producing more T-WLEs, supporting the idea that languaging mediates thinking, functioning as a learning tool (Swain, 2006, 2010).

It should be kept in mind, however, that not having produced any T-WLEs does not necessarily mean that the participants did not notice or pay attention to the target construction. Although it was not very successful in Hiroshi's case, learners can think internally without producing language externally. Regarding this point, in her aforementioned study which used learning journals as one of the research tools to investigate learners' noticing of feedback, Mackey (2006) aptly stated that a lack of evidence of noticing or attention does not necessarily mean that attention or noticing is not present, emphasizing that "absence of evidence is not the same thing as evidence of absence" (p. 409). Similarly, operationally defining noticing as availability for verbal reporting, Schmidt (1990) warned that "the lack of a verbal report cannot be taken as evidence of failure to notice" (p. 132), referring to memory limitations and lack of metalanguage knowledge as possible factors that may influence the availability of verbal reporting.

8. 5 Aptitude and WL (RQ5)

The fifth research question asked about the association between learners' language aptitude, i.e., grammatical sensitivity and inductive language learning ability, and the effect of WL on L2 learning in order to identify the aptitude profiles of learners who are likely to benefit from WL. As stated, WL requires learners to solve their linguistic issues on their own, presumably using their own resources, such as the two aptitude abilities. Therefore, although it was the first attempt to address this issue, these two aptitude constructs were expected to correlate closely with the +WL participants' learning. Nonetheless, of all the correlations between the scores on the three aptitude tests and the learning outcomes (i.e., the gain scores on the three assessment tests, essay tests, grammar production tests and recognition tests) for the +WL group, only one out of 36 (2.8%) was statistically significant, with a medium effect size. In sharp contrast, in terms of the –WL group, 11 out of 36 (30.6%) correlations turned out to be statistically significant, six (54.5%) being in the range of medium and the rest (five, 45.5%) small. In what follows, the findings are discussed in detail.

8.5.1 WL, an External Equalizer?

As reported, the –WL group demonstrated a greater number of significant correlations with aptitude constructs than the +WL group in general. To review the results briefly, the most striking differences between the two groups were observed in terms of the two grammar tests. In terms of the grammar production tests, seven out of 18 (38.9 %) of the total number of correlations were detected to be statistically significant for the –WL group across all aptitude tests (i.e., the MALT, LABJ, LLAMA), two with medium effect sizes and the rest with small effect sizes. Of the seven

correlations, six were with long-term gains. In contrast, the +WL group demonstrated merely one medium-sized significant correlation out of 18 (5.6 %), between the MLAT and the long-term overall gain score. Similarly, with respect to the recognition tests, exactly half of the total number of correlations (three out of six, one for each aptitude test) were identified to be statistically significant, all in the range of medium for the –WL group. One was with grammatical sensitivity (measured by the MLAT) and the short-term gain score, while the other two were with inductive language learning ability (measured by the LABJ and LLAMA) and long-term gain scores. No significant correlations were detected for the +WL group. Finally, here as well, the essay tests produced rather different results from the two grammar tests. The analyses revealed only one statistically significant correlation with grammatical sensitivity in the range of small for the –WL group, and none for the +WL group.

Taken together, the –WL group demonstrated a higher number of significant correlations than the +WL group across all aptitude measures, especially regarding the two grammar tests. What is more, of the 11 significant correlations for the –WL group, eight (72.2%) were observed with long-term gains, suggesting that the –WL participants needed to have higher grammatical sensitivity and inductive language learning ability to achieve higher gains, especially in the long term. (The one significant correlation found for the +WL group was also in the long term.)

As stated, compared to the –WL group, the +WL group demonstrated far fewer statistically significant correlations between their aptitude and L2 learning, thus not indicating any clear relationships. In other words, the +WL participants achieved similar gains regardless of their aptitude. Although the results were rather unexpected, the previous studies on ATI reviewed earlier seem to account for this outcome (e.g., Erlam, 2005; Li, 2013; Sheen, 2007a, 2007b; Stefanou & Révész, 2015; Trofimovich et al.,

2007; Y. Yilmaz, 2013). That is, as evidenced by ATI research, the impact of aptitude on learning has been identified to differ depending on the nature of instruction (i.e., the degree of implicitness/explicitness). In addition, as mentioned earlier, according to Li (2013), the effects of language aptitude can also vary in relation to the difficulty of target constructions (see Skehan, 2015, for a similar discussion). Therefore, these two factors, i.e., the nature of the treatment and the difficulty of the target construction, are likely to have contributed to the result in the present thesis.

In terms of the nature of the treatment, as mentioned, previous studies have produced mixed findings. That is, some studies have reported fewer statistically significant correlations between aptitude and explicit instruction than implicit instruction, suggesting that explicit instruction can minimise individual differences in language aptitude (e.g., Erlam, 2005; Li, 2013; Stefanou & Révész, 2015; Trofimovich et al., 2007). Others, however, have revealed more significant correlations between explicit instruction and aptitude abilities (e.g., Sheen 2007a, 2007b; Y. Yilmaz, 2013). In the case of the present thesis, the experience of WL, a conscious process (DiCamilla & Lantolf, 1994), might have made the nature of the treatment more explicit for the +WL participants than for the –WL participants, presumably enabling the +WL participants to acquire and/or consolidate "internally-derived explicit knowledge" (Roehr-Brackin, 2014) and minimise the individual differences in language aptitude of the +WL participants. Meanwhile, the acquisition of such explicit knowledge may have been difficult for the –WL participants who engaged in potentially silent languaging without a chance to engage in WL activity.

According to R. Ellis (2009), explicit knowledge can function as a tool that learners can utilise to mediate their performance when they are faced with linguistically challenging issues. Given this, the experience of WL might be argued to have equipped

the +WL participants with what R. Ellis calls “a tool,” thus allowing them to benefit from the treatment regardless of their aptitude profiles. In contrast, in terms of the –WL group, who had no access to WL, only the participants with higher aptitude are likely to have achieved gains from the treatment. If this interpretation is correct, it seems to account for the noticeably higher number of statistically significant correlations between the aptitude tests scores and the gain scores observed for the –WL group than the +WL group in terms of the two grammar tests, which are likely to have required more explicit knowledge than the essay tests.

This line of reasoning appears to be consistent with the claim of Cronbach (1967), an advocate of ATI, who suggested erasing individual differences as one way to adapt instructions to individual differences. Referring to the remedial instruction often employed by schools to minimise individual differences for practical reasons, i.e., to offer same instruction without any alterations, Cronbach argued that one of the ways for schools to adapt to individual differences is to employ differentiated instructional techniques according to learners’ individual differences. Given the findings of the current thesis, i.e., far fewer statistically significant correlations between aptitude and gain scores for the +WL group than the –WL group, WL can be said to have succeeded in erasing individual differences, functioning as a remedial treatment for the +WL participants with lower aptitude. Put differently, WL could be an instructional technique to reduce and/or adapt to individual differences for learners with lower aptitude, as suggested by Cronbach.

Meanwhile, as stated earlier, the difficulty of the target constructions has been suggested as another factor that influences the impact of aptitude (Li, 2013; Skehan, 2015). According to Li’s (2013) hypothesis, the target forms used in studies that observed fewer correlations between explicit instruction and language aptitude (e.g., Li,

2013; Trofimovich et al., 2007) were simpler than those used in studies that produced opposing findings (e.g., Sheen 2007a, 2007b; Y. Yilmaz, 2013). In his view, when target forms are simpler, learners can address tasks with explicit instruction alone, without resorting to their higher aptitude, resulting in no/fewer significant correlations between explicit instruction and aptitude. Conversely, when target forms are more difficult, learners probably need both explicit instruction and higher aptitude, resulting in significant correlations between these two variables (see Yalçin & Spada, 2016, for the relationship between explicit instruction and the difficulty of target structures). It is worth mentioning that for the participants with lower aptitude, even explicit instruction is not likely to be sufficient to achieve gains, resulting in no significant correlations (Sheen, 2007a, 2007b; Y. Yilmaz, 2013).

Related to the difficulty of target constructions, although not pointed out by Li (2013), prior knowledge seems to be another factor that influences the impact of aptitude. As mentioned already, the same target construction, i.e., English articles, was employed in the studies by Stefanou and Révész (2015) and Sheen (2007a, 2007b), which produced different findings. Among other things, the difference in the participants' familiarity with the target form seems to have contributed to the conflicting findings. Namely, the participants in Stefanou and Révész's study were EFL high-school students who had some level of knowledge of the target, whereas the ones in Sheen's studies were ESL learners who were not explicitly taught the target. Thus, not only the difficulty of the target constructions but also learners' familiarity with the constructions are likely to be influential factors to be considered. In addition, it should be pointed out that these studies looked into different meanings associated with articles. In Sheen's studies, learners' use of definite and indefinite articles (i.e., *a* as a first mention and *the* as an anaphoric reference) were examined, whereas Stefanou and Révész (2015)

investigated the use of specific and generic plural referents (i.e., definite plural, bare plural, demonstrative plural). Thus, Sheen's target might have been an easier target to learn.

In the case of the present thesis, the target construction, i.e., the present counterfactual conditional, was expected to be problematic for the participants because of its syntactic and semantic complexity as well as L1-L2 difference. That said, the target feature was supposedly familiar to the participants. As the background questionnaire results showed, they had received more than seven years of English education on average, including six years of traditional grammar-oriented education at high-school, in an EFL setting. Therefore, as evidenced by the pretest scores, although their understanding was not solid, they generally had some level of knowledge regarding the target construction at the outset of the experiment, which seems to be a similar condition to the one in Stefanou and Révész's (2015) study. Considering these points, the target construction might not have been as challenging as expected due to the participants' prior knowledge of it, probably being within their processing capacity (Li, 2013).

Taken together, as suggested by Li (2013), the nature of the treatment (i.e., the supposedly more explicit nature of the treatment due to the experience of WL) and the difficulty of the target construction (i.e., less challenging than expected because of potential prior knowledge) are likely to have contributed to fewer significant correlations between aptitude and L2 learning for the +WL group than the –WL group, at least, with respect to the two grammar tests. Meanwhile, as reported, no clear differences were found between the two groups in terms of the essay tests, revealing only one significant correlation for the –WL group, implying that another factor other than the ones mentioned above may be at play to explain the outcome.

8.5.2 The Impact of WL, Aptitude, and Assessment Task Effects

Echoing the results of the previous correlation analyses regarding the frequency (RQ2) and quality (RQ3) of T-WLEs and L2 learning, the essay tests showed rather different results from the two grammar tests here as well, producing almost no significant correlations, neither for the +WL group nor for the –WL group. As noted earlier, this discrepancy seems to be attributable to the more demanding nature of the essay tests compared to the two grammar tests, in that the participants had to draw on not only their explicit knowledge regarding the target construction, but also on other skills, such as lexical knowledge and planning in terms of the content in an open-ended situation, i.e., procedural knowledge of the target construction. Therefore, the combination of high aptitude and explicit knowledge of the target construction that potentially derived from WL experience for the +WL participants (and presumably just high aptitude for the –WL participants) might not have been enough to achieve gains on the essay tests. In other words, even those participants with both high aptitude and explicit knowledge of the target construction might not have managed to achieve significant gains on the essay tests. If this interpretation is correct, although three factors, i.e., aptitude, treatment and target construction, have been suggested as influencing the learning outcome, *task type* seems to be another key factor to be considered in interpreting the impact of aptitude on L2 learning.

As stated above and in section 8.1.3, procedural knowledge seems to be a key contributing factor to no significant differences between the two groups in terms of the essay tests here as well. That is, the 5-minute treatment with WL appears to have facilitated the development of declarative knowledge, enabling the +WL participants to outperform their –WL counterparts concerning the two grammar tests. Meanwhile, it is highly unlikely that the learners were able to proceduralise or further automatize their

declarative knowledge in such a short time regardless of the opportunity of WL, resulting in no significant differences between the two groups.

This speculation seems to be in line with Skehan's (2002) aptitude profile model, which explains SLA processing stages in relation to the potential aptitude components that each stage involves (see Table 3.1 in section 3.1.1 for details). Given that a greater number of significant correlations were observed for the –WL group than the +WL group between the gains of the two grammar tests and the two aptitude constructs (i.e., grammatical sensitivity and inductive language learning ability), it would appear that having engaged in WL compensated for these two abilities. According to Skehan, the two components are crucial in the earlier stages of development where "the capacity to detect and manipulate patterns in the target language" (p. 91) (i.e., patterning) is involved (Stages 2–5 out of the nine stages in his criteria). In contrast, automatization and proceduralisation, which were expected to be necessary for essay writing tasks, are predicted to play a greater role in later stages in the development (i.e., controlling) (Stage 6: becoming accurate, avoiding errors, and Stage 8: automatizing rule-based language, achieving fluency). As such, WL does not seem to have been very influential on learning in this stage, at least in the operationalization of WL in the current thesis (i.e., five-minute WL treatment).

8.5.3 Summary

The correlation between aptitude and L2 learning was examined in order to identify aptitude profiles that are likely to benefit the most from WL, with the aim of informing pedagogical decisions. The correlation analyses, however, detected generally fewer statistically significant correlations for the +WL group compared to the –WL group, which seems to indicate two important points. First and foremost, WL seems to benefit all learners irrespective of their aptitude profiles. More precisely, WL is likely to

benefit learners with lower aptitude more, as it is likely to minimise the effects of aptitude. This speculation may be explained by the heuristic nature of writing as opposed to speaking. That is, writing usually allows learners to plan their output as well as to process input under less time pressure. Moreover, they are expected to benefit from the product of writing, which functions as external memory and a further source of reflection, enabling them to reflect on and build on them (J. Williams, 2008, 2012). These facilitative characteristics of writing are likely to carry over to WL (W. Suzuki, 2012). Therefore, WL seems to facilitate learners' thinking and learning processes, benefiting even learners with lower aptitude/ability as an external equaliser. Although aptitude and learners' proficiency levels are not necessarily equivalent, the results obtained here seem to echo the findings of previous self-explaining studies which have reported that learners with lower proficiency benefit more from the facilitation of self-explaining than learners with higher proficiency, who probably do not need such facilitation (e.g., McNamara, 2004, 2017).

