CORRECTIVE FEEDBACK IN PERSPECTIVE: THE INTERFACE BETWEEN FEEDBACK TYPE, PROFICIENCY, THE CHOICE OF TARGET STRUCTURE, AND LEARNERS' INDIVIDUAL DIFFERENCES IN WORKING MEMORY AND LANGUAGE ANALYTIC ABILITY

BY

SHAOFENG LI

A DISSERTATION

Submitted to
Michigan State University
In partial fulfillment of the requirements
For the degree of

DOCTOR OF PHILOSOPHY

Second Language Studies

2010

ABSTRACT

CORRECTIVE FEEDBACK IN PERSPECTIVE: THE INTERFACE BETWEEN FEEDBACK TYPE, PROFICIENCY, THE CHOICE OF TARGET STRUCTURE, AND LEARNERS' INDIVIDUAL DIFFERENCES IN WORKING MEMORY AND LANGUAGE ANALYTIC ABILITY

BY

SHAOFENG LI

This study investigates the interaction between feedback type, proficiency, the choice of target structure, and learners' individual differences in working memory and language analytic ability in the learning of Chinese as a foreign language. Seventy-eight L2 Chinese learners from two large U.S. universities participated in the study. The participants were divided into two proficiency levels according to their performance on a standardized proficiency test. At each proficiency level, they were randomly assigned to three conditions: implicit (recasts), explicit (metalinguistic correction), and control. Treatment effects were measured by means of a grammaticality judgment test (GJT) and an elicited imitation (EI) test. Learners' working memory was measured by means of a listening span test, and the Words in Sentences subtest of the MLAT (Carroll & Sapon, 2002) was used to gauge learners' language analytic ability. The study had four sessions. In session 1, the learners took the proficiency test and the GJT pretests; in sessions 2 and 3, the learners took the EI pretest, received implicit or explicit feedback on their nontargetlike use of Chinese classifiers and the perfective *–le* in dyadic interaction, and in the end took the immediate posttest; in the final session (one week after session 3), the learners took the delayed posttests, the working memory test, and the test of language analytic ability.

Results revealed that implicit feedback had limited impact on low-level learners

in the learning of the perfective *–le*, but it was effective for high-level learners; implicit feedback was effective for the learning of Chinese classifiers at both proficiency levels. Explicit feedback was more effective than implicit feedback for low-level learners, but the two types of feedback were equally effective for more advanced learners. At the high proficiency level, the effects of implicit feedback were more sustainable than explicit feedback in the learning of the perfective *–le*. It was also found that in general, the effects related with classifiers were larger than the effects for the perfective *–le* and that EI tests showed larger effects than GJTs. With regard to the interaction between feedback type, the choice of target structure, and the two cognitive factors, language analytic ability correlated with the effects of implicit feedback in the learning of classifiers and the effects of explicit feedback in the learning of the perfective *–le*; working memory correlated with the effects of explicit feedback in the learning of classifiers. Interpretations for these results were sought from multiple perspectives and with reference to previous feedback research.

To my family

ACKNOWLEDGEMENTS

This dissertation project has benefited from many individuals, to whom I would like to extend my sincere gratitude. Dr. Susan Gass, chair of my advisory committee, has been unwavering in her support throughout the duration of my study in the SLS program. Her expertise, insights, inspirations, encouragement, and generous help are critical to my academic growth and will have a life-long impact on my professional development. Dr. Shawn Loewen provided invaluable comments on the research instruments as well as other aspects of my dissertation study; I have greatly benefited from his expertise in form-focused instruction and statistics. Dr. Paula Winke is a versatile and talented scholar, and the way she conducts research and helps students has made her a role model for me in terms of professorship. Dr. Patti Spinner's expertise in theoretical linguistics has allowed me to avoid pitfalls and stay on the right track in my search for the ideal target structures for my study. Dr. Xiaoshi Li has always encouraged me and provided emotional support whenever I experienced setbacks and difficulties. Her knowledge about Chinese linguistics is indispensible to the successful completion of this project. In addition to my advisory committee, I have received selfless assistance from Dr. Debra Friedman, Dr. Debra Hardison, and Dr. Charlene Polio. Joan Reid, secretary of the SLS program, has been very helpful with regard to the logistic aspects of my study.

My gratitude also goes to the Chinese instructors at Michigan State University and the University of Michigan, who encouraged their students to participate in my study. Among them, Shi Liren, Wang Qiongyao, Shi Taiheng, and Teng Chunhong are from Michigan State University; Chen Qinghai, Liu Wei, Tang Le, Yin Haiqing, and Laura Grande are from the University of Michigan. I would also like to thank all the

participants of my study, without whom the study would have been impossible.

Another group of individuals I must thank are my colleagues in the SLS and TESOL programs, who have helped me in various ways. They have provided me with either emotional support or academic assistance. I feel fortunate to be able to have the opportunity to complete my study together with such an excellent group of colleagues; they make my life more enjoyable and my study experience more fruitful. These people include Junkyu Lee, Luke Plonsky, Tomoko Okuno, Grace Lee Amuzie, Jennifer Behney, Soo Hyon Kim, Fei Fei, Kimi Nakatsukasa, Yeon Heo, Allyssa Chamberlain, and Mariah Shafer, to name only a few.

Last but by no means least, I am indebted to my wife Hong Wang and my daughter Ye Li. They are always behind me, and their support, help, and understanding have been a constant incentive for my study. My wife was directly involved in this project: She served as a second coder and provided interrater reliability for the data analysis. I would like to thank her for always being there to help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

This study is conducted in response to Ellis and Sheen's call for investigating variables constraining the effectiveness of corrective feedback (2006) and to the research gaps identified in recent meta-analyses related to the effectiveness of corrective feedback (Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Russell & Spada, 2006). Ellis and Sheen conducted a comprehensive review on previous research on recasts and pointed out that notwithstanding the abundance of research into the effectiveness of recasts, it remains to be seen how the efficacy of this corrective strategy is affected by variables such as the target structure and learners' individual differences. Corroborating Ellis and Sheen's statements, quantitative research syntheses by Li, Lyster and Saito, and Mackey and Goo showed that the effects of corrective feedback are constrained by both learner-internal and learner-external factors. Li's analysis also identified some gaps to be filled in feedback research including the need to examine how feedback facilitates the learning of non-Indo-European languages such as Chinese. This study seeks to address such issues and aims to answer the question of whether the effects of implicit and explicit feedback (operationalized as recasts and metalinguistic correction respectively) are mediated by learners' proficiency, the choice of target structure, and two components of language aptitude—grammatical sensitivity and working memory.

An Overview of Feedback Research in SLA

Corrective feedback in SLA takes the form of responses to learners' erroneous utterances. The responses may indicate that an error has been made, provide the correct linguistic form, supply metalinguistic information about the nature of the error, or contain any combination of these moves (Ellis, Loewen, & Erlam, 2006). There has been

controversy in the field of SLA as to whether corrective feedback plays a facilitative role in developing the learner's interlanguage. The anti-feedback camp (Krashen, 1981; Schwartz, 1993) takes a nativist approach to SLA and identifies L2 acquisition with L1 acquisition. They contend that children learn their first language through exposure to available input and make linguistic generalizations and computations using an inherently built-in Language Acquisition Device; adults learn a second language in the same manner. The L1-L2 equation leads to the argument that L2 acquisition is realized through mere exposure to input in the form of positive evidence (what is acceptable in the target language). Negative evidence (what is unacceptable in the target language) afforded by corrective feedback does not lead to linguistic competence.

Another group of researchers (DeKeyser, 1993; Ellis, 2008; Gass, 1997, 2003; Loewen & Philp, 2006; Long, 1996, 2007; Lyster & Ranta, 1997; Mackey, 2007; Sheen, 2010) voiced their opposition and pointed out that unlike L1 acquisition, adult L2 acquisition requires both positive and negative evidence. Positive evidence is available in the form of a "set of well-formed sentences to which learners are exposed" (Gass, 1997, p. 36). Negative evidence can be provided preemptively through rule-explanation or reactively through feedback to the learner's erroneous utterance. These scholars went on to argue that an optimal condition for L2 learning is negotiated interaction (between learners or between a learner and a language expert such as a native speaker) where the learner notices the gap between his/her erroneous production and the target form and makes subsequent interlanguage modifications. This argument constitutes the core of the Interaction Hypothesis (Gass, 1997; Long, 1996; Mackey, 2007) and lays the ground for the investigation of the effects of feedback. Interactional feedback affords opportunities

for both positive and negative evidence, noticing, and pushed output, all of which are essential to L2 development.

To resolve the controversy over the usefulness of corrective feedback, researchers conducted numerous studies. Studies conducted in laboratory as well as classroom settings (Ammar & Spada, 2006; Han, 2002; Li, 2009; Long, Inagaki, & Ortega, 1998; Lyster, 2004; Mackey & Philp, 1998; Sheen, 2008) have shown that corrective feedback is facilitative to L2 acquisition. Recently, several meta-analyses (Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Norris & Ortega, 2000; Russell & Spada, 2006) have been conducted on the empirical studies examining the effectiveness of corrective feedback and they all showed that feedback, be it oral or written, does benefit L2 learning. These findings undermine the nativist argument against the utility of feedback in L2 instruction.

Focuses of This Study

Corrective Feedback in L2 Chinese

Li's meta-analysis (2010) shows that the effectiveness of corrective feedback varies across different L2s. It is found that English, Spanish, and French are the most frequently investigated target languages in feedback-related research, and L2 Spanish studies showed larger effects than studies involving other target languages. Among the 33 retrieved primary studies in the meta-analysis, there is only one study (unpublished dissertation; Chen, 1996) that examines L2 Chinese. Despite the fact that some interesting findings were obtained, some methodological issues rendered the results less robust. For instance, the study did not include a pretest, and it had a small sample size. After the cut-off date for data collection of the meta-analysis, one study (Li, 2009) was published on how feedback enhanced the learning of Chinese by L1 English speakers.

However, due to the small sample size and failure to include a control group, the generalizability of the results is limited. Therefore, further empirical studies on how corrective feedback fares in L2 Chinese learning is warranted; such studies will enrich and complement interaction-driven SLA research and contribute to the understanding of L2 Chinese learning.

Feedback and Proficiency

Most feedback studies have only examined the effectiveness of one or more feedback types without taking learners' proficiency level into consideration. Mackey and Philp (1998) and Ammar and Spada (2006) are the only studies that investigated the role of developmental readiness in affecting the effects of corrective feedback. Mackey and Philp found that more advanced learners or learners with better mastery of English question formation benefited more from recasts in learning the target structure. Ammar and Spada found that learners with less previous knowledge about the English possessive determiners *his* and *her* benefited more from prompts; for students with more previous knowledge about the structure, prompts and recasts worked equally well.

In both studies, developmental readiness refers to learners' previous knowledge about the target structure. However, it is speculated that learners' general proficiency of the target language may also impact the effectiveness of feedback. Learners with greater proficiency have more attentional resources at their discretion and might therefore benefit more from corrective feedback (Li, 2009). More importantly, as with Ammar and Spada's findings about the interaction between feedback types and learners' previous knowledge about the target structure, different types of feedback may have differential effects on learners at different proficiency levels. Therefore, it would seem misleading

and arbitrary to make claims about the effectiveness of certain feedback types per se as each feedback type possesses characteristics that may work for learners at one proficiency level but not another. This study will investigate how the effects of implicit and explicit feedback (recasts and metalinguistic correction) are affected by learners' general L2 proficiency.

Linguistic Structure

There has been empirical evidence that corrective feedback worked differently for different linguistic structures (Ellis & Sheen, 2006). For instance, in Havranek and Cesnik's study (2001), feedback worked better for verb inflections and auxiliary use than for prepositions and tense choice; in Ishida's study (2004), recasts were more beneficial to the learning of the resultative meaning of the perfective form *–te i-(ru)* than to the learning of the progressive meaning of this structure. However, most of these studies did not examine the nature of linguistic structure as an independent variable.

One study that singled out linguistic structure as an independent variable affecting the effectiveness of corrective feedback is by R. Ellis (2007), who examined the effects of two feedback types (recasts and metalinguistic feedback) on the learning of two structures: the past form *–ed* and the comparative forms in English. He found that recasts did not differentially affect the learning of the two structures, but metalinguistic feedback did: it worked better for the comparative than for the past form. This, according to Ellis, was attributable to the less metalinguistic knowledge the learners had about the comparative than the past form prior to the treatment, which left more room for the increase of metalinguistic knowledge.

In light of the lack of research on the interface between feedback types and different

linguistic structures, this study includes two target structures: Chinese classifiers and perfective aspect marker -*le*, to ascertain whether implicit feedback and explicit feedback have differential effects on the learning of the two structures by learners at two proficiency levels.

Feedback and Language Aptitude

Ellis and Sheen (2006) pointed out that there has also been a paucity of research on how individual difference variables affect the effectiveness of corrective feedback. Among the individual difference variables, aptitude has been shown to be a strong predictor of second language achievements (Dörnyei, 2005). For instance, Oxford (1995) found that among all the individual difference variables she examined, aptitude had the strongest correlation with L2 proficiency. This study will investigate if the effects of implicit and explicit feedback are mediated by learners' language aptitude.

Aptitude-treatment interaction. Early language aptitude testing lays emphasis on its predictive function: It provides information about an individual's likelihood of success or rate of progress in attaining L2 proficiency. Traditionally, language aptitude is viewed as being fixed and is independent of learning conditions or teaching methods (Carroll, 1973, 1993). However, some researchers (Snow, 1994; Segalowitz, 1997) argue that aptitude should not be considered a static characteristic. Rather, it is situated in complex, dynamic, and communicative learning environments that have different processing demands on the learner's cognitive abilities. Robinson (2001) pointed out that the information-processing demands of different learning conditions might facilitate or inhibit learners' cognitive abilities. Similarly, Snow (1987, 1991) advanced the aptitude-treatment interaction hypothesis and suggested a link between aptitude and learning conditions.

Following this line of thinking, researchers have investigated how L2 learners' language aptitude interacted with different input conditions or instruction methods. Wesche (1981) found that learners with high analytic ability achieved more when they were exposed to an analytic teaching approach and that those with good memory and auditory abilities did better under a memory-based functional approach. Robinson (1997) examined the correlation between aptitude as measured by the MLAT (Carroll & Sapon, 1959, 2002) and different learning conditions: incidental, implicit, and explicit. It was found that aptitude correlated with the implicit and explicit conditions but not with the incidental condition. However, in a later study (2002) where he used another set of tests, aptitude correlated with incidental learning. Erlam (2005) found that deductive instruction that gave students opportunities for output minimizes the effect of aptitude variation on learning outcome and that learners with higher language analytic ability and greater working memory capacity benefited the most from instruction that focused on input and that did not require them to engage in language production.

Based on the above arguments and research findings, there is reason to believe that aptitude would correlate differently with the effects of implicit and explicit feedback, two very different learning conditions that supposedly implicate different cognitive processes. Also, aptitude might affect the learning of different linguistic structures as they may set different processing demands on learners. To date, there has been no research that examines the interface between aptitude, feedback, and the nature of the linguistic structure.

Language analytic ability and working memory as aptitude components. L2 aptitude researchers mostly use Carroll and Sapon's aptitude battery (1959, 2002), the MLAT,

which consists of five parts examining four constituent abilities of language aptitude: phonetic coding ability, grammatical sensitivity, rote learning ability, and inductive language learning ability. Accordingly, a composite score based on these four parts of the battery is usually used to gauge learners' language aptitude. However, it is suggested that separate components of aptitude should be examined as they relate to different stages of SLA and are sensitive to different learning conditions (Robinson, 2002; Skehan, 2002). Dörnyei and Skehan (2003) created a scheme that illustrates the aptitude constructs involved in different stages of SLA. For instance, phonemic coding ability is required in the "noticing" stage, and grammatical sensitivity is involved in the "pattern identification" stage.

This study examines the relationship between two feedback types (implicit and explicit) and two aptitude components: language analytic ability and working memory. Language analytic ability, or grammatical sensitivity, is measured by the 'Words in Sentences' subtest of the MLAT. In general, it has been found that language analytic ability has a greater role in classroom learning contexts than in naturalistic settings (Reves, 1983), that adult learners benefit more from analytic ability than child learners (DeKeyser, 2000; Harley & Hart, 2002; Rose, Yoshinaga, & Sasaki, 2002), and that it affects learners in explicit conditions more than learners in implicit conditions (Robinson, 1997).

There have been two studies that examined the interface between corrective feedback and language analytic ability. DeKeyser (1993) found that students with high language aptitude benefited the most from error correction. Sheen (2007a) questioned whether the effectiveness of recasts and metalinguistic correction in the learning of English articles

was mediated by learners' language analytic ability. She found that learners with higher analytic ability benefited more from metalinguistic feedback, but the performance of the recast group did not correlate with language analytic ability.

Among all the aptitude components included in the MLAT battery, the most studied is memory (Skehan, 2002). The MLAT was developed in the context of audiolingual teaching which relies heavily on rote memorization, hence the 'Paired Associates' part that measures associative memory. However, with the advent of communicative language teaching, which requires the learner to attend to form and meaning simultaneously, the predictive power of the memory part of the MLAT has been called into question. Instead of associative memory, working memory has been claimed to have better construct validity and is more predictive of language learning outcome (Robinson, 2002).

Working memory "involves the temporary storage and manipulation of information that is assumed to be necessary for a wide range of complex cognitive activities" (Baddeley, 2003, p.189). Miyake and Friedman stated that working memory is "one (if not the) central component of language aptitude" (1998, p.340). Working memory has been measured by means of digit- or word-span tests, where learners are required to repeat a sequence of digits, words, or syllables; it has also been thorough (reading or listening) sentence span tests that tax the processing as well as storage components (Juffs, 2004). It has been found that results based on sentence span tests are a better indicator of working memory than results based on digit- or word-span tests (Daneman & Carpenter, 1980; Harrington & Sawyer, 1992; Waters & Caplan, 1996).

With regard to working memory in SLA, it has been found that L2 working memory capacity as measured by reading and/or listening span tasks predict reading and listening

comprehension abilities (Osaka & Osaka, 1992; Harrington & Sawyer, 1992; Miyake & Friedman, 1998). Mackey, Philp, Egi, Fujii, and Tatsumi (2002) examined how working memory affected noticing and L2 development as a result of the provision of recasts to learners' erroneous production of English question formation. They found a positive correlation between working memory and noticing. In terms of L2 development, learners with low working memory capacity showed initial improvement and those with high working memory scores achieved more in delayed posttests. Mackey, Adams, Stafford, and Winke (2010) found that working memory is a strong predictor of modified output in dyadic L2 interaction.

To sum up, different feedback types, because of their unique characteristics, set different processing demands and involve different cognitive processes. Therefore, the effects of different types of feedback are likely to interact differently with different cognitive factors involved in SLA such as analytic ability and working memory. However, these factors have been under studied in feedback research and warrant further investigation (Ellis & Sheen, 2006).

In sum, notwithstanding a plethora of research on corrective feedback, it remains to be seen how factors such as proficiency, the target structure, and individual difference variables such as language analytic ability and working memory mediate the effects of feedback. In addition, second language Chinese learning is an understudied area (at least in terms of how feedback impacts the learning of this language), which is in disproportion with the rapid growth of the number of L2 Chinese learners. The need to address the research gaps in feedback research and the lack of L2 Chinese studies necessitates and justifies this study, which probes into how the included variables

contribute, jointly and independently, to the effects of feedback in the learning of a language that is typologically distinct from alphabetic languages such as English or Spanish.

This dissertation report has the following layout. The next chapter, Chapter 2, consists of a review of the literature related to the variables included. Chapter 3 reports on the research methodology of the study including the bio-data of the participants and information on the testing materials, treatment tasks, procedure, data coding, and analyses to be performed. The obtained results appear in Chapter 4, followed by Chapter 5, where the results are discussed and interpretations are sought with reference to previous research and SLA theories. The final chapter, Chapter 6, draws conclusions.

CHAPTER 2 REVIEW OF THE LITERATURE

This chapter provides an overview of previous research on the variables and constructs examined in this study and establishes the rationale underlying the current investigation. The research areas to be reviewed include corrective feedback, the relation of two aptitude components, language analytic ability and working memory, to corrective feedback, and the two target structures included in this study: Chinese classifiers and the Chinese perfective *–le*.

Corrective Feedback

Theoretical Background

Corrective feedback in SLA refers to the response a learner receives to his/her erroneous utterance in the target language, and the response, whether it is from a native speaker or nonnative speaker, is intended to correct the nontargetlike use of a particular linguistic structure. A distinction should be made between corrective feedback and feedback—whereas the former is corrective in nature and is often approached from a pedagogical point of view, the latter is an umbrella term that refers to any response following an erroneous utterance, regardless of whether it is intended to corrective or not. For instance, in either classroom or naturalistic conversations, it is by no means rare that a response occurs following a flawed utterance as a result of the failure to understand the message, in which case the response is a communication move that is not intended to be corrective despite the possibility that the nonnative speaker may perceive it to be. Therefore, corrective feedback in this study is approached from the interlocutor's perspective, that is, its purpose is for the nonnative speaker to be aware that (part of) an L2 utterance deviates from the correct form and/or to modify that utterance based on the

positive and/or negative evidence contained in the feedback.

Whether corrective feedback is useful for second language development is essentially a question of what type of input is necessary for learning to happen. The learner has access to two types of input (Gass, 1997): positive evidence and negative evidence. Positive evidence refers to what is acceptable in the target language and negative evidence informs learners of what is unacceptable. Corrective feedback contains negative evidence (although some feedback types might also contain positive evidence). Therefore, to acknowledge the role of corrective feedback is to endorse the value of negative evidence in second language learning. While the importance of input is recognized in all language learning theories, researchers and theorists are divided on whether input in the form of positive evidence is sufficient or both positive evidence and negative evidence are necessary.

The nativists (Krashen, 1995; Schwartz, 1993) insist that language is acquired as an abstract system of mental representation that is realized through a language acquisition device (Universal Grammar [UG]) that is inherent to human beings. It is argued that adults learn a second language in the same way as children learn their first language: Exposure to positive evidence is sufficient and no negative evidence is necessary. Therefore, any attempt to draw the learner's attention to linguistic forms, by either preemptive rule explanation or corrective feedback following learners' errors, is futile and should be avoided. As Krashen (1995) pointed out, "a safe procedure is simply to eliminate error correction entirely" (p. 76). An immersion class that is based on this model would be one where students read and listen to materials in the target language or learn the language through the subject matter and where the instructor does not address

linguistic forms or give any feedback to students' errors.

The interactionists (Gass, 1997; Long, 1996; Pica, 1988) believe that adults learn a second language differently from the way children learn their first language. In their view, both positive evidence and negative evidence are necessary, and hence the need to attend to linguistic form. In fact, the term "form-focused instruction" (FFI) is created in response to or contrast with meaning-based instruction that suppresses any attention to form (Ellis, 2001; Doughty & Williams, 1998; Spada, 1997). FFI refers to any attempt to draw the learner's attention to linguistic form (Spada, 1997). While there are various options of FFI (Loewen, 2005), one optimal condition in FFI, according to the Interaction Hypothesis (Gass, 2004; Long, 2007), is negotiated interaction where the learner notices the gap (such as through corrective feedback) between his/her wrong L2 production and the target form and makes subsequent modifications to his/her interlanguage. As Long (1996, p.414) stated:

> It is proposed that environmental contributions to acquisition are mediated by selective attention and the learner's developing L2 processing capacity, and that these resources are brought together most usefully, although not exclusively, during *negotiation for meaning* [emphasis original]. *Negative feedback* [emphasis added] obtained during negotiation work or elsewhere may be facilitative of L2 development, at least for vocabulary, morphology, and language-specific syntax, and essential for learning certain specifiable L1-L2 contrasts.

The usefulness of corrective feedback is also backed up by other SLA theories. According to Schmidt's Noticing Hypothesis (1990, 2001), unlike first language acquisition, second language acquisition is conscious. Schmidt stated that "subliminal

language learning is impossible…[and] noticing is the necessary and sufficient condition for converting input to intake" (p. 129). Corrective feedback contributes to the noticing of linguistic form. Another benefit of corrective feedback is the learner's responses following feedback (referred to as "uptake") (Loewen, 2004). Learner uptake is one form of output, which, according to Swain (1995, 2005), has three functions: noticing/triggering, hypothesis testing, and metalinguistic reflection. The effect of corrective feedback is also grounded in Socio-Cultural Theory, which holds that corrective feedback serves as a form of regulation in the zone of proximal development that can "be appropriated by learners to modify their interlanguage systems" (Aljaafreh & Lantolf, 1994, p. 480). Recently, the role of corrective feedback has been associated with skill acquisition theory (DeKeyser, 2007, 2008; Ellis, 2010; Lyster & Iquierdo, 2009), according to which L2 acquisition involves the transition from declarative knowledge to procedural knowledge and ultimately to automatic knowledge. And corrective feedback affords practice opportunities that contribute to this transition.

Taxonomy of Feedback

Empirical research on corrective feedback has mushroomed since the 1990s when the theoretical rationale had been established for its role in SLA. Early feedback research was conducted from the perspective of interaction, that is, how negotiated interaction where feedback is embedded facilitates second language development (Gass & Varonis, 1994; Mackey, 1999; Polio & Gass, 1998). Though feedback was not teased out as an independent variable in interaction studies, they supplied empirical evidence for the usefulness of feedback and provided impetus for subsequent feedback research because to a large extent, negotiated interaction contributes to L2 learning due to the presence of

feedback.

In terms of how feedback types are categorized, there are two schemes. Lyster and his colleagues (Lyster, 1998, 2001; Lyster & Ranta, 1997) conducted extensive research on the occurrence of corrective feedback in some French immersion classes in Canada and identified seven types of feedback: recasts, elicitation, clarification, metalinguistic comments, repetition, and explicit correction. These seven types of feedback are further divided according to whether they encourage learner repair: Elicitation, clarification, metalinguistic comments, and repetition are collectively called prompts; recasts and explicit correction provide the correct form and therefore lead to less learner repair. Sheen (2010) and Ellis (2010) made a similar distinction by pointing out that feedback can be input-providing (recasts and explicit correction) and output-prompting (prompts). In the other scheme, feedback is classified as implicit or explicit, depending on whether a feedback type explicitly draws the learner's attention to linguistic form (DeKeyser, 1993; Ellis, Loewen, & Erlam, 2006). Following this scheme, recasts, clarification, elicitation, and repetition are implicit; explicit correction and metalinguistic feedback are explicit (Lyster, 1998; Li, 2010).

While both categorization schemes are reasonable in their own right, they might have their respective problems. The implicit-explicit dichotomy is undermined by the fact that some implicit feedback types such as recasts can be explicit (e.g., Doughty & Valera, 1998). The taxonomy of feedback based on how much repair is generated masks the explicitness of feedback. For instance, in the "prompts" category, metalinguistic feedback and elicitation are explicit, but clarification and repetition are implicit. Another problem with prompts is that all four types of feedback are placed under this umbrella category,

which makes one question the extent to which it is reasonable to compare multiple corrective moves with a single move such as recasts (e.g., Lyster, 2004; Ammar & Spada, 2006). It is not easy to find a solution to the controversy or find a perfect way to categorize feedback types. Probably the best researchers can do is to maximize the implicit-explicit contrast when implicitness/explicitness is a key variable, and to interpret the differential effects of prompts and recasts based on the different cognitive processes involved as well as the amount of generated repair when these two feedback types are investigated (Yang & Lyster, 2010).