Second, the present thesis produced some evidence that aptitude and treatment interact, supporting the findings of previous ATI research (e.g., Erlam, 2005; Li, 2013; Stefanou & Révész, 2015; Trofimovich et al., 2007). The results obtained here also lend support to Li's (2013) hypothesis that the interaction between aptitude and treatment can differ depending on the difficulty of target structures. Related to this point, given the findings of this thesis and of previous research (Sheen, 2007a, 2007b: Stefanou & Révész, 2015), prior knowledge of target structures seems to be another influential factor. In addition, judging from the discrepancies observed among the three tests (essay tests, grammar production tests and recognition tests), the nature of the task is likely to remain another key to be considered.

8. 6 Metalanguage Knowledge and WL (RQ6)

RQ6 sought to identify the possible association between metalanguage knowledge and the effect of WL on L2 learning. As stated, WL was defined as "any language noted by learners to reflect on their language use, *with or without metalinguistic terminology*" and the use of metalanguage was not required for WL. That said, a higher level of metalanguage knowledge was expected to be influential in engaging in the activity of WL, which involves the externalisation of thoughts regarding linguistic issues by writing. Therefore, significant correlations between the metalanguage knowledge of the +WL participants and their L2 learning was expected. Nonetheless, echoing the results of the correlation analyses with aptitude (RQ5), a higher number of statistically significant correlations were identified for the –WL group rather than the +WL group. Namely, as for the +WL group, three statistically significant correlations were observed with the gains on the two grammar tests (two for grammar production tests and one for recognition tests) with small to medium effect sizes (25.0%, three out of 12). Meanwhile, the –WL group showed five significant correlations, four with the grammar production tests and one with the recognition tests, all in the range of medium (41.7%, five out of 12). No statistically significant correlations were detected in terms of the essay tests for either of the groups. The only difference between the two groups, therefore, lies in the grammar production tests. As the results of the correlation analyses again varied greatly from test to test, they are discussed by test below.

8.6.1 Grammar Production Tests

According to R. Ellis (2004, 2009), metalanguage knowledge is a constituent part of explicit knowledge along with analysed knowledge. Considering that the grammar production tests were designed to measure explicit knowledge of the target construction,

and that WL was hypothesized to facilitate explicit learning, it seemed natural to hypothesise that those +WL participants with sufficient metalanguage knowledge would be able to write their thoughts more easily than those without it, thereby showing significant correlations. In line with this thinking, Elder (2009), who evaluated the validity of a test of metalinguistic knowledge, used basic metalinguistic terminology, such as clause and verb, in her metalinguistic knowledge test, partly because "it is often difficult to formulate explanations without the use of such terms" (p. 118).

As reported, the analyses for the –WL group demonstrated four medium-sized statistically significant correlations (four out of six, 66.7%) with the accurate use and overall gain scores in both the short and long term, whereas the +WL group showed half such significant correlations (two out of six, 33.3%) with both the long-term gains in overuse and overall scores in the range of medium and small, respectively. These results (i.e., the higher number of significant correlations and their larger magnitude of effect in terms of the overall gain scores, the best indicator of learning of the three scores) suggest that those –WL participants who did not engage in WL generally needed to have a higher level of metalanguage knowledge in order to achieve higher gains compared to the +WL participants. Simply put, the +WL participants achieved similar gains regardless of their level of metalanguage knowledge under the +WL condition (i.e., when they engaged in WL), at least in the short term. Given this, like the results of the correlation analyses between WL and language aptitude for RQ5, WL may be argued to have neutralised the impact of possible individual differences in metalanguage knowledge on L2 learning for the grammar production tests to a certain extent, again functioning as an external equaliser.

If the above speculation is correct, the same claim made to explain the results for RQ5 might be applied here as well regarding the short-term gains. That is, having

engaged in WL in their own words, i.e., irrespective of the use of metalinguistic terminology, might have compensated for the potentially lower level of metalanguage knowledge of the +WL participants, levelling out their knowledge. In other words, the +WL participants might have benefited from the very experience of having engaged in WL, i.e., the externalisation of their thoughts on linguistic issues by writing, regardless of the level of their metalanguage knowledge. The results in this thesis also seem to be in line with the findings reported by Qi and Lapkin (2001) and Storch (2008), who speculated that learners might have benefited from the very experience of OL, as well as those from Sanz et al. (2009) and Yanguas and Lado (2012), who also state that the very experience of think-alouds might have contributed to its positive reactivity. Supposing this interpretation is correct, the aforementioned three factors induced by WL, i.e., enhanced noticing, deeper processing and a generation effect, might have been conducive to the better performance of the +WL group.

Meanwhile, the results obtained for the –WL group seem to support the perspectives of R. Ellis (2004, 2009) and Elder (2009), who state that explicit knowledge and metalanguage knowledge are likely to be correlated. Furthermore, the findings are consistent with those of Hu (2011), who identified a significant positive correlation between metalinguistic knowledge and metalanguage knowledge (see Ammar & Hassan, 2017, for a similar claim). In addition, they seem to be in line with Roehr-Brackin's (2015) perspective regarding the correlation between learners' levels of metalanguage knowledge and explicit learning difficulty. Namely, those –WL participants with lower metalanguage knowledge are likely to have found it more difficult to address the grammar production tests, which are likely to require mainly explicit knowledge, compared to those with higher metalanguage knowledge, at least in the short term.

8.6.2 Recognition Tests

In contrast to the grammar production tests, the recognition tests produced similar results for both groups, i.e., statistically significant correlations with a medium effect size were found between metalanguage knowledge and short-term gain for the –WL group and long-term gain for the +WL group. This is consistent with the results reported in section 8.1 (RQ1), where no significant difference was found between the +WL and –WL groups regarding their gain scores on the recognition tests. Given the similar results for the two groups, WL may not be as effective as it was for the grammar production tests, which are likely to involve largely explicit processing, in compensating for the lack of metalanguage knowledge when processing input, i.e., implicit processing.

8.6.3 Essay Tests

Echoing the results of the previous correlation analyses (RQ2 and RQ3), the essay tests demonstrated no statistically significant correlations between metalanguage knowledge and test gains for either the +WL group or the –WL group. As reported, similar results were found regarding the correlation analyses between aptitude constructs and the gain scores on the essay tests (RQ5), which detected only one statistically significant correlation for the –WL group. Judging from these findings, the same interpretations of the aptitude analyses may account for this outcome, too. That is, the essay tests involved the planning of content and paying attention to form as well as vocabulary items. Thus, their demanding nature might have exceeded the participants' attentional resources (VanPatten, 1990), not allowing them to employ their metalanguage knowledge regardless of their WL condition, leading to the same results (i.e., no significant correlations) for the two groups.

Another possibility is that, again as stated, because of their demanding nature, the essay tests were likely to have required procedural knowledge. Considering that metalanguage knowledge is part of explicit knowledge (R. Ellis, 2004, 2009), and that neither group demonstrated statistically significant correlations with metalanguage knowledge and gain scores, the participants are likely to have drawn less on their metalanguage knowledge in such a context or they might not have needed metalanguage knowledge to achieve gains, resulting in no significant correlations. This speculation seems to be in line with Elder's (2009) analogy of metalanguage knowledge with architecture, i.e., "the knowledge required to design or construct a building is clearly independent of the ability to label its parts" (p. 115). Like her analogy, the knowledge required to plan and write an essay seems to have been independent of the ability to label vocabulary items used in the essay, i.e., metalanguage knowledge.

8.6.4 Summary

Taken together, and echoing the previous findings for RQ5, a higher number of statistically significant correlations between metalanguage knowledge and pre- and posttests gains was observed for the –WL participants, suggesting that the experience of WL might have compensated for the lack of metalanguage knowledge of the +WL participants. Although it was hypothesised that metalanguage knowledge would have some effect on the quantity and quality of WL, further affecting its potentially facilitative impact, the results obtained here indicate that learners benefited from the very experience of WL, irrespective of their level of metalanguage knowledge. It should be pointed out, however, that a difference between the two groups was observed only in terms of the grammar production tests, especially in the short term. Therefore, the above claim can be extended neither to the recognition tests nor the essay tests, which are

likely to require less explicit processing and proceduralised knowledge, respectively. As stated earlier (RQ5), the positive impact of WL seems to vary depending on task type.

8.7 Frequency of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge (RQ7)

RQ7 was asked to investigate the frequency of T-WLEs in relation to aptitude and metalanguage knowledge in order to identify if producing T-WLEs with a higher frequency might be attributed to learners' individual differences. As reported, significant correlations were observed between the frequency of T-WLEs and grammatical sensitivity (measured by the MLAT), as well as metalanguage knowledge, in the range of medium and small, respectively. No significant correlations were found for inductive language learning ability (measured by the LABJ and LLAMA). The findings are discussed by construct, below.

Given that grammatical sensitivity is "the ability to recognize the grammatical functions of words (or other linguistic entities) in sentence structures" (Carroll, 1981, p. 105), it was hypothesized to be related to the frequency of T-WLEs. That is, in order to produce T-WLEs to a higher frequency, the participants were expected to possess the ability to focus on the target structure in the original text. Considering that there were only three target sentences in the original text (31.3% of total words), grammatical sensitivity should have been crucial in noticing the target construction. The observed medium-sized significant correlation between the frequency of T-WLEs and grammatical sensitivity seems to support the hypothesis.

Similarly, although in the range of small, a significant correlation was found between the frequency of T-WLEs and metalanguage knowledge. As stated, Elder (2009) used some simple metalinguistic terminology in her test of metalinguistic

knowledge because of the difficulty in preparing explanations without using such terms, which may apply to the frequency of T-WLEs. That is, the results suggest that having some level of metalanguage knowledge might have facilitated the production of T-WLEs. (The 17 items in the metalanguage knowledge test were also basic terms.)

In contrast to the above two abilities, no significant correlations were detected regarding inductive language learning ability (measured by the LABJ and LLAMA), which is "the ability to infer or induce the rules governing a set of language materials, given samples of language materials that permit such inferences" (Carroll, 1981, p. 105). The results suggest that this ability did not relate to the numbers of T-WLEs that participants produced. Considering that producing T-WLEs does not necessarily mean understanding, this outcome is not entirely surprising.

As mentioned, prior to the analyses for RQ7, the association between metalanguage knowledge and aptitude abilities was investigated. Although statistically significant correlations were detected between metalanguage knowledge and all the aptitude tests results (the MLAT, LABJ and LLAMA), the one with the MLAT (i.e., grammatical sensitivity) turned out to be the strongest, with a large effect size. As explained earlier, the MLAT was a multiple-choice type test, where the participants were not required to explain their decisions, i.e., they did not have to possess metalanguage knowledge. The outcome that showed a significant correlation despite the non-requirement for metalinguistic terminology seems to imply that having higher grammatical sensitivity facilitates learners acquiring metalanguage knowledge. It may be argued that the higher the grammatical sensitivity that learners possess, the more likely they are to learn grammar. As a result, they might also be more prone to acquiring related metalanguage knowledge, especially if they receive explicit instruction.

8. 8 Quality of T-WLEs and Individual Differences in Aptitude and Metalanguage Knowledge (RQ8)

RQ8 asked about the association between the quality of T-WLEs and learners' aptitude and metalanguage knowledge in order to discern if producing higher quality T-WLEs might be related to higher abilities. As stated, given that producing higher quality of T-WLEs is expected to involve factors such as noticing, understanding, and writing about the target construction, it was hypothesized that the participants would need all the abilities investigated (i.e., grammatical sensitivity, inductive language learning ability, and metalanguage knowledge), yielding significant correlations with all the tests results. In practice, only one significant correlation was identified, with grammatical sensitivity (measured by the MLAT). Interpreting this unexpected outcome, the following possibilities emerged.

First, although having no significant correlations with inductive language learning ability was rather unexpected, as stated under ATI, it might have been due to the participants' prior knowledge of the target construction. That is, as reported, the participants were university students with over seven years of formal English education on average. Therefore, as reflected in the pretest scores, they were considered to have some level of prior knowledge regarding the target construction, although their knowledge might not have been solid. Put differently, they were assumed to have declarative knowledge to some extent, even if they did not possess procedural knowledge. Given this, they might not have needed to rely on their ability to "infer or induce" the rule of the target construction, resulting in no significant correlations as a consequence.

Second, although metalanguage knowledge was also hypothesized to demonstrate a significant correlation with the quality of T-WLEs, this was not the case. That said,

given that metalanguage knowledge is considered to be “independent of grammatical knowledge per se” (Elder, 2009), the result might not be so surprising. In addition, as stated earlier, the quality of T-WLEs was operationalized with the level of awareness based on Schmidt’s (2001) distinction. Namely, regardless of the use of metalinguistic terminology, the T-WLEs that demonstrated a higher level of awareness, such as correct translations of the target sentences, were categorized as understanding, i.e., of the highest quality. Thus, the fact that a higher level of quality did not necessarily involve metalinguistic terminology might have contributed to the outcome.

Finally, the significant correlations with grammatical sensitivity observed here and under RQ7 indicate that the participants needed higher grammatical sensitivity in order to produce T-WLEs not only to a higher frequency, but also of a higher quality. As stated above, grammatical sensitivity is the ability to recognize grammatical functions (Carroll, 1981), which is likely to be necessary not just to focus on, but to attain a deeper understanding of, the target construction.

CHAPTER IX
CONCLUSION

In this final chapter, first, the main findings are briefly summarised, followed by a discussion of the theoretical and pedagogical implications of the results obtained. The chapter concludes with the limitations of the study and some possible directions for future research.

9.1 Summary

The present thesis has explored the potentially facilitative impact of WL on L2 learning and the associations between the frequency and quality of WL and L2 learning. Moreover, the effect of WL on L2 learning was examined in relation to learners' language aptitude and metalanguage knowledge. The main findings are summarised as follows.

To begin with, this thesis has produced some evidence to prove a direct link between WL and the learning of a target construction by comparing the test results with those of participants who did not engage in WL. It should be pointed out, however, that, of the three types of tests (i.e., essay tests, grammar production tests, and recognition tests), favourable results for the +WL group were only found for the production tests, especially the grammar production tests. The findings suggest that the impact of WL can vary depending on task/test types and that WL may be particularly effective for production tests depending on how much they require reliance on explicit knowledge.