The relationship between the two categorization schemes (explicit vs. implicit and input-providing vs. output-prompting) is illustrated in Figure 1 (also see Loewen & Nabei, 2007). However, it is evident that the implicitness or explicitness of feedback stands in a continuum and is contingent upon many factors. Therefore, the position of a certain feedback move as illustrated does not necessarily indicate it is more or less implicit or explicit than the feedback type next to it. Note that in Figure 1, metalinguistic correction (metalinguistic clue + correct form) is added to the list of feedback types identified by Lyster and Ranta (1997). Metalinguistic correction has been investigated in previous research (Sheen, 2007a) and it is also one type of feedback included in this study.

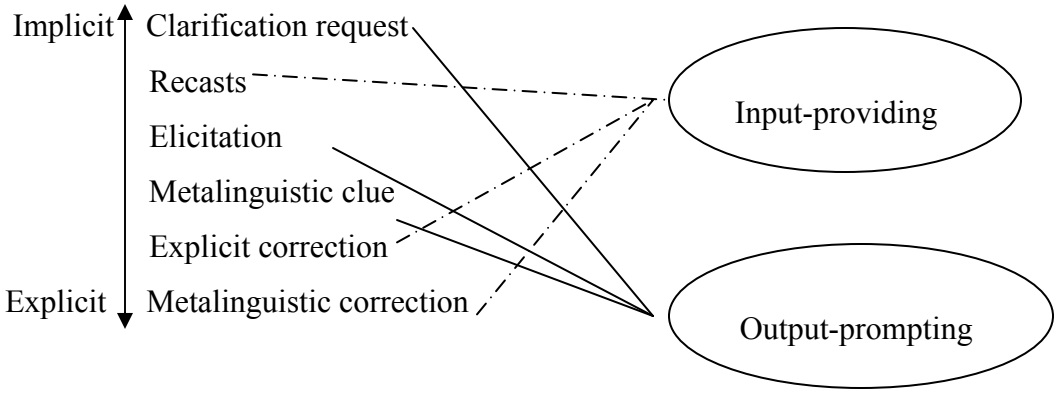


Figure 1. Taxonomy of feedback

Recasts

This study investigates the effects of two major feedback types: implicit feedback in the form of recasts and explicit feedback operationalized as metalinguistic feedback. Recasts refer to partial or full reformulation of the learner's erroneous L2 utterance. Among the various corrective strategies, the recast is the most studied, which is not surprising given its high frequency in the classroom as well as the sound theoretical justification for its usefulness. Classroom descriptive studies (Lyster, 1998, 2001; Lyster & Mori, 2006; Lyster & Ranta, 1997; Sheen, 2004; Sheen, 2006) showed that the recast was the most frequent feedback type in all instructional settings including immersion classes and classes of ESL and EFL. Long (1996, 2007) argued that the recast is optimal for form-focused instruction because it addresses linguistic forms when the primary focus is on meaning; it shifts the learner' attention away for a brief focus on form and juxtaposes the erroneous form with the target form, allowing for a cognitive comparison and priming the learner to notice the difference between the two. Also, the recast makes both types of input, positive evidence and negative evidence, available to the learner. This makes it possible for the learner to retrieve and rehearse pre-existing linguistic knowledge or benefit from the exposure to a provided language exemplar if the target form is unavailable or fails to be retrieved from the interlanguage repertoire.

Recasts have been shown to be effective in laboratory studies (Carroll & Swain, 1993; Egi, 2007; Ishida, 2004; Long, Inagaki & Ortega, 1998; Iwashita, 2003; Leeman, 2003; Li, 2009; Mackey & Philp, 1998; Lyster & Izquierdo, 2009; McDonough, 2007; Sagarra, 2007). These studies are typically carried out in dyadic interaction (except for Sagarra's study where feedback was provided through the computer) where learners

received intensive recasts on a single structure. Methodological features such as the lab setting, provision of feedback on a one-on-one basis, and targeting one structure might have made recasts relatively salient and therefore benefited L2 development. One might argue that the generalizability of laboratory findings to classroom contexts is questionable, which is to some degree legitimate. However, in laboratory studies, variables can be easily teased out and better controlled, distractions are minimized, and the obtained results might therefore be more reliable.

Quasi-experimental studies where students received feedback as a class or group showed that recasts were less effective than more explicit feedback types such as prompts and metalinguistic feedback (Ammar & Spada, 2006; Ellis, 2007; Ellis et al., 2006; Lyster, 2004; Sheen, 2007; Yang & Lyster, 2010). However, Han (2002) and Doughty and Varela (1998) showed that recasts can be very effective when targeting multiple learners if the feedback was intensive, targeted a single structure, and was made salient. In Han's study, learners received treatment in 11 sessions on the learning of past tense consistency. In Doughty and Varela's study, recasts were operationalized as repetition of the learner's nontargetlike utterance with a rising tone followed by a recast.

There has also been research on the level of uptake recasts generate and the characteristics of recasts that affect uptake (Ellis, Basturkmen, & Loewen, 2001; Loewen, 2004; Loewen & Philp, 2006; Lyster, 1998, 2001; Lyster & Mori, 2006; Lyster & Ranta, 1997; Panova & Lyster, 2002; Sheen, 2004, 2006). Uptake refers to the learner's response following feedback. Taken together, these studies showed that recasts led to more uptake in language programs than in immersion programs, and that uptake and successfulness of uptake related to the characteristics of recasts and the characteristics of

the form-focused episode that contains the recast. Also, uptake might relate to the nature of the target structure. For instance, recasting lexical or phonological errors might generate more uptake than recasting morphosyntactic errors.

Metalinguistic Feedback

The other feedback type this study investigates is meta-linguistic feedback. Meta-linguistic feedback takes two forms: It may refer to comments on the well-formedness of the learner's L2 production (metalinguistic comments/clues) (Carroll & Swain, 1993; Ellis et al., 2006) or to the provision of the correct form followed by metalinguistic comments (Li, 2009; Sheen, 2007). Some researchers (e.g., Ellis et al., 2006) opted for the former operationalization probably because the nature of the target structure is such that the provision of some metalinguistic comments was sufficient for the learner to retrieve and/or internalize the rule and make a correction (as in "You need past tense here"). Other researchers chose the latter operationalization probably because the target structure has some variants and providing some comments alone may not lead to the modification of the wrong form (such as in Li's study where the target structure was classifiers and a comment such as "You used a wrong classifier" was unlikely to make the learner use the correct classifier if it is not part of their interlanguage). Sheen (2007) justified her decision to combine metalinguistic comments with the provision of the correct form by claiming that it is more effective than supplying metalinguistic comments alone. Further support for Sheen's operationalization comes from Ellis (2007), who suggested the principle of "bias for best", that is, operationalizing a feedback type to maximize its potential effect. There is also empirical evidence that a brief metalinguistic comment may not lead to any linguistic development (Loewen & Nabei, 2007).

Researchers have studied the effectiveness of metalinguistic feedback as compared with recasts and/or other implicit feedback types (Carroll & Swain, 1993; R. Ellis et al., 2006; Loewen & Nabei, 2007, Sheen, 2007). Carroll and Swain (1993) investigated the effects of four feedback types—metalinguistic feedback, explicit hypothesis rejection, explicit utterance rejection, and recasts—on 100 ESL learners in the learning of English dative alternation. They found that metalinguistic feedback worked better than all other included feedback types. Kim and Mathes (2001) replicated Carroll and Swain's study but only included metalinguistic feedback and recasts in the replication. They failed to find any differences between the two feedback types, which might be attributable to the small sample size of the study (n = 10 in each group vs. n = 20 in each group in Carroll & Swain's study). Ellis et al. (2006) examined the differential effects of metalinguistic feedback and recasts on the learning of past tense *–ed* by 34 low-intermediate ESL learners. A superior effect was found for metalinguistic feedback over recasts and overall feedback contributed more to learners' implicit knowledge than their explicit knowledge. Loewen and Nabei's study (2007) included three feedback types: metalinguistic feedback, recasts, and clarification. Participants were two intact classes of Japanese EFL learners and the target structure was the English question formation. Results revealed that the learners only showed improvement on the timed grammaticality judgment test (the other measures are untimed grammaticality judgment test and oral production test) and that no differences were found between the experiment groups. The failure to find a superior effect for metalinguistic feedback was attributed to the insufficient metalinguistic information provided to the learner given the complex nature of the target structure. Sheen (2007a) studied the effects of metalinguistic feedback and recasts on the learning

of English article use by 80 ESL learners. A significant effect was found for metalinguistic feedback, but not for recasts.

Studies following Lyster's taxonomy of feedback (1998, 2001) classified metalinguistic feedback (provision of metalinguistic information) as a prompt (other prompts include clarification, elicitation, and repetition) and investigated prompts as one type of feedback in comparison with recasts. In general, these studies (Ammar & Spada, 2006; Lyster, 2004; Yang & Lyster, 2010) showed that prompts were more effective than recasts. Since metalinguistic feedback was conflated with other feedback types as prompts in these studies, it is difficult to know the extent to which it contributed to learning as a single corrective move.

The Effectiveness of Corrective Feedback: Toward an Integrated Model

There has been increasing evidence that the effectiveness of corrective feedback is subject to multiple factors and therefore should not be approached from the perspective of the properties of feedback per se. The accumulation of empirical research has made it possible for several meta-analyses to be conducted on how the role of feedback in SLA is constrained by various learner-internal and learner-external factors. Russell and Spada's meta-analysis showed that oral feedback was more effective than written feedback (2006). Lyster and Saito (2010) meta-analyzed 15 classroom-based studies and found that prompts worked better than recasts and feedback showed larger effects on oral production tests. The meta-analysts also found that younger learners benefited more from the effects of feedback.

Li (2010) included both published studies (n = 22) and Ph.D. dissertations (n = 11) in his meta-analysis and found that explicit feedback showed larger short-term effects but

the effects of implicit feedback were better retained. More importantly, the meta-analysis identified multiple factors mediating the effects of feedback. Specifically, studies conducted in foreign language contexts showed larger effects than studies conducted in second language contexts; lab-based studies shower larger effects than classroom-based studies; feedback provided in mechanical drills yielded a larger effect than feedback provided in communicative tasks; and similar to what Lyster and Saito found, feedback showed larger effects on free production tests (such as oral production) than on constrained production tests (such as grammaticality judgment test). The study also demonstrated a possible effect of interlocutor type (native speaker vs. nonnative speaker), mode of delivery (face-to-face vs. virtual), duration/intensity of treatment (short vs. long), and cross-linguistic differences on the effects of feedback.

Aside from the factors mentioned above, there has been evidence that the effects of corrective feedback are mediated by learners' proficiency (Mackey & Philp, 1998; Ammar & Spada, 2006), nature of the target structure (Ellis, 2007; Yang & Lyster, 2010), noticing (Egi, 2007; Philp, 2003) and individual learner differences (DeKeyser, 1993; Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Sheen, 2007, 2008). Narrative reviews by Nicholas, Lightbown, and Spada (2001) and Ellis and Sheen (2006) provided comprehensive and in-depth discussion of feedback related studies and the constructs involved in feedback research. These scholars also called for an integrated approach to the investigation of corrective feedback.

Based on the findings of quantitative and narrative syntheses on feedback research as well as those of primary research, I propose an integrated and interactive model on the constructs and variables underlying the effectiveness of corrective feedback (Figure 2).

This model recognizes the independent and joint effects of various factors affecting the role of feedback in L2 learning. These factors relate to the characteristics of feedback proper, linguistic properties of the target structure and target language, the context in which feedback is supplied, and learner differences. Acknowledging the interaction between these factors has two implications: One is to alert researchers to the possibility and necessity of interpreting the obtained results with reference to other relevant variables when a single variable is examined; the other is to prompt researchers to investigate the interaction effects of multiple variables.

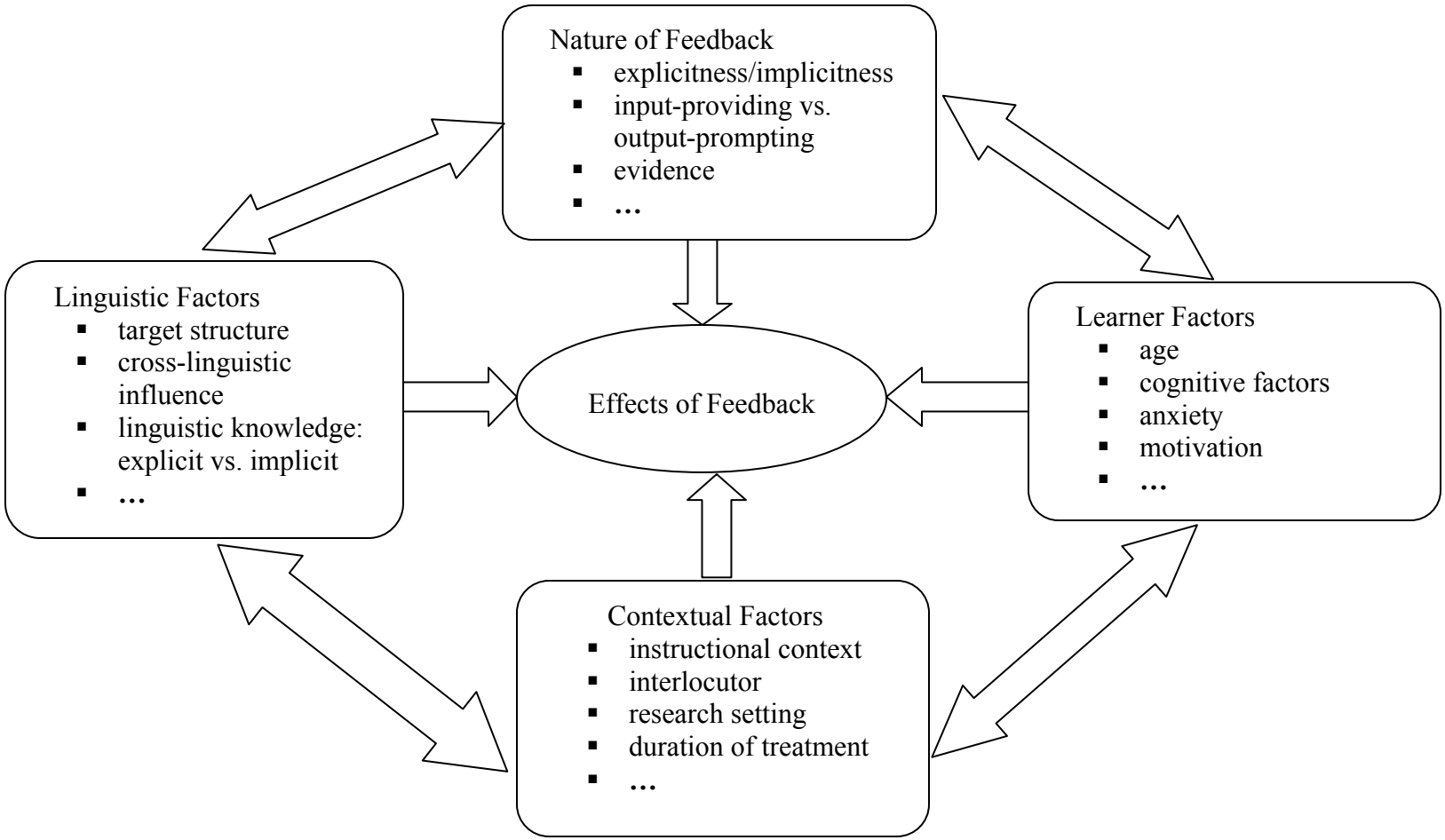


Figure 2. An integrated model of corrective feedback

Feedback, Proficiency, and the Target Structure

Proficiency and the Choice of Target Structure

Among the various variables that potentially mediate the effects of feedback, one that needs further investigation is learner's proficiency level. Philp (2003) found that more advanced learners were more likely to notice the reformulation of their wrong L2 production. Mackey and Philp (1998) found that ESL learners who were more developmentally ready benefited more from recasts in learning English question formation. Ammar and Spada (2006) investigated the differential effects of prompts and recasts on the learning of third-person possessive determiners in English (*his*/*her*) by 64 Francophone students. The study also examined whether students who scored higher on the pretests benefited more from feedback. It was found that lower-level learners benefited more from prompts but higher-level learners benefited equally from prompts and recasts. These studies showed that the effectiveness of feedback related to the learner's previous knowledge about the target structure.

While the effects of feedback have been shown to be mediated by how much the learner already knows about the target structure, the learner's general proficiency in the target language may also play a role. To date, there has been only one study that included the learner's proficiency level as a variable. Li (2009) investigated how recasts and metalinguistic feedback facilitated the learning of Chinese classifiers. The participants were 23 students from second- and fourth-year Chinese classes at a U.S. university. It was found that metalingistic feedback was more beneficial to the second-year students, but there was no significant difference between the two feedback types as far as the fourth-year students were concerned. The generalizability of the results is limited because

of the small sample size, failure to include a control group, and group assignment based on the students' enrollment status rather than a proficiency test.

The effects of feedback can also be constrained by the target structure. In other words, different feedback types may work differently for different structures (Nicholas, Lightbown, & Spada, 2001; Ellis & Sheen, 2006), and there is empirical evidence for this claim. For instance, in Long, Inagaki and Ortega (1998), recasts were effective for adverb placement but not for object topicalization in L2 Spanish learning. In Iwashita (2003), recasts benefited the learning of *te*-form verbs but not of the two locative-initial targets. Also, feedback worked differently in different studies although the research settings were similar, which might be attributable to, among other factors, the fact that different target structures were included. For instance, both Ammar and Spada (2006) and Lyster (2004) investigated the effects of recasts and prompts in immersion classes. While Ammar and Spada found that recasts facilitated the learning of English third person possessive determiners (his/her), participants in Lyster's study did not benefit from recasts when learning French gender agreement. Sheen (2007) conducted a classroom study on the effectiveness of metalinguistic feedback and recasts, and found that recasts did not work for the learning of English articles. It should be noted that in these studies, the choice of target structure is not an independent variable, despite the conflicting findings that are likely to result from the different linguistic structures they included.

To date there have been two studies that examined the choice of target structure as a variable affecting the effectiveness of corrective feedback. One is by Ellis (2007), and the other is by Yang and Lyster (2010). Ellis investigated whether recasts and metalinguistic feedback have differential effects on the learning of the English past tense *–ed* and

comparative *–er*. 34 adult ESL learners were randomly assigned to three conditions: recasts (*n* = 12), metalinguistic (*n* = 12), and control (*n* = 10). The study is quasi-experimental in that it was conducted in the classroom. Results showed that recasts did not promote the learning of either structure, but metalinguistic feedback did. Also, the effects of metalinguistic feedback on the comparative were immediate and its effects for the past tense *–ed* were delayed. Ellis speculated that this was because prior to the treatment, the learners did not have much explicit knowledge about the comparative but they did about the past tense *–ed*.

Yang and Lyster (2010) investigated the effects of prompts and recasts with 72 Chinese EFL learners. The target structures were regular and irregular English past tense. It was found that prompts showed an advantage over recasts on 8 measures and that while prompts worked better than recasts in assisting the acquisition of regular past-tense forms, both feedback types worked equally well for the learning of irregular past-tense forms. The researchers stated that the general superiority of prompts lied in the fact that they led to more self-repair and were more salient than recasts. Recasts were more effective for the learning of irregular past forms than that of regular past forms because of the greater saliency the former had. Prompts outperformed recasts in the learning of regular past forms due to the negative evidence and opportunities for self-repair afforded in prompts. The researchers continued to argue that the reason why learners benefited equally from prompts and recasts in learning irregular past forms was probably because the structure was item-based. Item-based learning profited from either the positive evidence available in recasts or negative evidence coupled with self-repair, which prompts entailed.

Taken together, these two studies as well as studies that did not include the choice of

target structure revealed that the effects of feedback indeed related to the nature of the target structure. The differential effects resulted from multiple factors that may include the unique attributes of different feedback types (explicitness/implicitness, evidence, and learner repair), the linguistic features of the target structure (saliency and/or rule-based vs. exemplar-based), and learners' individual differences (aptitude, anxiety, motivation, and so on). To date, there has been no study on how learner-related factors might affect the differential effects of different types of feedback on the learning of different linguistic structures. This study seeks to fill this gap by examining how two aptitude components, language analytic ability and working memory, might impact the effects of recasts and metalinguistic feedback in the learning of two different Chinese structures.

Chinese Perfective -le and Chinese Classifiers

Chinese perfective –le. Typologically, Chinese is different from Indo-European languages. One distinctive feature of the Chinese language is that it has a limited number of functional categories, among which the most studied is the aspect markers: the perfective *–le* and *–guo*, and the progressive *zheng-* and *–zhe.* Because of its prominence in the language, aspect has been considered one of the most defining features of Mandarin Chinese and it has been frequently utilized to exemplify aspect languages (Comrie, 1976; Smith, 1997; Xiao & McEnery, 2006). Probably the most extensively studied Chinese aspect marker is *–le*, which is mainly due to its high frequency and the controversy over its syntactic and/or semantic interpretations. This study investigates how two types of feedback impact the learning of *–le* by second language Chinese learners. Before exploring how instruction affects the acquisition of this structure, it is necessary to provide a detailed description about its linguistic characteristics. To have a

full understanding of how *-le* is used, it is important to define the concept of aspect and make a distinction between grammatical aspect and lexical aspect.

Not to be confused with tense, which indicates the relationship between event time, the time when the event actually takes place, and speech time, the time when the event is addressed, aspect is concerned with the relationship between event time and reference time, the time which is used as a reference point for the event. Aspect can be represented either grammatically or lexically. Grammatical aspect refers to aspectual distinctions realized through linguistic devices such as the use of auxiliaries and affixation (Li & Shirai, 2000). Lexical aspect, alternatively known as situation aspect, inherent aspect, or Aktionsart, is marked by the inherent characteristics of lexical items. To identify the lexical aspectual features of a verb, a binary system has been developed using such dimensions as telicity, punctuality, and dynamicity.

According to Vendler (1957), verbs are classified into four types according to the temporal attributes they display, which are states, activities, accomplishments, and achievements. Smith (1997) modified Vendler's system by adding "semelfactive" verbs. States verbs are used to describe situations that are homogeneous and have no successive phases or endpoints; activities verbs describe situations with successive phases but without endpoints; accomplishment verbs encode situations with successive phases and a natural endpoint; achievement verbs are also used to encode situations with a natural endpoint, but they are different from accomplishment verbs in that the events are punctual, instantaneous, and without time duration; semelfactives are punctual but they have no endpoint. In addition, the unique groups of verbs called resultative verb constructions (RVCs), according to Li and Shirai (2000), should be considered

achievements. RVCs are, as it were, combinations of accomplishments and achievement.

To better understand the semantics of the five types of verbs, the visual representation (Table 1) by Anderson (1990, in Li & Shirai, 2000) might be of assistance. To the original scheme, I added an illustration for RVCs and semelfactives.

Table 1. Schematization of the semantic properties of verb types

Type	*Illustration*	*Example*
State	————	*love, contain, know*
Activity	----------	*run, walk, swim*
Accomplishment	----------X	*paint a picture, build a house*
Achievement	X	*fall, drop, win the race*
Achievement (RVCs)	-------X-----X	*dǎkāi* (push + open), *shuāidǎo* (slip + fall)*
Semelfactive	---X---X---X---	*cough, tap, knock*

* These two examples are given in Pinyin, the Romanization system of the Chinese characters.

As a perfective aspect marker, *-le* encodes an event in its entirety. It occurs with situations that are [+bounded] or [+telic]. As to the interaction between lexical aspect and grammatical aspect, *-le* is naturally compatible with accomplishment and achievement verbs. For verbs (states, activity, and semelfactive verbs) that encode atelic situations, that is, situations without an endpoint, to be used with *–le*, an external device (usually a quantifier) needs to be added to set a beginning and end point or a boundary for the event. The following sentences illustrate how *–le* is used to indicate perfectivity.

(a) tā shuāidǎo le.

他 摔倒 了。
He fall-*Perf*
He fell.

(b) tā pǎo le shíwǔ fēnzhōng.
他 跑 了 十五 分钟。
He run-*Perf* fifteen minutes.
He ran for fifteen minutes.

In sentence (a), the verb *shuāidǎo* (fall) is an achievement verb and has a natural endpoint. In sentence (b), the verb *pǎo* (run) is an activity verb without a natural endpoint, but the time duration *shíwǔfēnzhōng* (15 minutes) delimits the situation to license the use of *–le*. It must be pointed out that although theoretically a delimiting device can be added to a situation encoded by a state verb such as *xǐhuān* (like) to allow the use of the perfective marker *–le*, the combination of *–le* with state verbs is rare in actual communication.

There has been a controversy over *le*'s interpretations in relation to the distinction between the verbal -*le* and the sentence final -*le* (Van den Berg & Wu, 2006). One view holds that there is only one *–le*, which marks either termination or completion (Chang, 2002; Shi, 1990; Thompson, 1968; Yang, 2003); others maintain that there are two *–le*s: a verbal *–le* which marks perfectivity and a sentence final *–le* which marks inchoativity or change of state of affairs (Li & Thompson, 1981; Liu, 2001; Van den Berg, 1989; Xiao & McEnery, 2004) (see Figure 3 for an illustration of the controversy over the verb *–le* and the sentence final -*le*).

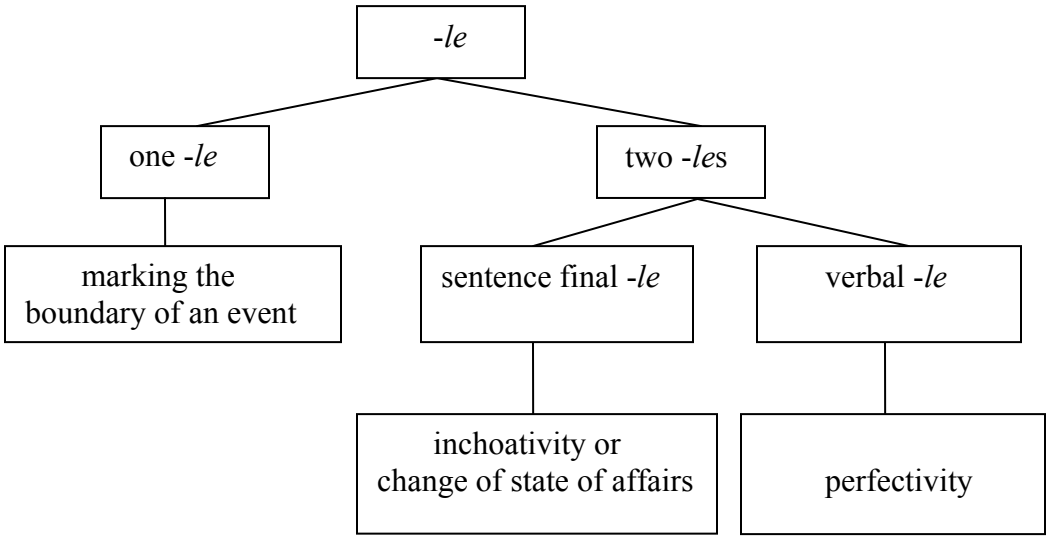


Figure 3. An illustration of the one-*le* and two-*le* controversy

The two-*le* view is more reasonable and the following examples show how the verbal *–le* differs from the sentence-final -*le*. As shown, in (a), (c), and (e), the verbal *–le* indicates completion and the sentence final *–le* in (b) and (d) describes current relevance. Sentence (a) suggests that "I did not eat dinner at first, but now I have." In (b), -*le* indicates that the activity of buying was completed. (c) means that it did not rain at first, but now, it has started to rain. In (d), the event of raining lasted for three days and it was completed some time in the past. (e) in fact involves a future event, indicating that the situation will change from "guests staying" to "guests leaving".

a. tā chī fàn le.
她 吃 饭 了。
He eat meal *Perf*
He has eaten dinner.

b. tā mǎi le yī běn shū.
他 买 了 一 本 书。
He buy-*Perf* one-*CL* book.
He has bought a book.

c. xià yǔ le.
下 雨 了。
Fall rain-*Perf*
It has started to rain.

d. xià le sāntiān yǔ.
下 了 三天 雨。
Fall-*Perf* three day rain.
It rained for three days.

e. kèrén yào zòu le.
客人 要 走 了。
Guest will leave-*Asp*.
The guests are leaving.