Second, the frequency and quality of T-WLEs were found to have some effect on learning in terms of two grammar tests, but not essay tests, again showing different findings depending on test types. Compared to the frequency of T-WLEs, the quality of

T-WLEs was found to be more closely related to gains with respect to the grammar production tests. Furthermore, a significant correlation was identified between the frequency and quality of T-WLEs, indicating that the more learners write about their linguistic issues, the higher the understanding they achieve.

Third, although language aptitude was examined in order to identify aptitude profiles (the characteristics of learners) that were likely to benefit the most from WL, it was found to have equalised learners' individual differences in language aptitude, benefiting all of them regardless of their aptitude, presumably benefiting learners with lower aptitude more. Similarly, the correlation analyses revealed that WL levelled out individual differences in metalanguage knowledge to some degree. These findings seem to suggest that WL can be used as a form of instruction to mitigate individual differences (Cronbach, 1967). It should be pointed out, however, as with the first three research questions (RQ1, RQ2 and RQ3), the results were not uniform across all the tests. That is, WL was found to be most effective in terms of the two grammar tests, especially the grammar production tests, presumably compensating for the lack of explicit knowledge, but not the essay tests.

Finally, the frequency of T-WLEs was identified to be related to learners' individual differences in grammatical sensitivity and metalanguage knowledge, whereas the quality of T-WLEs only demonstrated a significant correlation with grammatical sensitivity. The results suggest that metalanguage knowledge is related to the number of T-WLEs, but grammatical sensitivity is related to both the number and quality of T-WLEs. Although no significant correlations were found concerning inductive language learning ability, the difficulty of the target structure and learners' prior knowledge might have contributed to the results.

9.2 Theoretical Implications

Several important theoretical implications emerged from the findings of the present thesis. First of all, the positive impact of WL on learning observed here lends support to Swain's (2006, 2010) claim that languaging facilitates language learning, being consistent with the findings of previous studies that reported a facilitative effect of languaging, as well as self-explaining, in both oral and written modalities. The observed positive effect of WL on learning seems to be attributable to the noticing function and the reflective function of languaging (the articulation of thoughts with language), as Swain's Output Hypothesis (2005) claims. Moreover, the findings lend support to Schmidt's (1990) Noticing Hypothesis in that noticing, possibly triggered by WL, is likely to have contributed to the greater learning of the +WL participants compared to their –WL counterparts. In addition, the results that showed greater correlation for quality than frequency of T-WLEs and learning support his statement that a higher level of awareness leads to greater learning.

Second, the results of this thesis are also consistent with those of self-explaining research (Chi, 2000). Given the responses to the exit questionnaire, such as "WL helped me to think deeper" and "I can remember if I write," WL is likely to have induced deeper processing (Craik & Lockhart, 1978), triggering a generation effect (Slamecka & Graff, 1972) and a self-explanation effect (Chi, 2000), as hypothesised. Although no self-explaining research has been conducted in the field of SLA yet, the findings obtained here seem to indicate that self-explaining may be a domain-independent learning strategy.

Third, the results observed in this thesis indicate that languaging benefits learners regardless of its modality, again supporting Swain (2006, 2010). Compared to research on OL, WL is still an underexplored area. However, given the heuristic nature of

writing, i.e., a slower pace than speaking and its product being an external memory (Emig, 1977; Luria, 1982, 1999; Manchón, 2011; J. Williams, 2008, 2012), WL seems to offer optimal conditions for learning that are exclusively inherent to writing. This is not an argument to claim that WL is more advantageous than OL. Rather, the findings seem to indicate that WL offers different language learning opportunities that might complement the effects of OL.

Finally, the fewer correlations between the results of the aptitude tests and pre- and posttests of the +WL participants than those of the –WL participants indicate that WL can erase individual differences in language aptitude, functioning as an external equaliser. In other words, WL may function as a remedial treatment for learners with lower aptitude. The findings support the notion of ATI research (Cronbach, 1967) and previous findings on aptitude research (e.g., Erlam, 2005; Stefanou & Révész, 2015; Trofimovich et al., 2007), also lending support to Li's (2013) claim that the difficulty of target constructions is another factor to influence the interaction to be considered. Moreover, the findings of the present thesis indicate that prior knowledge (task familiarity) and the nature of the task may be additional key factors to be kept in mind.

9.3 Pedagogical Implications

The findings of the present thesis have several important pedagogical implications. First of all, WL can be used in regular classes as part of daily practice or may be employed as part of assignments, following Negueruela (2008), with the use of pedagogical tools such as reflective journals (Simard et al., 2007) or learner portfolios (Antonek et al., 1997). As WL does not require interlocutors or any electrical devices (unless typed WL), instructors can encourage their learners to engage in WL anytime and anywhere, with only pen and paper. More time is not a guarantee of more learning,

but given Bowles and Leow's (2005) report that participants who addressed a task while verbalising spent significantly more time on task than a silent control group, assigning WL might contribute to an increase in learners' study time outside the classroom.

Second, although the treatment in the current study only lasted for five minutes, given that learners are expected to benefit from WL through its process and product (W. Suzuki, 2012), preparing tasks with multiple exposures to the products of WL, that would enable learners not only to reflect on but also build on their WL, is likely to increase the observed positive impact of WL on learning. In addition, considering that the facilitative impact of WL was most noticeable in terms of the grammar production tests, which require the extensive use of explicit knowledge, the maximum effect of WL might be expected with tasks that are likely to involve explicit knowledge.

Third, considering the result that WL neutralised learners' individual differences in language aptitude and metalanguage knowledge, WL is likely to benefit learners irrespective of their aptitude or metalanguage knowledge. To be more precise, learners with lower aptitude and/or with less metalanguage knowledge are likely to benefit more from WL than those with higher aptitude and/or sufficient metalanguage knowledge who do not need such facilitation. Therefore, WL might be useful as remedial instruction for those learners who may be in need of extra support.

Last but not least, it is worth mentioning that the product of WL is likely to benefit not only learners, but instructors as well. That is, the product of WL can be a precious source of information regarding learners' "language learning in action" (Swain, 2005, p. 479), which enables instructors to reflect and/or adjust their teaching to learners' needs and to prepare lessons that are level-appropriate to learners, contributing to effective "tomorrow's instruction" (Brooks et al., 2010, p. 106).

9.4 Limitations and Directions for Future Research

There are a number of limitations to be acknowledged and considered in future research. First of all, as repeatedly stated, the present thesis included only one 5-minute treatment for the +WL participants to engage in WL. Considering that WL is hypothesised to benefit learners through its process and product, more ample opportunities to utilise the product of WL should have been given to the participants in order to fully investigate its effect.

Second, the original text of the dictogloss included only three target sentences, which was not likely to be enough exposure for the participants to develop their interlanguage knowledge of, and/or control over, the target construction. In addition, it influenced the investigation between the focus of WLEs, which was operationalised as the frequency of T-WLEs, and learning. That is, the maximum number of T-WLEs turned out to be three because of the limited number of target sentences. A higher number of target sentences should have been included in order to have more range, which could, potentially, have detected more subtle differences.

Third, although based on the experience of a pilot to the main study, an extra practice session of WL was included, this still seems to have been insufficient. As reported, some +WL participants commented on the exit questionnaire that they found WL "refreshing" as "they had never done it before," indicating a need for more preparation to familiarise learners with the concept and activity of WL. Given that the quality of self-explaining training influenced the learning outcome (McNamara et al., 2006), improving the quality of practice should be considered as well.

Fourth, also related to practice, in order to avoid teacher-class effects, each participating class (except for the control group) was divided into two groups (+WL group and –WL group). All the participants, therefore, including those who were later

assigned to the –WL group, experienced the WL training. Although it was ensured that the –WL participants did not engage in WL, the experience might have influenced the outcome.

Fifth, although an exit questionnaire was employed to gain information regarding the participants' perceptions and perspectives on the experiment and WL (only for the +WL participants), and this information was used to interpret the results, the validity of the data could be questioned on the ground that I was their instructor and the participants, at least some of them, might have felt it necessary to report what would interest me instead of their honest feelings. In a similar vein, given that the two interviewees were volunteers, the same claim might be made regarding their responses.

Future research should circumvent the shortcomings stated above. First, the impact of WL should be examined with tasks that offer learners enough opportunities for WL and potentially enable them to benefit from its process and product. Moreover, given that no longitudinal WL studies have been conducted (except for Simard et al.'s (2007) three-month investigation of written reflections with learning journals), the effect of WL should be investigated in the longer term, which should shed new light on the field.

Second, the current study employed a single treatment task (dictogloss) with a single target construction (the present counterfactual conditional). Given, however, that task types and the difficulty of target constructions are influential factors for outcomes (Li, 2013), future research should be conducted with various task types and target structures in order to find out ideal task types and target structures that are likely to maximise the facilitative impact of WL.

Third, listed as one of the limitations, the current study did not provide enough opportunities for WL practice. Given the positive impact of frequency and quality of T-

WLEs on learning as well as the significant correlations between the frequency and quality of T-WLEs, to elicit higher numbers of, and more importantly, higher quality of T-WLEs seems important. Considering that the positive impact of training programmes, such as iSTART (McNamara et al., 2006) and SERT (McNamara, 2004, 2017), has been reported, it seems worthwhile to develop training programmes to elicit WL that are likely to contribute to learning in addition to offering extra practice sessions.

Fourth, related to practise again, as the participants in the two treatment groups practised dictogloss four times (two times individually and two times combined with WL) beforehand, task familiarity might have boosted their success on the task (Bygate, 2001, 2018). That is to say, repeating the task might have enabled them to focus less on procedural issues and more on content and linguistic features when they received the treatment. Given this, in future research, it would be interesting to explore how repeating the same task type might influence learners' WL.

Fifth, another important avenue for future research could be the investigation of the impact of WL on L2 learning in relation to learner factors/individual differences other than aptitude, such as motivation and attitude/affect, as learners are agents who have control over their learning (Storch & Wigglesworth, 2010; Swain, 2006).

Although WL research has been growing, it still seems to be underexplored compared to OL. The goal of my future research is to further explore the role of WL in L2 learning, with the limitations mentioned above in mind, in order to contribute to effective and efficient teaching and learning.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near native second language acquisition. *Studies in Second Language Acquisition, 30*, 489–509.

Adams, R. (2003). L2 output, reformulation, and noticing: implications for IL development. *Language Teaching Research, 7*, 347–76.

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory, 25,* 764–771.

Aizawa, K, Ishikawa, S., & Murata, T. (2015). *JACET 8000 eitango* [JACET 8000 English vocabulary]. Tokyo: Kirihara Press.

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language Teaching Research*, *1*, 93–121.

Alegría de la Colina, A., & García Mayo, M. P. (2009). Oral interaction in task-based EFL learning: The use of the L1 as a cognitive tool. *International Review of Applied Linguistics, 47*, 325–345.

Ammar, A. & Hassan, R. M. (2017). Talking it through: Collaborative dialogue and second language learning. *Language Learning, 68*, 46–82.

Antonek, J. L., Mccormick, D. E., & Donato, R. (1997). The student teacher portfolio as autobiography: Developing a professional identity. *The Modern Language Journal*, *81*, 15–27.

Arita, S. (2014). Conditionals and modals in Japanese: 'Settledness' as an interface between tense and modality. *Rocznik Orientalistyczny, 67,* 26–41. Retrieved from http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.pan-ro-yid-2014-iid-1-art-000000000004/c/ROrient201-142004Arita.pdf

Arievitch, I. M., & Haenen, J. P. P. (2005). Connecting sociocultural theory and educational practice: Gal'perin's approach. *Educational Psychologist, 40*, 155–165.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189–208.

Bardovi-Harlig, K. (1994). Reverse-order reports and the acquisition of tense: Beyond the principle of chronological order. *Language Learning, 44*, 243–282.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use.* Malden, MA: Blackwell.

Becker, A. L. (1991a). A short essay on languaging. In F. Steier (Ed.), *Reflexivity: Knowing as Systematic Social Construction* (pp. 226–234). Newbury Park, CA: Sage.

Becker, A. L. (1991b). Language and languaging. *Language and Communication, 11*, 33–35.

Berry, R. (2005). Making the most of metalanguage. *Language Awareness, 14*, 3–20.

Berry, R. (2009). EFL majors' knowledge of metalinguistic terminology: A comparative study. *Language Awareness, 18*,113–128.

Berry, R. (2014). Investigating language awareness: The role of terminology. In A. Łyda & K. Szcześniak (Eds.), *Awareness in action: The role of consciousness in language acquisition* (pp. 21–33). Geneva, Switzerland: Springer.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*, 102–118.

Bitchener, J. (2012). A reflection on 'the language learning potential' of written CF. *Journal of Second Language Writing, 21*, 348–363.

Bitchener, J. & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics, 31,* 193–214.

Bitchener, J. & Storch, N. (2016). *Written Corrective Feedback for L2 Development.* Bristol, UK: Multilingual Matters.

Bowles, M. A. (2008). Task type and reactivity of verbal reports in SLA: A first look at a L2 task other than reading. *Studies in Second Language Acquisition, 30,* 359–387.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. London: Routledge.

Bowles, M. A., & Leow, R. P. (2005). Reactivity and type of verbal reports in SLA research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition, 27,* 415–440.

Brooks, L., & Swain, M. (2009). Languaging in collaborative writing: Creation of and response to expertise. In A. Mackey, & C. Polio (Eds.), *Multiple perspectives on interaction in SLA* (pp. 55–89). Mahwah, NJ: Lawrence Erlbaum.

Brooks, L., Swain, M., Lapkin, S., & Knouzi, I. (2010). Mediating between scientific and spontaneous concepts through languaging. *Language Awareness, 19,* 89-110.

British National Corpus retrieved from https://quizlet.com/8935711/british-national-corpus-top-1000-flash-cards/

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Harlow: Pearson Longman.

Bygate, M. (2018). *Language Learning through Task Repetition.* Amsterdam: John Benjamins.

Carroll, J. B. (1981). Twenty-five years of research in foreign language aptitude. In K. C. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83–118). Rowley, MA: Newbury House.

Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude: Then and now. In

T. Parry & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 11–29). Englewood Cliffs, NJ: Prentice Hall.

Carroll, J. B., & Sapon, S. M. (1959). *Modern Language Aptitude Test.* The Psychological Corporation. San Antonio, Texas.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course.* Boston, MA: Heinle & Heinle Publishers.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 students writing. *Journal of Second Language Writing, 12*, 267–296.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Chiu, J., & Chi, M. T. H. (2014). Supporting self-explanation in the classroom. In V. Benassi, C. Overson, & C. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the classroom* (pp. 91–103). Retrieved from http://teachpsych.org/ebooks/asle2014/index.php

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Conati, C., & VanLehn, K. (2000). Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education, 11*, 398–415.

Cook, V. (1991). *Second language learning and language teaching.* London, UK: Edward Arnold.

Craik, F., & Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behaviour, 11*, 671–684.

Cronbach, L. J. (1967). How can instruction be adapted to individual differences? In R. M. Gagne (Ed.), *Learning and individual differences* (pp. 23–39). Columbus, OH: Merrill Books.

Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning*, *39*, 81–135.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7*, 482–511.

Dancygier, B., & Sweetser, E. (2005). *Mental spaces in grammar: Conditional constructions*. Cambridge, UK: Cambridge University Press.

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

DeKeyser, R. (2003). Implicit and explicit learning. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 313–348). Oxford: Blackwell.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). New Jersey: Lawrence Erlbaum Associates, Inc.

DeKeyser, R. (2009). Cognitive-psychological processes in second language learning. In M. Long & C. Doughty (Eds.), *Handbook of second language teaching* (pp. 119–138). Oxford, UK: Willey-Blackwell.

DeKeyser, R. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*(Suppl. 2), 189–200.

DiCamilla, F. J., & Lantolf, J. P. (1994). The linguistic analysis of private writing. *Language Sciences*, *16*, 347–369.

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford: Oxford University Press.

Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro (Ed.), *Continuum companion to second language acquisition* (pp. 247–267). London: Continuum.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 612–630). Oxford: Blackwell

Dörnyei. Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. New York: Routledge.

Doughty, C. J. (2013). Optimizing post-critical-period language learning. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 153–175). Amsterdam: John Benjamins.

Doughty, C. J., Campbell, S. G., Bunting, M. F., Bowles, A. R., & Haarmann, H. J. (2007). *The development of the High-Level Language Aptitude Battery.* Technical Report. College Park, MD: University of Maryland Center for Advanced Study of Language.

Duncan, R. M., & Cheyne, J. A. (2001). Private speech in young adults: Task difficulty, self-regulation, and psychological predication. *Cognitive Development, 16*, 889–906.

Duncan, R., & Tarulli, D. (2009). On the persistence of private speech: Empirical and theoretical considerations. In A. Winsler, C. Fernyhough, & I. Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-*

regulation (pp. 176–187). New York: Cambridge University Press.

Egi, T. (2007). Recasts, learners' interpretations, and L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 249–267). Oxford: Oxford University Press.

Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79*, 67–89.

Elder, C. (2009) Validating a test of metalinguistic knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp.113–138). Bristol: Multilingual Matters.

Elder, C., & Manwaring, D. (2004). The relationship between metalinguistic knowledge and learning outcomes among undergraduate students of Chinese. *Language Awareness, 13*, 145–162.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305–352.

Ellis, R. (2001). Introduction: Investigating form-focussed instruction. *Language Learning*, *51* (Suppl. 1), 1–46.

Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning, 54,* 227–275.

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

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics, 27*, 431–463.

Ellis, R. (2008). Explicit form-focused instruction and second language acquisition. In B. Spolsky & F. M. Hult (Eds.), *The handbook of educational linguistics* (pp.

437–455). New York: Routledge.

Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal, 63*, 97–107.

Emig, E. (1977). Writing as a mode of learning, *College Composition and Communication, 28*, 22–128.

Ericsson, K. A. (2002). Toward a procedure for eliciting verbal expression of nonverbal experience without reactivity: Interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology, 16*, 981–987.

Ericsson, K.A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: The MIT Press.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9*, 147–171.

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing*, *8*, 1–10.

Ferris, D. R. (2004). The 'grammar correction' debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime…?) *Journal of Second Language Writing, 13*, 49–62.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th edition). Sage Publishing.

Fortune, A. (2005). Learners' use of metalanguage in collaborative form-focused L2 output tasks. *Language Awareness, 41*, 21–38.

Gal'perin, P. I. (1992). Stage-by-stage formation as a method of psychological investigation. *Journal of Russian & East European Psychology, 30*, 60–80.

García Mayo, M.P., & Loidi Labandibar, U. (2017).The use of models as written corrective feedback in English as a foreign language (EFL)witing. *Annual*

Review of Applied Linguistics, 37, 110–127.

Granena, G. (2013) Cognitive aptitudes for second language learning and the LLAMA Language Aptitude Test. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 105–129). Amsterdam: John Benjamins.

Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal, 84*, 390–405.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy, *Memory & Cognition, 36*, 93–103.

Gutiérrez, X. (2016). Analyzed knowledge, metalanguage, and second language proficiency. *System, 60*, 42–54.

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research, 11*, 459–479.

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: Addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing, 21*, 332–347.

Hausmann, R. G. M., & Chi, M. T. H. (2002). Can a computer interface support self-explaining? *Cognitive Technology, 7*, 4–14.

Hertel, P. T. (1993). Implications of external memory for investigations of mind. *Applied Cognitive Psychology, 7*, 665–674.

Hu, G. (2011). Metalinguistic knowledge, metalanguage, and their relationship in L2 learners. *System, 39*, 63–77.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and

explicit second-language learning. *Studies in Second Language Acquisition, 27,* 129–140.

Ionas, G. I., Cernusca, D., & Collier, H. L. (2012). Prior knowledge influence on self-explanation effectiveness when solving problems: An exploratory study in science learning. *International Journal of Teaching and Learning in Higher Education, 24,* 349–358.

Ishikawa, M. (2013). Examining the effect of written languaging: The role of metanotes as a mediator of second language learning. *Language Awareness, 22,* 220–233.

Ishikawa, M. (2015). Metanotes (written languaging) in a translation task: Do L2 proficiency and task outcome matter? *Innovation in Language Learning and Teaching, 9,* 115–129.

Ishikawa, M. (2018). Written languaging, learners' proficiency levels and L2 grammar learning. *System, 74,* 50–61.

Ishikawa, M., & Suzuki, W. (2016). The effect of written languaging on learning the hypothetical conditional in English. *System, 58,* 97–111.

Izumi, S. (2003). Comprehension and production processes in second language learning: In search of the psycholinguistic rationale of the output hypothesis. *Applied Linguistics, 24,* 168–196.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing in second language acquisition? *TESOL Quarterly, 34,* 239–278.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition, 21,* 421–452.

John-Steiner, V. (1991). Private speech among adults. In L. Berk & R. Diaz (Eds.), *Private speech: From social interaction to self regulation*. (pp. 545–553). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jourdenais, R. (2001). Cognition, instruction and protocol analysis. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 354–375). New York: Cambridge University Press.

Kastens, K. A., & Liben, L. S. (2007). Eliciting self-explanations improves children's performance on a field-based map skills task. *Cognition and Instruction, 25,* 45–74.

Kim, Y., & McDonough, K. (2008). The effect of interlocutor proficiency on the collaborative dialogue between Korean as a second language learners. *Language Teaching Research, 12,* 211–234.

Knouzi, I., Swain, M., Lapkin, S., & Brooks, L. (2010). Self-scaffolding mediated by languaging: Microgenetic analysis of high and low performers. *International Journal of Applied Linguistics, 20*, 23–49.

Kormos, J. (2013). New conceptualizations of language aptitude in second language attainment. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 131–152). Amsterdam: John Benjamins.

Kowal, M., & Swain, M. (1994). Using collaborative language production tasks to promote students' language awareness. *Language Awareness, 3,* 73–93.

Kowal, M., & Swain, M. (1997). From semantic to syntactic processing: How can we promote metalinguistic awareness in the French immersion classroom? In R. K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives* (pp. 284–309). Cambridge: Cambridge University Press.

Krashen, S. D. (1981). Aptitude and attitude in relation to second language acquisition and learning. In K. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 155–175). Rowley, MA: Newbury House.

Lado, R. (1979). Thinking and "languaging": A psycholinguistic model of performance and learning. *Sophia Linguistics, 12,* 3–24.

Lantolf, J. P. (2006). Sociocultural theory and second language learning: State of the art. *Studies in Second Language Acquisition*, *28*, 67–109.

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development.* Oxford: Oxford University Press.

Lantolf, J. P., & Poehner, M. E. (2014). *Sociocultural theory and the pedagogical imperative in L2 education: Vygotskian praxis and the research/practice divide.* New York: Routledge.

Lapkin, S., Swain, M., & Knouzi, I. (2008). French as a second language university students learn the grammatical concept of voice: A study design, materials development and pilot data. In J. P. Lantolf & M. E. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp. 228–255). London: Equinox Press.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York: Routledge.

Lee, J. (2008). Gesture and private speech in second language acquisition. *Studies in Second Language Acquisition*, 30, 169–190.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*, 467–506.

Leow, R. P. (2001). Do learners notice enhanced forms while interacting with the L2? An online and offline study of the role of written input enhancement in L2 reading. *Hispania*, *84*, 496–509.

Leow, R. P. (2015). *Explicit learning in the classroom: A student-centered approach.* New York: Routledge.

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research, 30,* 111–127.

Leow, R. P., Johnson, E., & Zárate-Sández, G. (2011). Getting a grip on the slippery

construct of awareness: Toward a finer grained methodological perspective. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 61–72). Washington, DC: Georgetown University Press.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26,* 35–57.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal, 97,* 634–654.

Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics, 36,* 385–408.

Lidstone, J. S. M., Meins, E., & Fernyhough, C. (2010). The roles of private speech and inner speech in planning in middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology, 107,* 438–451.

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process,* (pp. 22–47). Clevedon, UK: Multilingual Matters.

Linck, J., Hughes, M., Campbell, S., Silbert, N., Tare, M., Jackson, S., Smith, B., Bunting, M., & Doughty, C. (2013). Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning, 63,* 530–566.

Luria, A. R. (1982). Basic forms of the speech utterance: oral (monologic and dialogic) and written speech. In J. Wertsch (Ed.), *Language and cognition* (pp. 159–168). Washington, D. C. Wiley.

Luria, A. R. (1999). Speech development and the formation of mental processes. In P.

Lloyd & C. Fernyhough (Eds.), *Lev Vygotsky: Critical Assessments, Volume II. Thought and Language* (pp. 84–122). London, England: Taylor & Francis.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics, 27,* 405–430.

Mackey, A., McDonough, K., Fujii, A., & Tatsumi, T. (2001). Investigating learners' reports about the L2 classroom. *International Review of Applied Linguistics, 39,* 285–308.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback, and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–209). Philadelphia: John Benjamins.

Manchón, R. M. (2011). Writing to learn the language: Issues in theory and research. In R. M. Manchón (Ed.), *Learning-to-write and writing-to-learn in an additional language* (pp. 61–82). Amsterdam: John Benjamins.

McCafferty, S. G. (1992). The use of private speech by adult second language learners: A cross-cultural study. *The Modern Language Journal, 76,* 179–189.

McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on ESL question development. *Studies in Second Language Acquisition, 27,* 79–103.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30.

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes, 54,* 479–492.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147–171.

Meara, P. (2005). *LLAMA language aptitude tests: The manual*. Retrieved from http://www.lognostics.co.uk/tools/llama/llama_manual.pdf.

Mizutani, N. (1989). *Nihongo kyouiku no naiyou to houhou: Koubun no nichiei hikaku wo cyushin ni* [Content and method of Japanese language teaching: Focusing on the comparison between Japanese and English structures]. Tokyo: ALC Press.

Moradian, M. R., Miri, M., & Nasab, M. H. (2017). Contribution of written languaging to enhancing the efficiency of written corrective feedback. *International Journal of Applied Linguistics, 27*, 406–421.

Muñoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools. *Behavior Research Methods*, *38*, 211–217.

Nabei, T., & Swain, M. (2002). Learner awareness of recasts in classroom interaction: A case study of an adult EFL student's second language learning. *Language Awareness*, *11*, 43–63.

Nadeau, M., & Fisher, C. (2014). Expérimentation de pratiques innovantes, la dictée 0 faute et la phrase dictée du jour, et étude de leur impact sur la compétence orthographique des élèves en production de texte [Experimenting with innovative practices, zero−error dictation and the daily sentence dictation, and study of their impact on students' spelling ability in text production]. Research report. Quebec City, Canada: Fonds de recherché du Québec–Société et culture.

Negueruela, E. (2008). Revolutionary pedagogies: Learning that leads to second language development. In J. P. Lantolf & M. Poehner (Eds.), *Sociocultural theory and the teaching of second languages* (pp.189–227). London: Equinox Press.

Negueruela, E., & Lantolf, J.P. (2006). Concept-based instruction and the acquisition of L2 Spanish. In R. A. Salaberry & B. A. Lafford (Eds.), *The art of teaching Spanish: Second language acquisition from research to praxis* (pp. 79–102). Washington, DC: Georgetown University Press.

Norouzian, R., & Plonsky, L. (2017). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research, 34*, 257–271.

Norris, J., & Ortega, L. (2003). Defining and measuring L2 acquisition. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell.

Ohta, A. (2000). Rethinking recasts: A learner-centered examination of corrective feedback in the Japanese classroom. In J. K. Hall & L. Verplaeste (Eds.), *The construction of second and foreign language learning through classroom interaction* (pp. 47–71). Mahwah, NJ, Lawrence Erlbaum Associates.

Ohta, A. (2001). *Second language acquisition processes in the classroom setting: Learning Japanese*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ottó, I. (2002). Magyar Egységes Nyelvérzékmérő-Teszt [Hungarian language aptitude test]. Kaposvár: Mottó-Logic Bt.

Parry, T. S., & Stansfield, C. W. (1990). Introduction. In T. S. Parry & C. W. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 1–10). Englewood Cliffs, NJ: Prentice Hall.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*, 69–78.

Pinker, S. (1994). *The language instinct.* New York, NY: Harper Perennial Modern Classics.

Pimsleur, P. (1966). The Pimsleur Language Aptitude Battery. New York: Harcourt, Brace, Jovanovic.