It is not a goal of this paper to resolve the one-*le* versus two-*le* controversy. The objective of this project is to explore the effectiveness of corrective feedback in the learning of *–le* and the target structure is the verbal *–le* as the usage of *–le* is so complex that it would be difficult for the learner to acquire all the uses of the structure through a short instructional treatment. The following section reviews the previous studies on the acquisition of *–le* by second language speakers of Chinese.

There have been some studies on the acquisition of *–le*, all of which are descriptive and none concerns how instruction facilitates the learning of the structure (Duff & Li, 2002; Christensen, 1994; Wen, 1995, 1997; Yang, Huang, & Sun, 1999, 2000). These studies either investigated *–le* alone or the acquisition of *–le* with other aspect markers. Wen's studies dealt mainly with the acquisition of the aspect markers, and Yang et al.'s two studies examined the underuse of markers and the interaction between aspect markers and verb types. Wen's first study investigated L2 learners' acquisition of the perfective *–le*. The subjects were 14 L2 Chinese students at an American university who were L1 English speakers. Among them, six had studied Chinese for 14 months and were

considered beginners, and eight were advanced learners who had been in the program for 26 months. They were interviewed three times and engaged in three tasks: two conversational tasks and one picture-description task. It was found that there was no difference between the two levels of learners with regard to the accuracy rate in their use of the verbal *–le*, but the advanced learners performed better than the beginners in the use of the sentence final *–le*. The author claimed that this was because of the complex nature of the sentence final *-le*. It was further pointed out that the learners' correct use of the marker had to do with whether two events were involved, whether there was a duration of time, whether a sentence ended in a mono-syllabic verb, and whether the adverb *yǐjīng* (already) was present. The results were discussed in relation to L1 transfer. For instance, at the beginning level, the learners only used the perfective marker with past events and avoided using it with future events (to indicate inchoativity) because it was not allowed in their L1; some beginners also omitted the sentence final *–le* in obligatory contexts because there was not such a feature in their L1.

In another study (1997), Wen investigated L1 English speakers' acquisition of the perfective marker *–le*, the experiential marker *–guo*, and the durative marker *–zhe*. The participants were 19 students who were studying Chinese at an American university, and they were split into two levels: 10 from the lower level (who had studied Chinese for 15 months at the time of the study) and 9 from the higher level (who had studied Chinese for 27 months). The data was elicited through the use of two tasks: an interview and a picture description task. The results showed that in general there was no difference between the two levels in the number of the three aspect markers produced, but learners of higher proficiency were more accurate in using the markers. Furthermore, while the advanced

learners were more accurate in their use of *–zhe*, they did not outperform the less proficient learners in using *–le* or *–guo*. Wen also found that the perfective markers were acquired before the imperfective aspect and she argued that this is because of the semantic salience, syntactic simplicity, and pragmatic consistency of the former. One caveat about the study is that at the beginning of both tasks, as a data-soliciting prompt, a question was asked that contained the aspect marker to be used to perform the task. It is suspected that the question was likely to serve as a model for the use of the aspect marker, calling into question the reliability of the results.

Whereas Wen's studies investigated L2 Chinese learners at American universities, Yang et al.'s studies (1999, 2000) involved learners studying Chinese in mainland China, a second language setting. Yang et al.'s first study examined the use of three aspect markers *–le, -guo, and -zhe*, and the data was extracted from a corpus containing the narrative writings of the students at a 4-year Chinese program[1] at Beijing Languages and Cultures University. The narratives were from all levels of students (eight levels were identified) and were assorted: they were either timed or untimed, and were collected either inside or outside of class. It was found that the learners' use of *–le* did not improve with the increase of their proficiency, but they did make fewer errors in using *–guo* and *–zhe* when they reached higher levels. Among the three markers, *-le* had the highest frequency, followed by *–zhe* and *–guo.* Yang et al. also examined how the three markers encoded lexical aspect. It was found that most of the errors in the learners' use of *–le* occurred in situations where the marker was used with stative verbs. The imperfective *–zhe* was mostly used with statives and activity verbs, and *–zhe* was never used with achievement verbs. The researchers did not provide enough information about the use of

the experiential marker *–guo*. In discussing the results, the researchers mentioned the overuse of *–le* and *–zhe* in the learners' written production.

In a second study, Yang et al. (2000) investigated the underuse of aspect markers by L2 Chinese learners. The study included two types of data. One is based on the performance of 26 L1 Korean and L1 Japanese learners of Chinese at a Chinese university on a cloze test; the learners were divided into four levels. 120 narrative writings from the learners served as the basis for the naturalistic production data. According to the cloze test data, there was no difference between the four levels in terms of the frequency of *-le*, but the accuracy rate steadily improved with the increase of proficiency. The same pattern was found for the use of the durative marker *–zhe* and the experiential marker *–guo*. The correct rate of the use of *–guo* was higher than that of other markers across all the four levels, but its frequency was the lowest among the three. Regarding error types, the researchers found that overuse of the aspect markers decreased from lower to higher levels but underuse of the markers did not decrease as much as overuse. At higher levels, underuse occurred more frequently than overuse. The experimental data also showed that among the three markers, *-le* was underused the most, followed by *–zhe* and *–guo*. Compared with the cloze test data, the naturalistic data showed higher accuracy rate of the students' use of the three aspect markers. And overall, the experimental data was characterized by underuse whereas the naturalistic data by overuse. The differential results from the two types of data indicated the effect of task differences in the investigation of aspect acquisition.

Both Duff and Li's (2002) and Christensen's (1997) studies examined the use of *–le* by nonnative speakers of Chinese as compared with that by native speakers. In Duff and

Li's study, 9 native speakers and 9 nonnative speakers of Chinese were asked to complete three tasks: "an oral retelling of the Pear Story shown on the video (Chafe, 1980), a personal vacation narrative of vacation travel, and a written editing task of a past narrative that contained no aspect marking on verbs" (p.428). It was found that nonnative speakers tended to undersupply *–le* in oral narratives but oversupply it in the written cloze task. Christensen (cited in Duff & Li, 2002) found that more advanced learners used more perfective *–le*s and resultative verb compounds than beginners.

Taken together, these studies have identified the following facts about L2 learners' use of *–le*: (1) There is no difference between high- and low-proficiency learners in the accurate use of this aspect marker; (2) overall, learners tended to overuse *–le* in written tasks but underuse it in oral narratives; (3) as a perfective marker, *-le* is acquired earlier than progressive markers; (4) failure to correctly use *–le* resulted from learners' ignorance of the compatibility of *–le* with bounded situations; (5) cross-linguistic influence existed in the acquisition of *–le*: Native speakers of English tended to use *–le* as a past tense marker.

Chinese classifiers. Chinese is a classifier language. A classifier is a word that is used between a determiner (that is typically a number but can also be a demonstrative or quantifier) and a noun. The classifier is one of the most striking features of the Chinese language (Li & Thompson, 1981). The Chinese people started to use classifiers as early as 1400 B.C. (Erbaugh, 1986), and there are over 900 classifiers in the language (Zhang, 2007). The use of a classifier is both semantically and syntactically driven, and the choice of classifier depends on the noun. Semantically, a classifier is used to categorize and quantify a set of objects with the same or similar physical properties or characteristics.

The semantic representation of classifiers reflects how human beings perceive the world (Craig, 1986). Erbaugh (1986) divided Chinese classifiers into shape classifiers and function classifiers. There are two possibilities regarding the semantic motivation for the use of classifiers: (a) The construction has is fully predictable from the context, and (b) the construction is not fully predictable. For instance, the classifier *zhāng* is normally used with flat, smooth, and thin objects, hence *yī zhāng bǐng* (*one zhāng-CL pancake*, a pancake), or *yī zhāng zhuōzi* (*one zhāng-CL table* (because it has a thin, flat top), a table). In this case, the use of *zhāng* is predictable from the context since the referent following the classifier has all the perceptual features of the category of objects it refers to. However, in Chinese, the word "sofa" also takes the classifier *zhāng* although by no means a sofa is flat or thin (in which case the classifier is not predictable from the context). This might be due to the fact that when the furniture sofa (both the object and the word denoting it) was imported from the West, due to the absence of an appropriate classifier for it, the classifier that is used for table (which is also furniture), *zhāng*, was employed for this alien object. All classifiers, as Ahrens (1994) pointed out, have semantic connections with the physical properties or functions of the objects they refer to although the use of some appears to be arbitrary as a result of the evolution of the language.

Syntactically, "classifiers are units of enumeration employed to mark countability; their occurrence makes the semantic partitioning of nouns visible" (Wu & Bodomo, 2009, p.490). A nominal classifier is a bound morpheme that must occur with a determiner or quantifier. There are three permutations with respect to classifier use in the Chinese

language and in each permutation the use of a noun is optional if the referent is inferable from the discourse context (Li & Thompson, 1984):

(1) Number + Classifier + Noun
e.g.
yī gè rén.
一 个 人。
One-*CL* person.
One person.

(2) Demonstrative + Classifier + Noun
e.g.
zhè pǐ mǎ.
这 匹 马。
This-*CL* horse.
This horse.

(3) Quantifier + Classifier + Noun
e.g.
měi liàng chē
每 辆 车。
Every-*CL* vehicle.
Every vehicle.

Traditionally, no distinction is made between measure words and classifiers (Chao, 1968; Li & Thompson, 1984). At times, measure words are referred to as measure/count classifiers and classifiers are called special /mass classifiers (Chien, Lust, & Chiang, 2003; Erbaugh, 1986). However, this is not reasonable. A measure word often accompanies non-count nouns whose referents are not quantifiable as in *a glass of water*. When used with count nouns, the function of a measure word is to quantify, such as *a basket of pears*. A classifier is always used with count nouns to categorize as in *liǎng kē shù* (two-CL trees, meaning "two trees"). Also, while there is no semantic connection between a measure word and the accompanying noun, such connection exists between a classifier and the noun it co-occurs with. Measure words have equivalents in English but classifiers do not. It is the presence or absence of classifiers, not that of measure words,

that distinguishes classifier languages from non-classifier languages—classifiers are language specific but measure words are language universals (Erbaugh, 1986; Li, 2000; Tai & Wang, 1990).

The use of classifiers also relates to discourse factors. For example, Erbaugh (1986) explored adult classifier use based on the narratives of 19 native Mandarin speakers in Taiwan about a loosely-plotted seven-minute color film with sound but no dialogue (The Pear Film) (Chafe, 1980). She also examined classifier use in 877 utterances in casual conversations. It was found that the use of the general classifier *gè* dominated the subjects' classifier use. When special classifiers were used, they were used "to specify 1) the first mention of a 2) non-present object which was 3) unfamiliar or unclear to the hearer, especially in reference to a 4) new creation or as part of a 5) narrative, 6) pretend play scheme, or a 7) request" (p.425). Li (2000) also approached classifier use from a discourse perspective, arguing that classifiers served as a grounding mechanism to mark the salience of the related noun phrases.

How Chinese classifiers are used and acquired by second language speakers has been insufficiently studied. One representative study on L2 learners' use of Chinese classifiers is by Polio (1994). Her study involved 21 English and 21 Japanese speakers studying Chinese in Taiwan. They were students from three different levels of proficiency as measured through class placement, native speaker ratings, and an elicited imitation test. As in Erbaugh's study about L1 speakers' use of classifiers, the data in Polio's study were also collected using the Pear Film narratives. It was found that the nonnative speakers rarely omitted a classifier in obligatory contexts. However, when omission errors happened, they were invariably committed by English speakers (eight out

of 21 English speakers did), and the Japanese speakers never made such errors. This seemingly insignificant finding was not given further explanation but it is without doubt worth attending to in light of the fact that English is a non-classifier language whereas Japanese is a classifier language, although in Japanese a classifier is used after a noun but in Chinese it precedes a noun.

Chinese perfective –le versus Chinese classifiers. The choice of target structure is an independent variable in this study because it is speculated that different feedback types may have differential effects on the learning of different linguistic structures. As previously mentioned, there has been empirical evidence that feedback worked differently for different structures across studies (Ammar & Spada, 2006; Lyster, 2004; Sheen, 2007), but thus far, Ellis (2007) and Yang and Lyster (2010) are the only studies that have examined the choice of structure as an independent variable. Ellis's study included the English past tense *–ed* and comparative *–er* as the target structures, but the effects of the two feedback types (recasts and metalinguistic clues) did not differ substantially. Yang and Lyster's study investigated the differential effects of recasts and prompts on the learning of English regular and irregular past forms in a classroom setting. This study investigates whether implicit and explicit feedback, operationalized as recasts and metalinguistic correction, facilitates the interlanguage development of two very different Chinese structures in a lab setting.

Previous sections provided extensive, separate discussions on the two included target structures of this study. This section seeks to juxtapose the two structures and offer a comparison between them along various dimensions. DeKeyser (2005) pointed out that difficulty in grammar learning may relate to at least three factors: form, meaning, and

form-meaning mapping (see Table 2 for different schemes on structural difficulty). Difficulty of form may result from the competition between available choices when the learner is selecting the right morpheme and allomorph. The meaning of a form may constitute a source of difficulty because of its novelty or abstractness or both. A major source of difficulty is form-meaning mapping, that is, the link between form and meaning is not transparent. There are three contributing factors to the difficulty related to form-meaning mapping: 1) redundancy—the form is not semantically necessary, 2) optionality — the supply of the form is not obligatory, and 3) opacity — a morpheme has different allomorphs and the same form stands for different meanings. DeKeyser went on to state that if form-meaning mapping is clear, minimal exposure may be enough for acquisition; if it is obscure, the structure may pose a great challenge for adult learners. Goldschneider and DeKeyser (2005) explored the factors contributing to the sequence of L2 English morpheme acquisition through a meta-analysis. A large portion of variance in acquisition order was accounted for by five determinants: perceptual salience, semantic complexity, morphophonological regularity, syntactic category, and semantic complexity. The authors pointed out that all these factors related to saliency to varying degrees.

Ellis (2007) developed a set of criteria to determine the difficulty of a linguistic form. These criteria include (a) grammatical domain—whether a form is morphological or syntactic, (b) input frequency, (c) learnability/processibility (Pinneman, 1998)—linguistic forms are processable at different stages of development, (e) explicit knowledge—the complexity of the rule explanation of a structure, (f) scope (Hustijn & De Graaff, 1994)—the scope of a rule is large if it covers more than 50 cases, (g) reliability—a rule has high reliability if it applies to more than 90% of all cases, and (h) formal semantic

redundancy—forms that are not necessary for meaning processing are semantically redundant.

In light of the difficulty of approaching structural difficulty theoretically, some researchers based their judgment on the ratings by language instructors. For instance, Robinson (1997) asked 15 ESL teachers to rate the complexity of some selected rules and identified an easy rule and a hard rule for instructional treatment. Drawing on the schemes developed by previous researchers and given the characteristics of the two target structures of this study, a comparison is made between them based on the following criteria:

(1) Redundancy. Redundancy refers to the fact that the use of a linguistic structure is semantically superfluous although it might be syntactically obligatory. Whether a certain structure is redundant is dependent upon whether it is indispensible to the accurate interpretation of the utterance that contains the structure. Alternatively, one can determine the redundancy of a linguistic feature by examining whether the meaning the feature encodes can also be encoded through other linguistic features or devices in the utterance. A classifier is critical to the accurate interpretation of the determiner phrase where the classifier is situated. Missing a classifier or using a wrong classifier is likely to distort the meaning of the utterance or make the utterance unintelligible. For instance, the utterance **sān hé* (three rivers), where the classifier *tiáo* is missing, is hardly intelligible to a Mandarin speaker. And **sān zuò hé*, where the classifier for "bridge" is used for rivers, sounds equally foreign.

Unlike the classifier, the perfective *–le* is redundant in many cases. This is because the Chinese language is heavily discourse- or topic-oriented, and semantic

interpretation is largely dependent upon the context rather than the syntax of the utterance (Yang, 1995; Duff & Li, 2002). The absence of the perfective *–le* can be compensated for by the use of time expressions, the sequence of events, and other discourse or linguistic devices. For example, in the sentence ** zuó tiān wǒ chī le sān gè píng guǒ* (Yesterday I ate three apples), *-le* is used with the activity verb *chī* (eat) to encode perfectivity. However, the time expression and the number can jointly mark perfectivity in the absence of *-le*. Furthermore, linguists noticed that *–le* is more often used with monosyllabic words than disyllabic words to meet the disyllabic feature of the Chinese language (Chang, 1986; Yang, Huang, & Cao, 2000). The use of *–le* in these cases is obviously phonologically rather than syntactically or semantically driven. In conclusion, classifiers are more meaning-loaded than the perfective *–le* as incorrect classifier use is a source of communication breakdown but absence of the perfective -le may not impede information exchange.

(2) Perceptual saliency. According to Goldschneider and DeKeyser (2005), perceptual saliency refers to the ease in hearing or perceiving a given structure. The perfective *–le* is always affixed to the verb in the post-verbal position and is always pronounced in a neutral tone (Li & Shirai, 2000). A Chinese classifier always precedes the noun and its tone is not neutralized. Therefore, classifiers would seem to be more perceptually salient than the perfective *–le*.

(3) Form-meaning mapping. Form-meaning mapping can be transparent or opaque. In the case of the perfective *–le*, the form-meaning mapping is opaque because of the fact the form has two variants that have different interpretations. The verbal *–le* encodes completion and boundedness, and the sentence-final *–le* indicates current

relevance/change of situation. So this same form may occur in different positions of a sentence and stands for different meanings. The form-meaning mapping of classifiers is transparent in that a certain classifier is usually used with one object or objects that fall into the same category because of the physical properties they have in common.

(4) Explicit knowledge. The rule explanation for the use of the perfective *–le* is complex because it involves at least two components: (a) the event is completed, and (b) the situation must be bounded or have an endpoint. Thus, the rule is difficult to understand and learn as explicit knowledge. The rule of classifiers, conversely, is relatively easy: It only states that a certain classifier must be used with a particular noun.

(5) Learnability /teachability. Using Pienemann's Processiblity Theory (1998), Zhang (2005) investigated the emergence of five Chinese structures in the speech production of three L2 Chinese learners, who were enrolled in a first-year Chinese course at an Australian university. Among the five structures were two aspect markers (progressive and experiential) and classifiers. Zhang argued that according to the processibility theory, aspect markers are lexical morphemes and require the Category Procedure (Stage 2; Stage 1 is called Word/Lemma) to implement. The processing of the classifier, however, involved the numeral, the classifier proper, and the head noun. Thus, it was processed through the Phrasal Procedure (Stage 3). However, contrary to the prediction of the Processibility Theory, classifiers emerged earlier than the aspect markers. Zhang failed to find a reasonable explanation for the finding. Both Wen (1997) and Yang et al. (2000) showed that there was no difference between high- and low-proficiency learners in terms of their accuracy in using *–le*. This was ascribed to

the difficulty in acquiring the structure. With regard to classifiers, Li (2009) found that fourth-year Chinese learners did not differ from second-year learners in the use of classifiers in their pretest scores. Li conjectured that this might be due to the reduced occurrence of classifiers in the textbook for the advanced learners. Taken together, these studies showed 1) both structures emerged very early in learners' interlanguage, 2) there seemed to be no difference between beginners and advanced learners in their accurate use of the two structures, but this counter intuitive finding was likely caused by different factors: -*le* was difficult and classifiers were less frequent in textbooks for learners at the higher level.

Table 2. Schemes on structural difficulty

DeKeyser '05	***Goldschneider & DeKeyser '05***	***Ellis '07***	***Robinson '97***
form: choice between morphemes and allomorphs	*morphonological regularity:* extent to which a form is affected by its phonological environment	*grammatical domain:* whether a form is morphological or syntactic	*complexity of the structure* described by pedagogical rules
meaning: novelty/abstractness	*frequency:* number of times a form occurs	*input frequency:* how frequently a form occurs in the input	*complexity of pedagogical rules* describing the structure
form-meaning mapping: a. redundancy b. optionality c. opacity	*semantic complexity:* number of meanings a form expresses	*learnability:* extent to which a form is processable	*expert opinion* from instructors
	perceptual salience: ease in hearing or perceiving a given structure	*explicit knowledge:* complexity of rule explanation	
	syntactic category: whether a form is lexical or syntactic	*scope:* number of cases a rule covers	
		reliability: percentage of all cases a rule applies to	
		formal semantic redundancy: indispensability in expressing meaning	

Language Aptitude

As previously mentioned, the effects of corrective feedback have to do with learner-internal as well as learner-external factors. Whereas learner-external factors include the characteristics of feedback, the target structure, and the instructional context, learner-internal factors relate to age, proficiency, noticing (how learners perceive feedback or whether learners notice feedback), and individual differences in aptitude, anxiety, motivation, attitude toward feedback, and the like. As Ellis and Sheen (2006) noted, among the various factors affecting the effects of feedback, there had been very little research on the moderating effects of individual difference variables. Among the various individual difference variables, language aptitude is worthy of special attention.

According to Robinson (2005), "second language (L2) aptitude is characterized as strengths individual learners have—relative to their population—in the cognitive abilities information processing draws on during L2 learning and performance in various contexts and at different stages" (p.46). Carroll and Sapon claimed that learners' language aptitude is stable and not subject to training or environmental factors; it is "largely independent of intelligence, and is distinct from motivations and attitudes of the learner" (2002, p.24). Also, as Sawyer and Ranta (2001) observed, aptitude is not susceptible to learners' previous language learning experience and is not "a matter of skill development" (p. 334). Aptitude has received much attention in SLA because it is the most predictive of L2 proficiency among individual difference variables (Ehrman & Oxford, 1995; Hummel, 2009; Reves, 1983; Robinson, 2005; Skehan, 1998; Sparks, Patton, Canschow & Humbach, 2009). Studies using the Modern Language Aptitude Test (MLAT) developed

by Carroll and Sapon (1959) showed that the correlation between aptitude and L2 success ranged from .4 to .6 (Robinson, 2005).

The MLAT (Carroll & Sapon, 1959, 2002) has been used as a standard measure of second language aptitude. It consists of five parts that measure four dimensions of aptitude including learners' phonetic coding ability, grammatical sensitivity, rote learning ability, and inductive language learning ability. It should be noted that the distinction between grammatical sensitivity and inductive language learning ability is fuzzy, and that the latter in fact is not measured in the MLAT (Carroll, 1962; Erlam, 2005; Sawyer & Ranta, 2001). Skehan (1998) contended that these two abilities have the same underlying construct: language analytic ability. Hence, in this study, the term "language analytic ability" is adopted for grammatical sensitivity. Other test batteries have been developed such as the PLAB (Pimsleur, 1966) and the DLAB (Petersen & Al-Haik, 1976), but the MLAT has so far proved to be the best instrument to measure language aptitude (Sawyer & Ranta, 2001; Dörnyei, 2005). Carroll and Sapon (2002) stated that the MLAT can be use to select students for foreign language courses, estimate individual students' probability of L2 learning success so that counselors can provide appropriate guidance, achieve placement purposes, and diagnose students' learning abilities so as to match learner types with instructional approaches.

Initially, the primary purpose of L2 aptitude research was to examine the extent to which aptitude could differentiate learners in terms of the rate at which to achieve L2 gains. However, aptitude research waned in the 1970s because of two reasons (Robinson, 2002): (a) By emphasizing the role of aptitude, learners' individual efforts are diminished, and (b) some researchers (Cook, 1996; Gardner, 1985; Spolsky, 1989) claimed that with

the advent of communicative language teaching, the predictive power of aptitude tests, which were developed in audio-lingual instructional contexts, no longer obtained. Despite a temporary slowdown, aptitude research has resurrected in recent years. On one hand, researchers found that aptitude predicted L2 success in all sorts of learning environments including communicative language classes (Ehrman & Oxford, 1995; Ranta, 2002), meaning-based immersion classes (Harley & Hart, 1997), informal learning settings (Reves, 1983), and the laboratory (Robinson, 1997)[2]. On the other hand, acknowledging the limitation of using aptitude measures only for prediction or selection purposes, researchers embarked on exploring new venues of investigation.

Aptitude-Treatment Interaction

Snow (1987, 1991; Cronbach & Snow, 1977) argued that aptitude should not be only considered from the learner's perspective but also from the perspective of how it interacted with situational constraints. Building on Snow's concept of aptitude-treatment interaction, researchers (Robinson, 2005; Segalowitz, 1997) pointed out that aptitude should not be viewed as a fixed characteristic because the learner is situated in a complex, dynamic environment that imposes different cognitive demands on learners with different experiences and/or at different stages of learning. Also, instead of a monolithic construct, aptitude is composed of multiple components, and these (sets of) components interact differently with different learning conditions (Robinson, 1997, 2002) and are drawn upon at different stages of learning (Dörnyei & Skehan, 2003; Skehan, 2002).

Empirical research has shown support for the above claims. For instance, research showed that the role aptitude and aptitude components played varied depending on the stage the learner was at. DeKeyser (2000) and Sasaki (1996) found that aptitude had little

to do with pre-critical period language learning and that it only affected post-critical period learners. Harley and Hart (1997) found that pre-critical period learning related with memory and post-critical learning had to do with analytic ability. In terms of aptitude-treatment interaction, Wesche (1981) found that learners benefited more from the instruction that matched their cognitive strengths than from the instruction that did not. For instance, students with high analytic ability benefited more from the analytical teaching approach than from other approaches.

Robinson's work (1997, 2002) provided further evidence for the importance of examining the interaction between aptitude and different learning conditions. In the 1997 study, Robinson investigated whether aptitude related to awareness and learner gains in four conditions: (a) implicit (memorizing examples only), (b) incidental (processing examples for meaning), (c) rule-search (trying to find rules), and (d) instructed (applying a rule explanation to examples). 104 intermediate ESL learners participated in the study. Aptitude measures included the Paired-Associates Subtest (for memory) and the Words in Sentences Subtest (for grammatical sensitivity) of the MLAT. The combined scores of the two subtests were used as the global score of aptitude. It was found that in the implicit condition, learners' posttest scores and awareness correlated with grammatical sensitivity; in the instructed condition, memory was positively related to awareness; in the rule-search condition, grammatical sensitivity was positively related to awareness. The only condition that was unaffected by aptitude in terms of learning or awareness is the incidental condition. Robinson speculated that it was because the incidental condition did not draw on the learner's rote memorization ability. To address this issue, he included a working memory test in the 2002 study and found significant correlations between

working memory and learning in the incidental condition. Robinson concluded that (a) adult learning in all conditions is similar and is sensitive to individual differences, (b) the extent to which learning is affected by individual differences is determined by whether the processing demands of the learning condition match the cognitive ability under question, and (c) different learning conditions draw on different aptitude complexes or different sets/combinations of aptitude components.