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878–912.

Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing, 21*, 375–389.

Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*, 277–303.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855–863.

Rees, J. (2000). Predicting the future of foreign language aptitude. In S. Cornwell & P. Robinson (Eds.), *Individual differences in foreign language learning: Effects of aptitude, intelligence and motivation* (pp. 187–197). Tokyo: Aoyama Gakuin University.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1–29.

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning, 64*, 615–650.

Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review, 24*, 1501–1510.

Robinson, P. (1995). Aptitude, awareness and the fundamental similarity of implicit and explicit second language learning. In R.W. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 303–343). University of Hawaii: Second Language and Teaching Curriculum Center.

Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition, 18*, 27–67.

Robinson, P. (2002). Learning conditions, aptitude complexes, and SLA: A framework for research and pedagogy. In P. Robinson (Ed.), *Individual differences in instructed language learning* (pp. 113–133). Amsterdam, the Netherlands: Benjamins.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25*, 45–73.

Robinson, P. (2012). Individual differences, aptitude complexes, SLA processes and aptitude test development. In M. Pawlak (Ed.), *New perspectives on individual differences in language learning and teaching* (pp. 57–76). Oxford, UK: Springer.

Robinson, P. (2013). Introduction to the encyclopedia—The scope and methods of inquiry. In P. Robinson (Ed.), *The Routledge encyclopedia of second language acquisition* (pp. xxii–xxiv). New York/London: Routledge.

Roebuck, R. (2000). Subjects speak out: How learners position themselves in a psycholinguistic task. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 79–95). Oxford: Oxford University Press.

Roehr-Brackin, K. (2014). Explicit knowledge and processes from a usage-based perspective: The developmental trajectory of an instructed L2 learner. *Language Learning, 64*, 771–808.

Roehr-Brackin, K. (2015). Explicit knowledge about language in L2 learning: A usage-based perspective. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages (Studies in bilingualism)* (pp. 117–138). Amsterdam: Benjamins.

Rosa, E. M., & Leow, R. P. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics, 24*, 269–292.

Rossomondo, A. E. (2007). The role of lexical temporal indicators and text interaction format in the incidental acquisition of the Spanish future tense. *Studies in Second Language Acquisition, 29*, 39–66.

Sachs, R. R. (2010) *Individual differences and the effectiveness of visual feedback on reflexive binding in L2 Japanese* (Unpublished doctoral thesis). Georgetown University, Washington DC.

Sachs, R. R., & Polio, C. (2007). Learners' uses of two types of written feedback on a L2 writing revision task, *Studies in Second Language Acquisition, 29*, 67–100.

Sachs, R. R., & Suh, B.R. (2007). Textually enhanced recasts, learner awareness and L2 outcomes in synchronous computer-mediated interaction. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp.197–227). Oxford University Press.

Sanz, C., Lin, H-J., Lado, B., Bowden, H.W., & Stafford, C.A. (2009). Concurrent verbalizations, pedagogical conditions and reactivity: Two CALL studies. *Language Learning*, *59*, 33–71.

Sasaki, M. (1991). *Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modelling approach* (Unpublished doctoral thesis). University of California, Los Angeles, Los Angeles.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses.* New York: Peter Lang.

Saville-Troike, M. (1988). Private speech: Evidence for second language learning strategies during the 'silent' period. *Journal of Child Language*, *15*, 567–590.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 129–158.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–226.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. *AILA Review*, *11*, 11–26.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–17). Honolulu, HI: University of Hawai'i Press.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (Eds.), *Proceedings of CLaSIC 2010* (pp. 721–737). Singapore: National University of Singapore, Centre for Language Studies.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a foreign language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn* (pp. 237–326). Rowley, MA: Newbury House.

Sharwood Smith, M. (1976). A note on 'writing versus speech.' *ELT Journal, 31*, 17–19.

Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics, 2,* 159–168.

Sheen, Y. (2007a). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41*, 255–281.

Sheen, Y. (2007b). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301–322). Oxford: Oxford University Press.

Shintani, N., & Ellis, R. (2015). Does language analytical ability mediate the effect of written feedback on grammatical accuracy in second language writing? *System, 49*, 110–119.

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures: Effects of written feedback and revision. *Language Learning*, *64,* 103–131.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Garnott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31–58). New York: Cambridge University Press.

Simard, D., French, L., & Fortier, V. (2007). Elicited metalinguistic reflection and

second language learning: Is there a link? *System, 35*, 509–522.

Simard, D., Guénette, D., & Bergeron, A. (2015). L2 learners' understanding of written corrective feedback: Insights from their metalinguistic reflections. *Language Awareness, 24*, 233–254.

Skehan, P. (1986). The role of foreign language aptitude in a model of school learning. *Language Testing, 3*, 188–221.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2002). Theorizing and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–94). Amsterdam: John Benjamins.

Skehan, P. (2012). Language aptitude. In S. Gass, & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 381–395). New York: Routledge.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics, 36*, 367–384.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Slobin, D. I. (1987). Thinking for speaking. *Proceedings of the Thirteenth Annual Meetings of the Berkeley Linguistics Society, 13*, 435–444.

Stefanou, C. (2014). *L2 article use for generic and specific plural reference: The role of written corrective feedback, learner factors and awareness* (Unpublished doctoral thesis). Lancaster University, UK.

Stefanou, C., & Révész, A. (2015). Direct written corrective feedback, learner differences, and the acquisition of second language article use for generic and specific plural reference. *The Modern Language Journal, 99*, 263–282.

Storch, N. (1999). Are two heads better than one? Pair work and grammatical accuracy.

System, 27, 363–374.

Storch, N. (2001). Comparing ESL learners' attention to grammar on three different collaborative tasks. *RELC Journal 32*, 104–124.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52,* 119–158.

Storch, N. (2008). Metatalk in a pair work activity: Level of engagement and implications for language development. *Language Awareness, 17*, 95–114.

Storch, N. (2013). *Collaborative writing in L2 classrooms.* Bristol, UK: Multilingual Matters.

Storch, N., & Wigglesworth, G. (2003). Is there a role for the use of the L1 in an L2 setting? *TESOL Quarterly, 37*, 760–770.

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake and retention of corrective feedback on writing. *Studies in Second Language Acquisition, 32*, 303–334.

Suzuki, M. (2008). Japanese learners' self revisions and peer revisions of their written compositions in English. *TESOL Quarterly, 42*, 209–233.

Suzuki, W. (2009a). Improving Japanese university students' second language writing accuracy: Effects of languaging. *Annual Review of English Language Education in Japan, 20*, 81–90.

Suzuki, W. (2009b). *Languaging, direct correction, and second language writing: Japanese university students of English* (Unpublished doctoral thesis). University of Toronto, Canada.

Suzuki, W. (2012). Written languaging, direct correction, and second language writing revision. *Language Learning, 62,* 1110–1133.

Suzuki, W. (2016). The effects of quality of written languaging on second language learning. *Writing & Pedagogy, 8,* 461–482.

Suzuki, W., & Itagaki, N. (2007). Learner metalinguistic reflections following output-

oriented and reflective activities. *Language Awareness, 16,* 131–146.

Suzuki, W., & Itagaki, N. (2009). Languaging in grammar exercises by Japanese EFL learners of differing proficiency. *System, 37,* 217–225.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Cambridge, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). Cambridge: Cambridge University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–484). Mahwah, NJ: Lawrence Erlbaum Associates.

Swain, M. (2006). Languaging, agency and collaboration in advanced second language proficiency. In H. Byrnes (Ed.), *Advanced language learning: The contribution of Halliday and Vygotsky* (pp. 95–108). London: Continuum.

Swain, M. (2010). Talking it through: Languaging as a source of learning. In R. Batstone (Ed.), *Sociocognitive perspectives on language use and language learning* (pp. 112–130). Oxford: Oxford University Press.

Swain, M. & Deters, P. (2007). "New" mainstream SLA theory: Expanded and enriched. *The Modern Language Journal, 91,* 820–836.

Swain, M., Kinnear, P., & Steinman, L. (2011). *Sociocultural theory in second language education: An introduction through narratives*. Clevedon, UK: Multilingual Matters.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*, 371–391.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal, 82,* 320–337.

Swain, M., & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research, 37*, 285–304.

Swain, M., & Lapkin, S. (2007). 'Oh, I get it now!' From production to comprehension in second language learning. In D. M. Brinton & O. Kagan (Eds.), *Heritage language acquisition: A new field emerging* (pp. 301–319). Mahwah, NJ: Lawrence Erlbaum Associates.

Swain, M., Lapkin, S., Knouzi, I., Suzuki, W., & Brooks, L. (2009). Languaging: University students learn the grammatical concepts in French. *The Modern Language Journal, 93*, 5–29.

Swain, M., & Watanabe, Y. (2013). Languaging: Collaborative dialogue as a source of second language learning. In M. Swain (Ed.), *The encyclopedia of applied linguistics*. Hoboken, NJ: Blackwell.

Tajika, H., Nakatsu, N., Nozaki, H., Neumann, E., & Maruno, S. (2007). Effects of self-explanation as a metacognitive strategy for solving mathematical word problems. *Japanese Psychological Resea*rch, *49*, 222–233.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and SLA. *Studies in Second Language Acquisition, 16*, 183–203.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytical ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 171–195). Oxford University Press.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, *46*, 327–69.

Uggen, M. (2012). Reinvestigating the noticing function of output. *Language Learning, 62*, 506–540.

VanPatten, B. (1990). Attending to form and content in the input. *Studies in Second Language Acquisition, 12,* 287–301.

VanPatten, B., & Smith, M. (2015). Aptitude as grammatical sensitivity and the early stages of learning Japanese as an L2: Parametric variation and case-marking. *Studies in Second Language Acquisition, 37,* 135–165.

Vatz, K., Tare, M., Jackson, S. R., & Doughty, C. J. (2013) Aptitude-treatment interaction studies in second language acquisition: findings and methodology. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 272–292). Amsterdam and Philadelphia: John Benjamins.

Vocate, D. R. (1994). Self-talk and inner speech: Understanding the uniquely human aspects of intrapersonal communication. In D. R. Vocate (Ed.), *Intrapersonal communication: Different voices, different minds* (pp. 3–32). Hillsdale, NJ: Erlbaum.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Vygotsky, L. S. (1987) Thinking and speech. In R. W. Rieber, A. S. Carton, & N. Minick (Eds.), *The collected works of L. S. Vygotsky. Volume 1: Problems of general psychology* (pp. 37–285). New York: Plenum Press.

Wajnryb, R. (1990). *Grammar dictation*. Oxford: Oxford University Press.

Wesche, M. B. (1981). Language aptitude measures in streaming, matching students with methods, and diagnosis of learning problems. In K. Diller (Ed.), *Individual differences and universals in language learning aptitude* (pp. 119–154). Rowley, MA: Newbury House.

Wigglesworth, G., & Storch, N. (2012). What role for collaboration in writing and writing feedback. *Journal of Second Language Writing, 21,* 364–374.

Winsler, A. (2009). Still talking to ourselves after all these years: A review of current research on private speech. In A. Winsler, C. Fernyhough, & I. Montero (Eds.), *Private speech, executive functioning, and the development of verbal self-regulation* (pp. 3–41). New York: Cambridge University Press.

Wilkinson, K. (2009). *Les effets de la dictée 0 faute sur la compétence en orthographe d ·élèves de troisième secondaire* [Effects of zero-error dictation on secondary three students' spelling ability]. Unpublished master's dissertation, Université du Québec à Montréal, Montréal, Québec, Canada.

Williams, J. (2001). Learner-generated attention to form. *Language Learning, 51,* 303–346.

Williams, J. (2008). The speaking-writing connection in second language and academic literacy development. In D. Belcher & A. Hirvela (Eds.), *The oral/literate connection: Perspectives on L2 speaking, writing, and other media interactions* (pp. 10–25). Ann Arbor: University of Michigan Press.

Williams, J. (2012). The role(s) of writing and writing instruction in L2 development. *Journal of Second Language Writing, 21*, 321–331.

Williams, J. N. (2004). Implicit learning of form-meaning connections. In B. VanPatten, J. Williams, S. Rott, & M. Overstreet (Eds.), *Form-meaning connections in second language acquisition* (pp. 203–218). Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language*

Acquisition 27, 269–304.

Yalçın, Ş. (2012). *Individual differences and the learning of two grammatical features with Turkish learners of English* (Unpublished doctoral thesis). University of Toronto, Canada.

Yalçın, Ş., & Spada, N. (2016). Language aptitude and grammatical difficulty: An EFL classroom-based study. *Studies in Second Language Acquisition, 38*, 239–263.

Yang, L. (2010). Doing a group presentation: Negotiations and challenges experienced by five Chinese ESL students of commerce at a Canadian university. *Language Teaching Research, 14*, 141–160.

Yang, L. (2016). Languaging in story rewriting tasks by Chinese EFL students. *Language Awareness, 25*, 241–255.

Yanguas, I., & Lado, B. (2012). Is thinking aloud reactive when writing in the heritage language? *Foreign Language Annals, 45*, 380–399.

Yilmaz, M. (2016). Improving Turkish EFL learners' writing accuracy: Effects of written languaging and languaging type. *Procedia - Social and Behaviour Sciences, 232*, 413–420.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics, 34*, 344–368.

Yoshida, R. (2008). Functions of repetition in learners' private speech in Japanese language classrooms, *Language Awareness, 17*, 289–306.