In sum, to move aptitude research forward, researchers (Robinson, 2005; Skehan, 2002) have called for more research into the interaction between language aptitude and learning conditions. To echo this call from aptitude researchers, scholars in feedback research (e.g., Ellis & Sheen, 2006) called for more research into the role of individual difference variables such as language aptitude in affecting the efficacy of corrective feedback. The need for more research into individual differences is evident in Ellis's statement that "[t]he vast bulk of CF studies has ignored learner factors, focusing instead on the relationship and the effect of specific CF strategies and learning outcomes" (2010, p. 339). This study addresses the relevant issues by investigating the relationship between corrective feedback and two aptitude components: language analytic ability and working memory.

Language Analytic Ability and Feedback

Language analytic ability is often measured with the Words in Sentences subtest of the MLAT. Carroll defined language analytic ability (grammatical sensitivity) as "the ability to recognize the grammatical functions of words (or other linguistic entities) in sentence structures" (1981, p.105) or "the individual's ability to demonstrate his awareness of the syntactical patterning of sentences in a language and of the grammatical

functions of individual elements in a sentence" (1973, p.7). Previous research showed that among the four components included in the MLAT, language analytic ability is probably the most predictive of L2 proficiency (Ehrman & Oxford, 1995; Hummel, 2009; Ranta, 2002).

Research has demonstrated that language analytic ability interacted with contextual factors and affected learners at different acquisition stages. Reves (1983) found that language analytic ability played a greater role in formal classroom contexts than in naturalistic learning settings. Robinson (1997) found that language analytic ability did not affect the incidental learning condition but it related to learning in both the implicit and explicit conditions. Erlam (2005) studied the relationship between language analytic ability and three learning conditions: deductive instruction, inductive instruction, and structured input instruction. Language analytic ability was found to be positively correlated with learning in the inductive condition and the structured input condition, but it was not related to learning in the deductive condition. Erlam also found that the correlations between language analytic ability and learning conditions were subject to test format. Finally, there is also evidence that language analytic ability did not affect child learners but it related to adult L2 learning (DeKeyser, 2000; Harley & Hart, 2002; Rose, Sasaki, & Yoshinaga, 2002).

In terms of how language analytic ability interacts with the effectiveness of corrective feedback, the picture is far from clear because there have been only a few relevant studies (DeKeyser, 1993; Sheen, 2007a, 2007b; Trofimovich, Ammar, & Gatbonton, 2007, which will be reviewed below in the "Working memory and feedback" section). DeKeyser's longitudinal study investigated the relationship between two

feedback types (implicit and explicit) and three individual difference variables: language analytic ability, extrinsic motivation, and anxiety. Participants were two classes of Dutch-speaking high school learners of French (n = 19; n = 16). During a full school year, the instructor of one class was told to correct mistakes as frequently and explicitly as possible, and the instructor of the other class was directed to avoid error correction. Posttest results showed that the class which received feedback did not outperform the no-feedback class and that language analytic ability did not correlate with the effectiveness of feedback, despite the fact that feedback correlated with the other two individual difference variables. The absence of a link between feedback and language analytic ability was ascribed to the strong effect of anxiety, which might have neutralized the role of aptitude.

Sheen (2007a; 2007b) conducted two similar studies to explore how the effects of feedback were mediated by learners' language analytic ability. In one study (2007b), she investigated the extent to which ESL learners benefited from two types of written feedback: direct-only correction (provision of correct form) and direct metalinguistic correction (correction + metalinguistic explanation), in the learning of two uses of English indefinite and definite articles: *a* as first mention and *the* as anaphoric reference. The study involved 92 students from 6 classes and these students formed two experiment groups and one control group. There were two treatment sessions, during which the students read a story, the instructor discussed the moral of the study, and finally the students rewrote the story. Students' writings were turned in to the researcher, who provided different types of feedback or no feedback to the mistakes the students made in using the target structures. 2-4 days later, the students attended a feedback session where they went over the comments provided on their writings. Three tests were used to

measure the effects of feedback: speeded dictation, narrative writing, and error correction. Language analytic ability was measured with a test used by Schmitt et al. (2003) that consisted of 14 multiple choice questions asking the students to choose the correct translation for a sentence in an artificial language after they were familiarized with a list of exemplars in the artificial language and the English equivalents. The results indicated that language analytic ability correlated with the gains of both feedback groups, but that the correlations were stronger for the delayed effects for feedback.

In the other study (2007a), Sheen replicated the afore-reviewed study in the oral mode, that is, participants received oral rather than written feedback on their mistakes in using English articles. The two feedback types provided in this study were recasts and metalinguistic correction. The results obtained regarding the relationship between feedback and language analytic ability were somewhat different from those in the other study. Whereas language analytic ability related to the effects of both feedback types in that study, it only correlated with the effects of metalinguistic feedback in this study, indicating different results for oral and written feedback. What merits attention is that a negative, albeit insignificant, correlation was found between the effects of recasts and language analytic ability.

As shown, there is a very limited amount of research on how the effects of feedback are constrained by the learner's language analytic ability, a major component of language aptitude. The few previous studies were carried out in the classroom, which might not be an ideal setting to investigate individual difference variables. Also, previous research only addressed certain aspects of the relation between feedback and language analytic ability, and many questions remained to be answered. DeKeyser's study (1993) examined

two broad categories of feedback (implicit and explicit); how specific feedback types interact with aptitude is not clear. Sheen investigated how recasts and metalinguistic correction, provided as written and oral feedback in two respective studies (2007a, 2007b), related to language analytic ability. The results from the two studies were different. Also, the target structure in both studies was English definite and indefinite articles, a non-salient linguistic feature. One question that needs further exploration is whether aptitude interacts with different feedback types in the learning of different linguistic structures. This study seeks to answer the question.

Working Memory and Feedback

The term "working memory" has been adopted for short-term memory to reflect the fact that instead of being merely a warehouse to store incoming data, it is also responsible for information processing. Miyake and Friedman (1998) rightly pointed out the difference between working memory and the traditional conception of short-term memory:

> "Unlike the traditional conception of short-term memory (STM) as a fixed set of slots that passively store to-be-maintained information…the conception of WM is more closely tied to the dynamic nature of the processing and storage activities, such as executing various language processes and maintaining intermediate products of the processing". (p.341)

There are two views on the architecture of the working memory construct (Conway, Jarrold, Kane, Miyake, & Towse, 2007; French, 2006): the unitary approach and the multicomponential or multifaceted approach. Researchers embracing the unitary approach believe that working memory is a single construct that performs both storage

and processing functions (Daneman, 1991; Daneman & Carpenter, 1980). There are others who hold that working memory consists of a central executive and several slave systems (Baddeley, 2003, 2006, 2007; Baddeley & Hitch, 1974). The central executive is responsible for the control and regulation of the working memory system, and the subcomponents include the phonological loop that stores phonological/auditory information, the visuospatial sketchpad that involves the generation and storage of visual information, and the episodic buffer that integrates information from a variety of systems and from long-term memory.

Working memory is operationalized in two ways and is measured accordingly. One way is to define it as phonological working memory, which is measured through digit span or nonword repetition tests where the learner is asked to repeat a sequence of digits, words, or nonsense syllables (e.g., Baddeley et al., 1998). However, some researchers argued that digit span or nonword repetition tests only measure the storage function of working memory and a good working memory test should also measure the processing function (Daneman & Carpenter, 1980; Walters & Caplan, 1996). In their seminal study, Daneman and Carpenter (1980) developed a reading span test that taps into both the storage and processing components. The test has been used in numerous studies as a standard measure of working memory. During the test, subjects were required to read sets of unrelated sentences and recall the sentence-final words in each set. The researchers also developed a listening span test as a variant of the reading span test, where participants were asked to listen to some sentence stimuli read by the presenter and recall the final words of the sentences. The rationale behind the reading/listening test is that participants had to process the meaning of a sentence when reading or listening to it and

at the same time memorize the final word of the sentence. Daneman and Carpenter found that the reading and listening span test scores correlated with college students' reading comprehension ability and verbal SAT scores, but traditional word span measures did not. The finding that complex span tests that tax both processing and storage are better predictors of L1 and L2 learning was also obtained by other researchers (Harrington & Sawyer, 1992; Lehto, 1996; Miyake & Friedman, 1998; Waters & Caplan, 1996).

Although sentence span tests have proven to be one step forward compared with traditional word- or digit-span tests, they are not unproblematic. The problem lies in the way the tests are scored. Despite the claim that sentence span tests measure both the processing and storage functions of working memory, it is usually only the recall component that is scored. Researchers argued that learners might trade off between processing and recall accuracy, that is, they might sacrifice the speed and accuracy of processing to achieve better recall scores (Waters & Caplan, 1996; Leeser, 2007). To verify this hypothesis, Waters and Caplan administered a test during which subjects were asked to view some sets of sentence stimuli, judge whether each sentence made sense in the real world, and recall the sentence-final words in each set. Scores for reaction time, plausibility judgment, and recall accuracy were all calculated, and negative correlations were found between the three scores, showing that the subjects did trade off between the three components. It was also found that the global score of the three components was a better predictor than the score for recall accuracy alone.

Working memory has been found to correlate with both L1 and L2 learning whether it is measured as phonological short-term memory using word/digit span tests or as a construct that is responsible for processing and storage activities and that is measured

with sentence span tests. In L1 research, it is found that phonological short-term memory is a strong predictor of vocabulary acquisition (Avons, Wragg, Cupples, & Lovegrove, 1998; Gathercole, Frankish, Pickering, & Peaker, 1999; Michas & Henry, 1994), and that learners' performance on sentence span tests is a strong predictor of reading comprehension ability (Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Just & Carpenter, 1992; Waters & Caplan, 1996). In L2 research, phonological short-term memory is found to be associated with vocabulary learning (Papagno, Valentine, & Baddeley, 1991; Service & Kohonen, 1995) and grammar learning (N. Ellis & Sinclair, 1996; Williams & Lovatt, 1999; Hummel, 2009). Working memory measured with sentence span tests are shown to predict reading comprehension ability (Harrington, 1991; Harrington & Sawyer, 1992; Leeser, 2007), listening comprehension ability (Miyake & Friedman, 1998), acquisition of morphosyntax (Mackey et al., 2002; Sagarra, 2007; Trofimovich, Ammar, & Gatbonton, 2007), and modified output in L2 interaction (Mackey, Adams, Stafford, & Winke, 2010).

In L2 research, there has been a call to investigate working memory as an aptitude component (Miyake & Friedman, 1998; Robinson, 2005; Skehan, 1982). Robinson argued that aptitude as measured by the MLAT and other test batteries were developed in audiolingual teaching where rote-learning was a major feature. However, in communicative language teaching, linguistic forms are addressed in meaning-focused instruction, and the processing demands of this type of instruction are different from those of audiolingual classes. Thus, "for these learning conditions, a measure of aptitude that reflects the processing demands of simultaneous attention to form and meaning, with its attendant demands on *working memory* [emphasis added] would seem to be

necessary" (p.215). Skehan also argued that the MLAT subtest that is concerned with the memory component of aptitude measures learners' associative memory, which may not be most predictive of language learning. Furthermore, there is empirical evidence to justify working memory as an aptitude component. For instance, Robinson (1997) examined the correlation between aptitude and four learning conditions and found that aptitude did not relate to the treatment effects in the incidental condition. However, in a later study (2002), a working memory test was used and aptitude was found to correlate with the incidental condition.

With respect to the relationship between working memory and the effectiveness of corrective feedback, there have been three published studies (Mackey et al., 2002; Sagarra, 2007; Trofimovich, Ammar, & Gatbonton, 2007). Mackey et al. investigated the relationship between working memory, noticing of recasts, and the effects of recasts in the learning of English question formation. The participants were 30 Japanese EFL learners. The learners' working memory capacities were based on their scores on three measures: a non-word recall test, an L1 listening span test, and an L2 listening span test. The noticing data were based on the learners' metalinguistic comments during a stimulated recall and the learners' responses to an exit questionnaire. Learning of the target structure was determined by way of the production of targetlike higher-stage questions after the learners received recasts on their nontargetlike production of English questions while engaged in communicative tasks. Results showed that more noticing was reported by learners with higher working memory capacities (the result was only obtained for the composite working memory score) and by learners at lower developmental level of the target structure (the result was obtained only for the non-word recall working

memory subtest). In terms of the contribution of working memory to learner outcome, learners with lower working memory capacities showed more improvement at the immediate posttest; learners with higher working capacities demonstrated more interlanguage development at the delayed posttest. Mackey et al.'s study is important in that it is the first attempt to address the relationship between working memory and the effects of corrective feedback in SLA research. However, due to the small sample size, the authors cautioned against the generalizability of their findings.

Trofimovich, Ammar, & Gatbonton (2007) investigated the role of attention, memory, and analytic ability in affecting the effects of computerized recasts. During the study, 32 adult Francophone learners of English were presented with some pictures on a one-on-one basis, the description of which required the use of the target structures. The learner's description of each picture was followed by a recorded native speaker response that served as a recast. Two memory measures were used: one was a non-word repetition test measuring phonological short-term memory, and the other, called a working memory test, was the Letter-Number Sequencing subtest of the Wechsler Adult Intelligence test (Psychological Corporation, 1997). The Words in Sentences subtest of the MLAT was used to measure analytic ability, and attention control was tested using the Trail Making Test of the US Army Individual Test Battery. It was found that recasts were effective and that learners' individual differences in attention control, analytic ability, and phonological working memory were predictive of the learners' interlanguage development; working memory was not a significant predictor.