Appendix A: Informed Consent Documentation for the Pilot Study (Japanese)

Institute of Education

UCL

実験情報シート

筆記ランゲージング・学習者の習熟度レベル・第二言語学習

私は皆さんの英語の授業を担当する石川正子です。今回は皆さんに私の研究プロジェクトへの参加をお願いしたいと思います。私は英語を教えることに加えて、第二言語習得の分野で「ランゲージング」（学習者が問題に直面した時に使用する副次的な言語）の研究を行っていて、特にその筆記版、筆記ランゲージングに注目しています。現在はロンドン大学に在籍し、今回皆さんに参加をお願する実験はその博士課程での研究の一部です。

この実験は3週間に渡って行われます。第1週目には、文法問題を行います。第2週目は1つのタスクの後、別の文法問題のセットに取り組んでいただきます。第3週目には、また別の文法問題のセットに解答してください。全てのテストを提出するようお願いしますが、テスト結果が成績に影響することはありません。実験は全て通常の90分授業の一部（1・3週目）又は全て（2週目）の時間で、いつもの教室で行われます。

この実験は純粋に研究目的であり、実験に参加するか否の決断が成績に影響することはありません。更に、実験途中にいつでも罰則なしに参加をやめることもできます。この実験で得られる結果は研究目的のみに使用され、皆さんの個人的な情報が公開されることはありません。実験後、私は全テストをチェックして結果を2016年1月最後の授業でお知らせします。

このシートを読んでいただき、ありがとうございます。実験に参加するかどうかは皆さん次第ですが、実験に参加し、良い経験をしたと感じることを願っています。実験に参加していただける場合は同意書に記入して、2015年11月27日までに提出してください。もし質問があれば、ご連絡ください（masako.ishikawa.14@ucl.ac.uk）。このプロジェクトはロンドン大学倫理委員会の承認を得ています。

ロンドン大学博士課程　石川正子 [redacted]
ロンドン大学　カルチャー・コミュニケーション・メディア学部
ベッドフォードウェイ20、ロンドン　WC1H OAL

Institute of Education

UCL

実験参加同意書

筆記ランゲージング・学習者の習熟度レベル・第二言語学習

2015/12/01 ~ 2016/01/22

実験に参加していただける場合はこの同意書に記入して2015年11月27日までに石川正子まで提出してください。

	はい	いいえ
私はこの研究に関する情報シートを読んで理解しました。	☐	☐
私は情報シートに説明されている実験に参加することに同意します。	☐	☐
私の解答が報告や発表で使用される際には個人名が明らかにならないことを理解しています。	☐	☐
私は途中で参加をやめられること、その際は私のデータが使われないことを理解しています。	☐	☐
私は実験結果が成績に影響しないことを理解しています。	☐	☐
私はいつでも石川正子[redacted]に連絡できることを理解しています。	☐	☐

氏名: _____

署名: _____

日付け: _____

Appendix A: Informed Consent Documentation for the Pilot Study (English)

Institute of Education

UCL

Information Sheet

Written Languaging, Learners' Proficiency Level, and L2 Learning

I am your English instructor, Masako Ishikawa, and I am inviting you to take part in my research project. In addition to teaching English, I conduct research in the field of second language acquisition, focusing on languaging, i.e., secondary language use which people make when they are faced with linguistic problems. I am especially interested in the written version of languaging, written languaging. Currently I'm enrolled at the the UCL Institute of Education, University College London, and the experiment I am inviting you to take part in is part of my doctoral studies there.

The experiment will be conducted in three sessions over three weeks. In the first session, you will be asked to work on a set of grammar tests. In the second session, you will perform a communicative task, followed by another set of grammar tests. In the third session, you will be asked to work on another set of grammar tests. Although I will ask you to submit all the tests, this experiment is for purely research purposes and none of the tests results will be reflected in your grade. All the sessions will be conducted in part (1st and 3rd sessions) or all (2nd session) of our regular 90-minute class time in our regular classroom.

This experiment has nothing to do with your English course, and your decision as to whether or not to participate will not have any effect on your grades. Moreover, you can stop anytime without penalty. The results which will be obtained in this study will be used only for research purposes and your personal information will never be disclosed. After the experiment, I will check all the tests and share the results of the tests in the last class of your course (the last week of January, 2016).

Thank you very much for taking the time to read this information sheet. It is entirely up to you whether or not you choose to take part. That said, I hope that you choose to participate and that you will find it a valuable experience if you do so. If you would like to be involved, please complete the consent form and return to me by 2015/11/27. If you have any further questions before you decide whether to take part, you can reach me at masako.ishikawa.14@ucl.ac.uk. This project has been reviewed and approved by the UCL IOE Research Ethics Committee.

Masako Ishikawa, PhD student [redacted]
Department of Culture, Communication, and Media
UCL Institute of Education, University of London
20 Bedford Way, London WC1H OAL

Institute of Education

UCL

Consent Form

Written Languaging, Learners' Proficiency Level, and L2 Learning

2015/12/01 ~ 2016/01/22

If you are happy to participate, please complete this consent form and return to Masako Ishikawa by 2015/11/27.

	Yes	No
I have read and understood the information sheet about the research.	☐	☐
I agree to take part in the experiment as outlined on the information sheet.	☐	☐
I understand that if any of my words are used in reports or presentations they will not be attributed to me.	☐	☐
I understand that I can withdraw from the project at any time, and that if I choose to do this, any data I have contributed will not be used.	☐	☐
I understand that the results will not have an effect on my grade.	☐	☐
I understand that I can contact Masako Ishikawa [redacted] at any time.	☐	☐

Name: _____

Signed: _____

Date: _____

Appendix B: Background Questionnaire for the Pilot and Main Studies

英語に関するアンケート

A questionnaire regarding your English

名前：_____年齢：_____

Name：_____ age：_____

1. 今までに英語をどのくらいの期間、勉強していますか？

How long have you been studying English?

_____年_____か月くらい勉強 している

For _____year(s) _____ month(s)

２．今までに英語圏に行ったこと、又は生活したことがあれば、場所・期間・時期を書いてください。（例：１０歳の時、ロンドンに２週間旅行）

If you have visited/lived in English speaking countries, write the place(s), duration and time (e.g., trip to London for 2 weeks when I was 10 years old)

３．英語以外に勉強している（していた）言語があれば書いてください。

If you have studied/studied foreign languages besides English, write below.

_____語を、_____年_____か月くらい勉強 している・した

I have studied/studied _____(name of the language)

for _____ year(s)_____ month(s).

(２か国語以上あれば下の空欄に上と同じように書いてください。)

(If there are more than one, use the space below.)

４．あなたの TOEIC スコアを書いてください。他にも英検や TOEFL など受験したことがある英語のテストがあればスコア・級を書いてください。

Please write your TOEIC score below. If you have taken any English tests (e.g., the TOEFL test, STEP) besides the TOEIC test, write the name(s) of the test(s) and your score(s)/grade(s).

Appendix C: Vocabulary Checklist

Vocabulary Checklist

I.

A. 人名：以下の人物を知っていますか？

People's names: Are these people familiar to you?

1. Pablo Picasso
2. Bill Gates & Paul Allen
3. LeBron James
4. Prince William
5. Akio Toyoda
6. Keith Haring
7. Taylor Swift

B. 地名：これらの場所を知っていますか？

Place names: Are you familiar with all these places?

✧ Argentina, Australia, China, Korea, India, Malaysia, Mexico
✧ London, Paris, New York, Alaska

C. 他

Others

- bedtime stories
- besides
- an entrance exam
- fall colors
- a graduate school
- location
- a model agency
- none
- a precious mirror
- a roof
- salmon fishing spots
- typhoon
- weather forecast

Name: _____

II.

A. 以下の動詞の過去形と過去分詞形を確認しましょう！

Let's check if you remember past and past perfect forms!

1. become – became - become
2. break - _____ - _____
3. come - _____ - _____
4. do - _____ - _____
5. drive - _____ - _____
6. eat - _____ - _____
7. fall - _____ - _____
8. give - _____ - _____
9. go - _____ - _____
10. grow - _____ - _____
11. know - _____ - _____
12. see - _____ - _____
13. shake - _____ - _____
14. take - _____ - _____
15. write - _____ - _____

B. 他

Others

- do nothing but~
- forgive
- found
- go hiking
- keep one's promise
- lose some/a lot of weight
- miss a class/presentation, a train/flight
- prepare for
- taste good
- wear heavy make-up

Appendix D-1: A Sample Version of Grammar Production Tests

English Exercises A

Name: _____

I. （　　）内の語を適当な形にして空所に入れ、各会話を完成させましょう。
（必要が無ければ形を変えずに入れること。必要ならば語を補うこと。）
Complete each conversation by putting an appropriate form of a word given in each parenthesis. (If not necessary, put the word as is. Add words if necessary.)

1. A: What a surprise! Taylor Swift is coming to our campus tonight!
   B: I know! Lisa is a big fan, but she is sick in bed today.
   If she _____ (know) about it, she would come in her pajamas.

2. A: Emi, are you an only child or do you have any brothers or sisters?
   B: I have twin brothers. They are 25.
   A: Are you _____ (young) in your family, then?

3. A: I have two tickets for tonight's movie. I need to find someone to go with me.
   B: If you _____ (want) me to go with you, I will go. What do you think?

4. A: I know that I need to exercise to lose some weight, but my knee hurts…
   B: You simply eat too much. If you _____ (eat) less, you would lose a lot of weight.

5. A: Welcome back to Japan! How was Korea?
   B: I had a lot of meetings and didn't have much free time.
   However, I _____ (enjoy) good Korean food every night.

6. A: *Doraemon* was very popular when I was a child. Is he still popular in Japan?
   B: Yes, he _____ (love) by many people. You can watch his TV anime every Friday.

7. A: Bob, it's time. You should stop writing.
   B: Oh, no! If you _____ (give) me a little more time, I could finish this test.

8. A: My parents often say to me, "Don't come home too late."
   B: My parents, too! They tell me _____ (not/come) home late all the time. I'm so sick of it…

9. A: Look! Kiyoshi is drinking again. We should get out of here.
   B: Yes, let's go! If he _____ (drink) too much, he will start to sing. He is a bad singer…

10. A: Where do you want me to move this mirror?
    B: Be careful! It's really precious. If you _____ (break) it, it would be very hard to find another one.

11. A: I really like this picture. Would you tell me who _____ (paint) it?
    B: I'm not sure, but I think Picasso. Let's ask the staff of this museum.

12. A: Is Alice going to the party tonight?
    B: I think she is still thinking. If she _____ (join) the party, she will wear her new dress.

13. A: I can never remember Amy's birthday… I'm sure it's coming soon, but I need to ask her again.
    B: Donald, that's not very nice. If you _____ (write) it down, you wouldn't have to do that every year.

14. A: You may not believe it, but our baseball team won 1st prize!
    B: What _____ (excite) news! I'm very happy for you.

15. A: Is it true that people speak English and French in Canada?
    B: Yes. English and French _____ (speak) in Canada. Many signs are in two languages.

16. A: Oh, no… The last train is leaving soon. If I _____(drive) you to the station, you could take it. Sorry, I'm too sleepy to drive.
    B: It's almost midnight. I understand.

17. A: I would like to go to Kyoto. When do you think is the best time?
    B: That's a difficult question. I enjoy every season. If you _____ (visit) next month, you will enjoy the fall colors.

18. A: Oh, no! We just missed the train. We should have left a little earlier…
    B: I hate to wait here in the rain. Do you know when the next train _____(arrive)?

19. A: I wonder why Amanda wears heavy make-up every day.
    B: I don't know, but she looks very different without it. If you _____ (see) her without make-up, you wouldn't know it's her.

20. A: Jane, please listen. Steve just called me and said "I need to talk with you."
    B: What? What did he say?
    A: He told me that he _____ (need) to talk with me.

21. A: Is your grandfather feeling better?
    B: No… He's been sick for a month. If he _____ (go) to the hospital, he would feel better. But he doesn't like going to the hospital.

22. A: Hi, Nick. What are you doing?
B: I am preparing for tomorrow's class. If the professor _____ (ask) me a question, I will be able to answer easily.

23. A: Hi, Lisa. Where are you? We were just talking about you.
B: Hi, Akira. I'm sorry but I will be late. I'm calling from the bus stop in front of my apartment. I _____ (wait) for a bus for 30 minutes now…

24. A: Is Kenji really traveling to Europe during the vacation?
B: I think so, but he may not have enough money. If he _____ (travel) to Malaysia, it will be much cheaper.

Appendix D-2: A Sample version of Recognition Tests

II.各文の状況を最もよく表している選択肢を一つ選び文字を丸で囲みましょう。(Read each sentence and choose one of the four choices that best describes the situation by circling the letter next to it.)

1. If my brother were at home, he could drive us to the airport.
a. My brother is at home.
b. My brother isn't at home.
c. My brother was at home.
d. My brother wasn't at home.

2. Hanako visited Argentina to learn tango last summer.
a. Hanako visited Argentina last summer.
b. Hanako has never learned tango.
c. Both (a) and (b) are OK.
d. Don't know.

3. I'll take whoever wants to come with me to Disneyland.
a. If you want to come with me, I will take you to Disneyland.
b. If you take me, I will come with you to Disneyland.
c. Both (a) and (b) are OK.
d. Don't know.

4. If Anna had time today, she could meet her friends for lunch.
a. Anna has time today.
b. Anna doesn't have time today.
c. Anna had time today.
d. Anna didn't have time today.

5. Ichiro cannot swim and Jiro cannot, either.
a. Either Ichiro or Jiro can swim.
b. Neither Ichiro nor Jiro can swim.
c. Both (a) and (b) are OK.
d. Don't know.

6. If you are hungry, you can have lunch now.
a. You may be hungry.
b. You aren't hungry.
c. You were hungry.
d. You were never hungry.

7. If Kate asked us, we would help her any time.
a. Kate asks us.
b. Kate doesn't ask us.
c. Kate asked us.
d. Kate didn't ask us.

8. No book is more interesting than this book.
a. This book is more interesting than any other book.
b. No book is as interesting as this book.
c. Both (a) and (b) are OK.
d. Don't know.

9. The class started at 10 am. We went to the classroom at 10:05 am.
a. The class started before we went to the classroom.
b. We went to the classroom after the class started.
c. Both (a) and (b) are OK.
d. Don't know.

10. If Bill needed money, he would get a job.
a. Bill needs money.
b. Bill doesn't need money.
c. Bill needed money.
d. Bill didn't need money.

11. If Akiko loves Mexican food, she should try this restaurant.
a. Akiko may love Mexican food.
b. Akiko never loves Mexican food.
c. Akiko loved Mexican food.
d. Akiko didn't love Mexican food.

12. Jack is 80. Mary is 85. Paul is 90.
a. Jack is the youngest of the three.
b. Mary is younger than Paul, but older than Jack.
c. Both (a) and (b) are OK.
d. Don't know.

13. If I lived close to the station, I wouldn't take a bus.
a. I live close to the station.
b. I don't live close to the station.
c. I lived close to the station.
d. I didn't live close to the station.

14. Cristiano does nothing but play football with his friends.
a. Cristiano's friends don't play football with him.
b. Cristiano plays football a lot.
c. Both (a) and (b) are OK.
d. Don't know.

15. If Takeshi comes for dinner, I will cook his favorite food.
a. Takeshi may come for dinner.
b. Takeshi never comes for dinner.
c. Takeshi came for dinner.
d. Takeshi didn't come for dinner.

16. If you bought a car, you could drive to school.
a. You buy a car.
b. You don't buy a car.
c. You bought a car.
d. You didn't buy a car.

17. Mike will have worked for the company for 10 years next month.
a. Mike has worked for the company for 10 years now.
b. Mike has worked for the company for a little less than 10 years now.
c. Both (a) and (b) are OK.
d. Don't know

18. The first novel in the Harry Potter (HP) series was published in 1997. The second one was published one year later.
a. The first novel in the HP series was published one year before the second one.
b. The second novel in the HP series was published in 1998.
c. Both (a) and (b) are OK.
d. Don't know

19. If Megumi practiced harder, she would be a good skater.
a. Megumi practices harder.
b. Megumi doesn't practice harder.
c. Megumi practiced harder.
d. Megumi didn't practice harder.

20. Not all the apples are sweet.
a. Some apples are sweet, but the others aren't.
b. None of the apples are sweet.
c. Both (a) and (b) are OK.
d. Don't know.

21. If Mark worked longer hours, he would make more money.
a. Mark works longer hours.
b. Mark doesn't work longer hours.
c. Mark worked longer hours.
d. Mark didn't work longer hours.

22. Jake promised us to finish the report by Friday, and he kept his promise.
a. Jake finished the report by Friday.
b. Jake didn't finish the report by Friday.
c. Both (a) and (b) are OK.
d. Don't know.

23. If Tom has a lot of work to do, he will stay home this weekend.
a. Tom may have a lot of work to do.
b. Tom may never have a lot of work to do.
c. Tom had a lot of work to do.
d. Tom never had a lot of work to do.

24. The guests are arriving soon. We should stay home and welcome them.
a. The guests are here.
b. The guests are not here yet.
c. Both (a) and (b) are OK.
d. Don't know

Appendix E: Dictogloss Reconstruction Sheet

Dictogloss

I.

太郎はもうすぐ卒業を控えた大学 4 年生ですが、まだ卒業後の進路が決まらずあれこれ考えています。これから彼の心の中のつぶやきを 2 回聞いていただきます。1 度目は内容に注意して、2 度目は下の余白にメモを取りながら聞いてください。（メモは英語でも日本語でも構いません。）聞き終わったら、そのメモを元にして出来るだけ正確につぶやきを再現してください。

Taro is a college student who is graduating soon. It's December but he still doesn't know what to do after his graduation. You are going to listen to his inner thoughts twice. First, just listen. Second, take notes in the space below. You will be asked to reconstruct what you hear based on the notes later.

もしも～	可能性・選択肢
if~	possibilities・options

✧

✧

✧

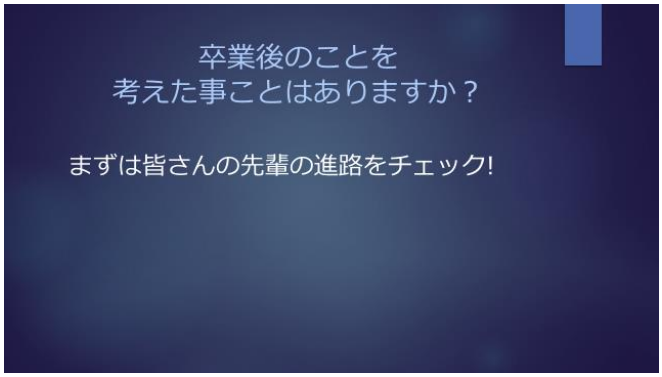
Name: _____

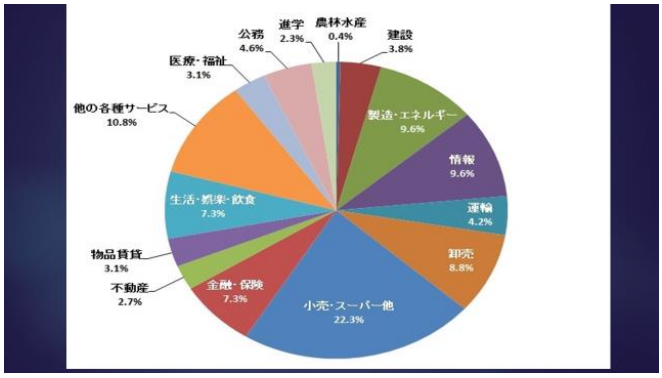
II. 太郎のつぶやき：メモを元に出来るだけ正確に再現しましょう。

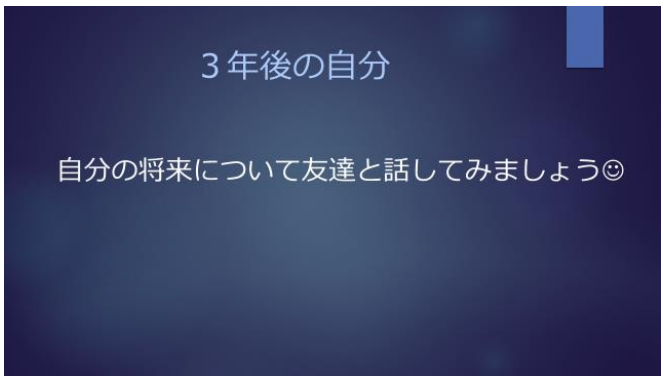
Taro's whisper: Reconstruct what you heard as accurately as possible with your notes.

It's December and I still don't know what to do after I finish college.

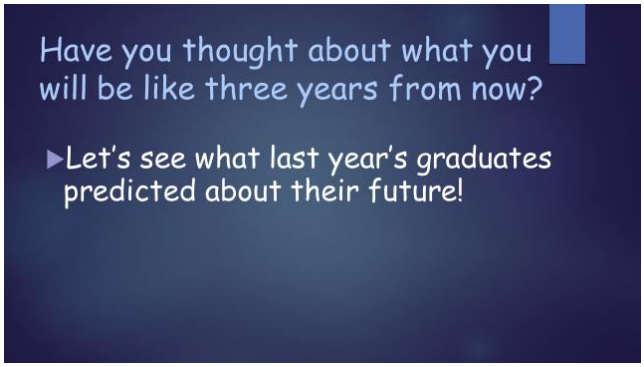
Appendix F-1: PowerPoint Slides for Warm-up (Japanese)

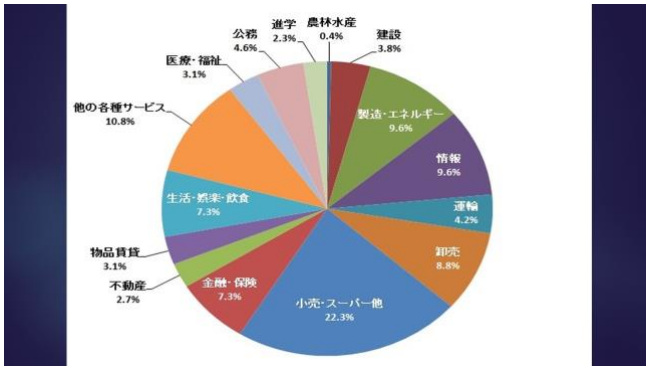






Appendix F-2: PowerPoint Slides for Warm-up (English)





Appendix G-1: WL Sheet for the +WL Group

Name:_____

III．原文と比較しながら考えたことを書く：

(Comparison with the original text with WL):

以下は先ほどの太郎の心のつぶやきです。時制や動詞の形などに特に注意しながらあなたが再現したものと注意深く見比べてください。違いはありますか？あるとすれば、どう違っていますか？（なぜ原文ではそのような形が使われているのでしょうか？文法のルールなどが思いつきますか？）気が付いたことや考えたことを何でも原文の行間、または下の余白に書いてください。

（全てこの用紙に書くこと！　先ほどの用紙に書き込んではいけません。）

Below is what you heard earlier. Please compare it with your reconstruction, paying special attention to forms such as tense and verb forms. Are there any differences? If so, how is it different? (Why do you think whatever you noticed is used in the original text? Can you think of the grammar rule?) Think hard and write whatever comes to your mind between the lines or the blank space below. (Write on this sheet! Please do not write on the sheet you used earlier.)

It's December and I still don't know what to do after I finish college.

I need to think about my options one more time. If I had better grades, I could go to graduate school. However, my grades are not that high. Besides, I don't like to study that much. If I were good at computers, I could find a good job. Again, no. I can go to a computer school, but maybe it's too late. How about going abroad? Yes, I don't have to work in Japan! Wait. My English is not very good… If I spoke English better, I would work abroad. None of these options work. I have to think about something else.

Appendix G-2: –WL Sheet for the –WL Group

Name:_____

III．原文との比較：Comparison with the original text

以下は先ほどの太郎の心のつぶやきです。時制や動詞の形などに特に注意しながらあなたが再現したものと注意深く見比べてください。違いはありますか？あるとすれば、どう違っていますか？（なぜ原文ではそのような形が使われているのでしょうか？文法のルールなどが思いつきますか？）何も書かずに原文をじっくりと読んでよく考えてください。この後エッセーを書いていただきます。

Below is what you heard earlier. Please compare it with your reconstruction, paying special attention to forms such as tense and verb forms. Are there any differences? If so, how is it different? (Why do you think whatever you noticed is used in the original text? Can you think of the grammar rule?) Think hard without writing. You are going to write an essay after this.

It's December and I still don't know what to do after I finish college.

I need to think about my options one more time. If I had better grades, I could go to graduate school. However, my grades are not that high. Besides, I don't like to study that much. If I were good at computers, I could find a good job. Again, no. I can go to a computer school, but maybe it's too late. How about going abroad? Yes, I don't have to work in Japan! Wait. My English is not very good… If I spoke English better, I would work abroad. None of these options work. I have to think about something else.

Appendix H: Informed Consent Documentation for the Main Study (Japanese)

Institute of Education

UCL

実験情報シート

「筆記ランゲージング・学習者の適性・第二言語学習」

私はみなさんの英語の授業を担当する石川正子です。今回はみなさんに私の研究プロジェクト「**筆記ランゲージング・学習者の適性・第二言語学習**」への参加をお願いしたいと思います。私は城西大学で英語を教えることに加えて、第二言語習得の分野で「ランゲージング」（例えば独り言をつぶやくなど学習者が問題に出くわした時に使用する言語）の研究を行っています。私は特にその筆記版、筆記ランゲージングに興味があり、これまで5年以上研究を行ってきました。現在はロンドン大学に在籍し、今回みなさんに参加をお願する実験はその博士課程での研究の一部です。

この実験は4週間に渡って行われます。第1週目には、適性テストを実施予定です。第2週目は作文と文法問題のセットに取り組んでいただきます。第3週目は1つのタスクに取り組んだ後、別の作文と文法問題のセットに取り組んでください。第4週目には、また別の作文と文法問題のセットを行います。全ての実験は通常の90分授業の一部（1・2・4週目）または全て（3週目）の時間で、いつもの教室で行われます（ただし、1週目のみコンピュータールーム使用予定）。3週目の授業後、同意が得られた場合に限り数名の方に実験の感想を聞くために短時間残っていただくかもしれません。

この実験は純粋に研究目的であり、実験に参加するか否が成績に影響することはありません。全てのテストを提出するようお願いしますが、テスト結果は成績には全く反映されません。更に、実験途中にいつでも罰則なしに参加をやめることもできます。この実験で得られる結果は研究目的のみに使用され、みなさんの個人的な情報が公開されることはありません。

このシートを読んでいただき、ありがとうございました。実験参加はみなさん次第です。とは言え、みなさんが実験に参加し、良い経験をしたと感じることを願っています。もしも何か質問があれば、ご連絡ください（masako.ishikawa.14@ucl.ac.uk）。このプロジェクトはロンドン大学倫理委員会の承認を得ています。

ロンドン大学博士課程　石川正子 [redacted]
ロンドン大学　カルチャー・コミュニケーション・メディア学部
ベッドフォードウェイ20、ロンドン　WC1H OAL

Institute of Education

UCL

同意書

筆記ランゲージング・学習者の適性・第二言語学習

2016/11/28 ~ 2017/01/16

	はい	いいえ
私はこの研究に関する情報シートを読んで理解しました。	☐	☐
私は情報シートに説明されている実験に参加することに同意します。	☐	☐
私はインタビューされる際には、その内容を録音されることに同意します。	☐	☐
私の解答が報告や発表で使用される際には個人名が明らかにならないことを理解しています。	☐	☐
私は途中で参加をやめられること、その際は私のデータが使われないことを理解しています。	☐	☐
私は実験結果が成績に影響しないことを理解しています。	☐	☐
私はいつでも石川正子[redacted]に連絡できることを理解しています。	☐	☐

氏名 _____

署名 _____

日付け _____

Appendix H: Informed Consent Documentation for the Main Study (English)

Institute of Education

UCL

Information Sheet

Written Languaging, Learners' Aptitude, and L2 Learning

I am your English instructor, Masako Ishikawa, and I am inviting you to take part in my research project, 'Written languaging, learners' aptitude, and L2 learning'. In addition to teaching at Josai University, I conduct research in the field of second language acquisition, focusing on languaging, that is, secondary language use which people make when they are faced with linguistic problems (such as talking to themselves to solve problems). I am especially interested in the written version of languaging, namely, written languaging, and have been conducting research for over five years. Currently I'm enrolled at the UCL IOE, and the experiment I am inviting you to take part in is part of my doctoral studies there.

The experiment will be conducted in four sessions over four weeks. In the first week, you will be given the aptitude tests. In the second week, you will be asked to work on a set of essay and grammar tests. In the third week, you will perform a communicative task, followed by another set of essay and grammar tests. In the final week, you will be asked to work on another set of essay and grammar tests. All the sessions will be conducted in part (1st, 2nd, and 4th sessions) or all (3rd session) of our regular 90-minute class time in our regular classroom (except for the 1st session, which will be conducted in a computer room). In the 3rd week, I may ask some of you to stay for a short time after class for an interview about the experiment only when you agree to do so.

This experiment has nothing to do with your English course, and your decision as to whether or not to participate will not have any effect on your grades. Although I will ask you to submit all the tests, none of the tests results will be reflected in your grade. Moreover, you can stop anytime without penalty. The results which will be obtained in this study will be used only for research purposes and your personal information will never be disclosed.

Thank you very much for taking the time to read this information sheet. It is entirely up to you whether or not you choose to take part. That said, I hope that you choose to participate and that you will find it a valuable experience if you do so. If you have any further questions before you decide whether to take part, you can reach me at masako.ishikawa.14@ucl.ac.uk. This project has been reviewed and approved by the UCL IOE Research Ethics Committee.

Masako Ishikawa, PhD student [redacted]
Department of Culture, Communication, and Media
UCL Institute of Education, University of London
20 Bedford Way, London WC1H OAL

CONSENT FORM

Written Languaging, Learners' Aptitude, and L2 Learning

2016/11/28 ~ 2017/01/16

	Yes	No
I have read and understood the information leaflet about the research.	☐	☐
I agree to take part in the experiment as outlined on the information sheet.	☐	☐
I am happy for my interview to be audio recorded if chosen for an interview.	☐	☐
I understand that if any of my words are used in reports or presentations they will not be attributed to me.	☐	☐
I understand that I can withdraw from the project at any time, and that if I choose to do this, any data I have contributed will not be used.	☐	☐
I understand that the results will not have an effect on my grade.	☐	☐
I understand that I can contact Masako Ishikawa (rtnvish@ucl.ac.uk) at any time.	☐	☐

Name: _____

Signed: _____

Date: _____

Appendix I: Essay Tests (Three Versions)

Essay A:

もしあなたが過去に戻って誰にでも１人だけ会うことができるとしたら、誰に会いますか？　もしその人物に会ったら、何をしますか？　楽しい可能性を色々と考えて、少なくとも3つあなたがその人物とするであろうことを書いてください。「もしも私が過去に戻る (travel back to the past) ことができるとしたら、、、」で文章を始めること。

If you could travel back to the past and meet one person, who would you meet? If you saw the person, what would you do? Think of fun possibilities and write at least three things that you would do with the person. Start your essay with "If I could travel back to the past…"

書き始める前に Before you start:

✧ まずは、誰か1人会いたい人を考えましょう☺ _____
First, pick one person you would like to meet☺

✧ その人としたいことを少なくとも3つ（もちろんもっと多くても！）考えましょう。（文で書く必要はありません。）
Write at least three things (of course, you can write more) you would like to do with the person. (You do not have to write in sentences.)

✓

✓

✓

✧ 考えがまとまったら、右側のページに書きましょう！
When you think you are ready, start writing on the right side of this sheet.

Essay A: Name: _____

Essay B:

もしあなたが１００万円をもらって１週間以内に使うように言われたら、どう使いますか？　楽しい可能性を色々考えて、少なくとも3つあなたがするであろうことを書いてください。「もし私が１週間で１００万円使わなければ (spend one million yen in a week) ならなかったとしたら、、、」で文章を始めること。（貯金はせずに楽しく使い切りましょう☺）

If you were given one million yen and told to use it in a week, how would you use it? Think of fun possibilities and write at least three things that you would do with the money. Start your essay with “If I had to spend one million yen in a week…” (Don’t think of saving and use it up☺)

Essay C:

もしあなたがどこの都市でも訪問してそこに１週間滞在することができるとしたら、どこに行きますか？　そこでの滞在中に何をしますか？　楽しい可能性を色々考えて、少なくとも3つあなたがするであろうことを書いてください。「もし私がどこの都市でも１週間訪問 (visit any city for a week)できるとしたら、、、」で文章を始めること。

If you could visit any city for a week, where would you go? What would you do during your stay there? Think of fun possibilities and write at least three things that you would do. Start your essay with “If I could visit any city for a week…”

Appendix J: Adapted Version of the Modern Language Aptitude Test (Japanese)

名前＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

各番号の上の文を下線部に注意して読み、下の文からその部分と同じ働きをする語句を１つ選び〇で囲みましょう。

例　これはおいしいりんごだ。
　　ボブは世界的に　有名な　アメリカの　(歌手)だ。

１．私たちの学校のチームが、その試合で金メダルを勝ち取った。
　　メアリーは　その歌手が大好きで、コンサート　前から　興奮していた。

２．この図書館には、親切な図書館員が沢山いる。
　　多くの子供たちが澄んだ　水の流れるこの川で毎夏泳ぐ。

３．昨日洗濯機が壊れたために私の母は技術者を呼んだ。
　　先生はクラスの　委員長に　部屋をきれいにするように指示を出した。

４．ヘレンは私に私のお気に入りのクッキーをくれた。
　　私達の美術の先生はリサに　彼女の　絵に対する賞を　手渡した。

５．アンナは毎週末のんびりと家でくつろぐ。
　　この教室ではたくさんの　学生達が熱心に　英語を　勉強する。

６．突然、怒った顧客が入ってきた。
　　失望した　画家はやる気をなくして絵筆を下に投げた。

７．エリックは日本食を好むが、納豆は嫌う。
　　私は運よくタクシーをつかまえたので、時間通り　空港に　到着した。

８．キースはいつも注意深く運転する。
　　学生たちは先生に指示をゆっくりと繰り返すように頼んだ。

９．その少年は彼の子猫をじっと見つめた。
　　その歌手の後ろには、小さな集団の人々がいた。

１０．彼らの父は作家だ。
ユリコは５年にわたり その委員会の 会長をつとめている。

１１．マリアは昨夜すてきな夢をみた。
ブラジルではサッカーが人気だが、トニーは柔道が好きだ。

１２．たった今ジローが通り過ぎて、私に手を振った。
村では 一羽の 雄鶏が毎朝 村の人々を目覚めさせた。

１３．ケイトは日本の冬が好きだ。
このカフェのほとんどの お客が カフェ特製の クレープを注文する。

１４．そのドアは自然に閉まる。
リサはいつも早く起きる、そして、ゆっくり 朝ごはんを食べる。

１５．家を離れている間、ジョンは毎日家族に電話した。
その医師は看護婦に少年の 脚の傷を消毒するように頼んだ。

１６．エミは昨日友人に１０００円貸した。
私の祖父母は毎年私の誕生日に 私に 沢山の プレゼントをくれる。

１７．海外旅行をするには、有効なパスポートが必要だ。
私はあまり お金がないので、中古の コンピューターを買った。

１８．ヒデキは毎朝公園を散歩する。
西の空には真っ赤な 夕日がゆっくりと 沈んでいった。

１９．新キャプテンは責任の重さに耐えぬいた。
そのゴルフ選手は常に 激しい 腰の 痛みを訴えている。

２０．彼の名前はケンタロウだが、みんな彼をケンと呼ぶ。
驚いたことに、コーチは昨日ハナコを次のキャプテンに任命した。

Appendix J: Adapted Version of the Modern Language Aptitude Test (English)

Name_____

Circle one word or phrase in the second sentence that plays the same role as the underlined word in the first sentence.

Example: This is a delicious apple.
Bob is an internationally famous American singer.

1. Our school team won the gold medal in the game.
   Mary loved the singer and was already excited even before the concert.

2. In this library, there are many kind librarians.
   Many children swim in this river with clear water every summer.

3. Because the washing machine broke yesterday, my mother called a mechanic.
   The teacher gave the class president instructions to clean the room.

4. Helen gave me my favorite cookies.
   Our art teacher handed Lisa an award for her picture.

5. Anna relaxes leisurely at home every weekend.
   In this classroom, many students study English hard.

6. All of a sudden, an angry client came in.
   Having lost motivation, a disappointed artist threw the brush down.

7. Erik likes Japanese food, but he hates *natto*.
   Fortunately, I took a taxi and arrived at the airport on time.

8. Keith always drives carefully.
   The students asked the teacher to repeat the instruction slowly.

9. The boy stared at his kitten.
   A small group of people was behind the singer.

10. Their father is a writer.
Yuriko has been a chairperson of the committee for five years.

11. Maria had a nice dream last night.
In Brazil, soccer is popular, but Tony likes judo.

12. Just now, Jiro passed and waved at me.
In the village, a rooster woke up the villagers every morning.

13. Kate likes winter in Japan.
Most of the customers at this café order its signature crepe.

14. The door closes automatically.
Lisa always gets up early and eats breakfast slowly.

15. While he was away, John called his family every day.
The doctor asked the nurse to sanitize the wound on the boy's leg.

16. Emi lent her friend ¥1000 yesterday.
My grandparents give me many presents on my birthday every year.

17. You need a valid passport to travel abroad.
I don't have much money, so I bought a second-hand computer.

18. Hideki takes a walk in the park every morning.
The bright red sun set slowly in the western sky.

19. The new captain managed the burden of his responsibility.
The golf player always complains about his acute back ache.

20. His name is Kentaro, but everyone calls him Ken.
To our surprise, the coach named Hanako the next captain yesterday.

Appendix K: LLAMA_F Worksheet

LLAMA_F

＊LLAMA_F は皆さんが外国語学習上どのような個性や適性を持っているか
を調べ、それぞれに適した学習方法を研究するために開発されたものです。
＊学習パート（5 分間）と試験パート（時間無制限）の２つのパートがあります。
＊パート毎に説明をしますから、質問があれば聞いてください。
*LLAMA_F was developed to examine your characteristics and language aptitude in order to offer instructions that match your characteristics.
*It consists of two parts, a learning part (five minutes) and a test part (no time limit).
* I will explain about each part. If you have any questions, please ask me.

メモ欄: 学習パートで学ぶことをメモするために使ってください。

Memo: You can take notes of what you learn in the learning part below

Name: _____

メモ欄続き：必要ならば裏面も活用してください。

Memo: Please use the other side if necessary.

Appendix L: Metalanguage Knowledge Test

Name: _____

以下の文を読んで、表の文法用語の例を文中から1つ選び書いてください。
（例が1つ以上あることもあります。1 度以上使われる語もあるかもしれません。）
Read the passage below and choose one example for the grammatical features below.
(There may be more than one example. Some words may be used more than once.)

Matilda lived in a small village in England. She was five years old, but she was a genius. Unfortunately her parents were foolish and did not notice it. Although she was unhappy at home, there were two places where she could be happy: her local library and her school. At the library, she could read many books and travel anywhere inside those books. At school, Miss Honey, who was her teacher, noticed her surprising intelligence and treated her kindly.

例）Example) 固有名詞 proper noun: Matilda
可算名詞 countable noun : _____
不可算名詞 uncountable noun : _____
代名詞 pronoun : _____
定冠詞 definite article : _____
不定冠詞 indefinite article : _____
自動詞 intransitive verb : _____
他動詞 transitive verb: _____
動詞の原形 verb root : _____
動詞の過去形 verb past : _____
現在分詞 present participle : _____
関係代名詞 relative pronoun: _____
関係副詞 relative adverb : _____
形容詞 adjective : _____
副詞 adverb : _____
接続詞 conjunction : _____
助動詞 auxiliary verb : _____
前置詞 preposition: _____

Appendix M: Exit Questionnaire for the +WL Group

+ 最後のアンケート + (Exit Questionnaire)

Name: _____

Part I. 実験を振り返って：Reflecting on this experiment:

１．この実験の目的は何だったと思いますか？(言語学習の研究という目的以外に、特に何か調べていると思ったことはありますか？)
What do you think was the purpose of the experiment? (Besides just doing research on language learning, was there something specific you thought I might be studying?)

２．この３週間の実験で英語に関して何か学んだと思いますか？その場合、それは何ですか？
Do you think that you learned something about English during the 3-week experiment? If yes, what?

３．２と関連して、今回学んだ単語や文法のルール等があれば具体的に書いてください。
Related to Q2, are there any vocabulary items or grammar rules you learned? If so, please write them below in detail.

４．特に注目した単語や文法項目等があれば具体的に書いてください。
If there are any vocabulary items or grammar points you focused on, please write them in detail.

５．今回の経験について、何でも自由にあなたの意見を書いてください。（上の回答と重なってもかまいません。）
Please write any other comments you may have regarding this experience. (Your comments may overlap with what you wrote above.)

Part II. 先週原文を確認しながら、考えたこと・疑問に感じたことを書いた経験について：Regarding last week's task of writing your thoughts or questions while checking the original text:

１．自分の考えや疑問を書くということについて、どう感じましたか？ それはなぜですか？
What feelings do you have about the experience of writing your thoughts and questions? Why?

２．今回は原文を確認しながら考えを書いていただきましたが、何も書かずに確認するのと違いがあると思いますか？ それはなぜですか？
In the task, you wrote your thoughts while checking the original text. However, do you think there would be a difference if you hadn't written your thoughts while checking? Why?

3. 書くことで何かを学んだと思いますか？ それはなぜですか？
Do you think you learned something from the experience of writing while checking? Why or why not?

4. 考えや疑問を書くという経験について、何でも自由にあなたの意見を書いてください。
Please write any comments you may have regarding the experience of writing your thoughts while checking the original text.

ご協力ありがとうございました☺
Thank you for your cooperation☺

Appendix N: Skewness and Kurtosis Ratios for All the Tests

RQ1: Essay, Grammar Production and Recognition Tests

	pre-post gains		pre-del post gains	
	skewness	kurtosis	skewness	kurtosis
Essay				
P/C	.831	1.618	1.902	1.599
OC	1.737	.042	.977	1.059
Grammar production				
AU	2.143	1.928	1.030	3.500
OU	-.530	2.580	1.026	3.970
OA	.959	.643	1.744	.029
Recognition				
	-1.214	1.646	-2.380	.962

Note. P/C: points per context, OC: obligatory contexts, AU: accurate use, OU: overuse, OA: overall

RQs 2, 3, 4: Frequency and Quality of T-WLEs

	skewness	kurtosis
Frequency	.173	-.927
Quality	.432	-.869

RQ5: Aptitude Tests

	skewness	kurtosis
MLAT	-2.108	-.237
LABJ	-1.781	-1.386
LLAMA	-1.152	-1.628

RQ6: Metalanguage Knowledge Test

skewness	kurtosis
-1.542	-1.509