Similar to Trofimovich, Ammar, & Gatbonton (2007), Sagarra (2007) examined the effects of recasts that were provided via the computer and the effect of working memory

on the effectiveness of recasts. 82 L1 English speakers enrolled in first-semester Spanish classes at a U.S. university participated in the study. They were asked to fill in the blanks in some Spanish sentences using the correct forms of the given adjectives. A recorded recast was provided when an error was made. The effects of recasts were tested with a written test as well as an oral production test. The working memory test was adapted from Waters and Caplan's sentence span tests (1996), and scores were computed only based on the items where the learner was accurate in plausibility judgment, the reaction time was not an outlier, and the final word of the sentence was correctly recalled. The results revealed that recasts were effective and the effects were associated with the learners' working memory capacities.

Previous research has established a link between corrective feedback in the form of recasts and working memory. However, further research is warranted to address remaining issues. Mackey et al.'s study revealed some interesting and thought-provoking findings, but these findings need to be verified and tested with more learners and in different contexts. Trofimovich, Ammar, and Gatbonton (2007) and Sagarra (2007) obtained some valuable results, but in both studies, recasts were provided in the computer mode and in discrete item practice. How working memory interacts with feedback in meaningful communication remains to be seen. All three studies investigated recasts, so how working memory relates to the effects of other feedback types needs further exploration. Also, in previous research, working memory was either operationalized as phonological short-term memory, or when it was measured using complex, sentence-span tests, a score that included all three components of the measure (reaction time, plausibility judgment, and word recall) was not used to reflect both the processing and

storage functions of working memory. Finally, it is speculated that the learning of different linguistic structures might impose different processing demands on the learner's working memory. To date, no study has examined the interaction between the choice of target structure and working memory in feedback research. This study was undertaken to address this gap by including two very different structures: Chinese perfective *–le* and Chinese classifiers.

2.5 Research Questions

The review of the literature shows that the facilitative role of corrective feedback is theoretically justified (Gass, 1997, 2003, 2004; Long, 1996, 2007) and empirically verified (Li, 2010; Lyster & Saito, 2010; Mackey & Goo, 2007; Norris & Ortega, 2000; Russell & Spada, 2006); it is also abundant in second language classes (Lyster & Ranta, 1997; Loewen, 2004; Sheen, 2006). Now that the effects of feedback have been established, the question arises as to what factors, be they learner-external or learner-internal, mediate the effects. The identification of the constraining factors of the effects of feedback is equally, if not more, important than the establishment of its effects. This study investigates how the effectiveness of implicit and explicit feedback is affected by the learner's proficiency, working memory capacity, and language analytic ability in the learning of two Chinese structures. The following research questions are formed:

RQ1: Do explicit feedback and implicit feedback facilitate the learning of Chinese perfective *–le*? If so, do they have differential effects on learners at different proficiency levels in the learning of the structure?

RQ 2: Do explicit feedback and implicit feedback facilitate the learning of Chinese

classifiers? If so, do they have differential effects on learners at different proficiency levels in the learning of the structure?

RQ 3: Do the two feedback types work differently in the learning of the two target structures?

RQ 4: What is the relationship between feedback type, the nature of linguistic structure, and learners' language analytic ability?

RQ 5: What is the relationship between feedback type, the nature of linguistic structure, and learners' working memory capacity?

CHAPTER 3 METHOD

The previous chapter laid out the theoretical framework and provided the rationale for the investigation of the variables included in this study. Previous studies on corrective feedback were discussed and issues were identified that need to be addressed in further research. This chapter details how the study was conducted with regard to the characteristics of the participants, tasks where the target structures were elicited and feedback was provided, the procedure, testing materials, coding schemes, and the statistical analyses that were performed.

Participants and Grouping

The participants of this study were 78 learners of Chinese from two large Midwestern U.S. universities. Among them, 75 were native speakers of English and 3 reported Korean as their native language[3]. Heritage speakers of Chinese were not included in the study and they were identified by being asked to respond to the question of whether their parents were Chinese and whether they spoke Chinese at home. The instructors of the classes that contributed participants were also consulted to verify the participants' linguistic background. At the time of data collection, the learners were in their 4^{th} ($n = 41$), 6^{th} ($n = 20$) and 8^{th} ($n = 17$) semesters of their Chinese study. 34 of the learners were female and 44 were male. With respect to the learners' enrollment status, 6 were freshmen, 20 were sophomores, 28 were juniors, 21 were seniors, and 3 were graduate students. They were aged between 18 and 38, and the average age was 20.78 (SD = 2.48). The learners volunteered to participate in the study and were provided monetary compensation and extra credit points in return for their time commitment.

A standardized Chinese proficiency test named HSK (see the "testing" section for

details about this test) was administered to each participant because proficiency is an independent variable in this study and a major goal of this study is to explore whether different types of feedback affect high-proficiency and low-proficiency learners differently. Using a proficiency test also made it possible to recruit students from two academic institutions. Based on their performance on the proficiency test, the learners were divided into two large groups: high and low. The full score of the test is 60 and the median of the learners' scores, 29, was set as the cut-off point for the high-low division: Learners who scored 29 or higher were labeled "high", and those who scored 28 or lower were labeled "low". An Independent-Samples t-test was performed and showed that the two resultant proficiency groups were significantly different in terms of their test scores, $t(76) = -11.65$, $p < .00$ (the statistics for grouping information appear in Table 3).

At each level (high and low), the learners were divided into three subgroups: implicit, explicit, and control, depending on the type of feedback they received. Consequently, six groups were generated, three at each proficiency level: low-implicit, low-explicit, low-control, high-implicit, high-explicit, and high-control. One-way ANOVAs were conducted to make sure that the three groups at each level were comparable in terms of proficiency. The analyses showed no significant difference between the three groups at the low level, $F(2, 36) = .71$, $p = .51$, or at the high level, $F(2, 36) = 0.36$, $p = .70$.

Table 3. Descriptive statistics for groups

Low proficiency									High proficiency								
n			*M*			*SD*			*n*			*M*			*SD*		
39			23.31			3.01			39			36.67			6.49		
LI			LE			LC			HI			HE			HC		
n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
14	24.07	3.29	15	22.80	3.00	10	23	2.67	14	35.64	5.83	14	36.71	5.70	11	37.91	6.51

Note. LI = low implicit; LE = low explicit; LC = low control; HI = high implicit; HE = high explicit; HC = high control

Feedback Operationalization

Implicit Feedback

Implicit feedback was operationalized as recasts, that is, the reformulation of the learner's nontargetlike production of the target structures (Chinese classifiers and the Chinese perfective -*le*). There are several issues regarding the implicitness/explicitness of recasts. Recasts can be explicit if the interlocutor uses linguistic (such as repeating the wrong utterance in Doughty and Varela (1998)) and/or paralinguistic signals (such as through prosodic features) to convey the corrective intention. The second factor relates to the characteristics of recasts, that is, whether they are partial or full, whether they involve a single move or multiple moves, and so on (Loewen & Philp, 2006; Sheen, 2006). Still another factor has to do with the receiver of recasts, that is, whether the learner notices the corrective force of the feedback move. Of course whether recasts are noticed is, at least partly, contingent on the corrective intention of the interlocutor and the characteristics of recasts, but it is not entirely so. Other factors may also contribute to noticing such as the context where they are provided. For instance, recasts provided in mechanical drills are more noticeable than recasts provided in communicative tasks where the primary focus is on the exchange of information; recasts that target only one structure throughout are more noticeable than recasts directed toward multiple structures.

The above caveats do not undermine the relative implicit nature of recasts, at least in comparison with other corrective strategies such as metalinguistic feedback and explicit correction (Long, Inagaki, & Ortega, 1998; Long, 2006; Lyster, 1998). In this study, recasts have the following characteristics:

(1) They were provided in meaning-focused tasks where the target structures were

attended to in information exchange (details on treatment tasks are provided below).

(2) At the intra-utterance level, the recasts were mostly partial recasts reformulating the errors related to the target structures. However, it was not always possible to isolate the parts that contained the target structures, especially when local reformulation did not lead to a meaningful utterance. Therefore, recasts that involved the reformulation of the whole utterance were not rare in the dataset.

(3) At the inter-utterance level, aside from the utterances containing the target structures, utterances that subsumed errors related to non-target structures were responded to with recasts as well as other feedback types when the errors caused communication breakdown or misinterpretation. Attending to forms other than the target forms helped maintain the flow of communication and mask the linguistic foci.

(4) No linguistic or paralinguistic signals were utilized to convey the corrective intention on the interlocutor's part.

The following two episodes, which were extracted from the dataset of this study, exemplify how recasts were provided.

Episode 1

NNS: *wǒ zuótiān wǎnshang zhǐ shuì wǔ gè xiǎoshí
我 昨天 晚上 只 睡 五 个 小时。
I yesterday night only sleep-* [missing *Perf*] five-*CL* hour.
I only slept for five hours last night.

NS: zhǐ shuì le wǔ gè xiǎoshí
只 睡 了 五 个 小时。
Only sleep-*Asp* five-*CL* hour.
Only slept five hours.

NNS: shuì le wǔ gè xiǎoshí

睡 了 五 个 小时。
Sleep-*Asp* five-*CL* hour.
Slept five hours.

Episode 2

*NNS: zhè gè zhàopiàn shì liǎng gè zhū
这 个 照片 是 两 个 猪。
This-*CL* [wrong] photo be two-*CL* [wrong] pig.
This photo is two pigs.

NS: liǎng tóu zhū
两 头 猪。
Two-*CL* pig.
Two pigs.

*NNS: tóu zhū. zhū hěn pàng
头 猪。猪 很 胖。
CL pig. Pig very fat [inappropriate word choice].
(Two) pigs. The pigs are fat.

NS (laughs): zhū hěn féi dòngwù yīnggāi yòng féi
猪 很 肥。 动物 应该 用 肥。
Pig very obese. Animals should use obese.
The pigs are very obese. For animals, we should use *obese*.

In episode 1, the learner (NNS = nonnative speaker) failed to use the perfective *–le* to mark the completed and bounded event *slept for five hours*. The interlocutor (NS = native speaker) responded by reformulating the part that contained the error and adding the aspect marker. The learner then repeated the reformulation and incorporated the correct form in her utterance. Episode 2 is more complex: it has four moves and contains several errors and corrections. In the first move, the learner's utterance contains three errors: the wrong classifier for photos, the inappropriate use of the *be* verb, and the inappropriate use of *gè* as the classifier for pigs. In the second move, the native speaker only reformulated the noun phrase headed by *pigs* and replaced the wrong classifier with

the correct one; the nontargetlike use of the *be* verb and the wrong use of the classifier for *photo* (which is not on the list of target classifiers) were ignored. In the next utterance, the learner repeated the correct classifier and the noun, followed by a descriptive statement about the pigs in the photo where *pàng* (fat), a word used for human beings, was used for animals. The error relates to pragmatics and the native speaker responded by providing some metalinguistic explanation. This episode exemplified a recast on the use of classifiers as well as an additional corrective move that was utilized for the sake of natural communication and to hedge the linguistic focus.

Explicit Feedback

Carroll and Swain (1993) defined explicit feedback as "any feedback that overtly states that a learner's output was not part of the language-to-be-learned" (p.361). In their study, explicit feedback includes explicit hypothesis rejection, where a learner was told that she/he made a mistake followed by rule explanation, and explicit utterance rejection, where the learner was simply told that she/he made a mistake. Ellis et al. (2006) stated that explicit feedback can take two forms: metalinguistic feedback and explicit correction. Metalinguistic feedback refers to the linguistic information on the well-formedness of the learner's utterance (Lyster & Ranta, 1997) such as in "You need past tense" (Ellis et al., 2006). In this case, metalinguistic feedback is similar to Carroll and Swain's explicit hypothesis rejection. Explicit correction entails the message that the utterance is incorrect followed by the provision of the correct form as in "No, not goed—went". In Sheen's (2007a) and Li's (2009) studies, explicit feedback referred to the combination of explicit correction and metalinguistic feedback, that is, supplying the correct form followed by explicit rule explanation.

It is obvious that explicit feedback, whatever form it takes, must include a beacon message that unequivocally informs the learner about the ill-formededness of his or her L2 production. The message can be conveyed by simply stating that a mistake occurred and/or providing some sort of rule explanation, which can be brief or detailed. While it is without doubt that all forms of explicit feedback discussed above are "explicit", it is worth noting that they might facilitate L2 learning in different ways because of the different types of input they provide. Signaling the presence of a mistake and providing metalinguistic comments constitute negative evidence, but obviously the two types of negative evidence are different. Even when providing metalinguistic rule explanation, there can be much variation: it can be very brief or very detailed—"You need present perfect tense" vs. "You need past perfect tense because it is completed and has some effect on the present". Providing the correct form, on the other hand, constitutes positive evidence and does not involve the retrieval and processing of previously acquired forms. The point here is not to deny the legitimacy of the different ways to operationalize explicit feedback, but to bring to light the fact that it is critical to realize the different learning processes involved in varied forms of this feedback type.

Following Sheen (2007) and Li (2009), explicit feedback was operationalized as metalinguistic correction, that is, the provision of the correct form followed by explicit rule explanation. This operationalization is motivated by several reasons. First, as Sheen pointed out, metalinguistic correction is potentially more effective than metalinguistic feedback because of the availability of positive evidence in the former. Metalinguistic feedback, which only contains some rule explanation, might facilitate the learning of forms that are easily acquired, that do not involve complex rule explanation, and that

involve transparent complex form-meaning mapping (e.g. English regular past forms (Ellis et al., 2006)). It may not work for forms that involve complex form-meaning mapping and that may also require positive evidence in addition to negative evidence for the development of linguistic competence (e.g., English question formation (Loewen & Nabei, 2007)).

The two structures included in this study are Chinese classifiers and the Chinese perfective *–le*. The metalinguistic explanation for classifiers is simple[4], but informing the learner that a wrong classifier is used is unlikely to lead to the learner's use of the correct classifier if it is not part of his/her current interlanguage. This is because classifier use is to a large extent exemplar-based and the rule that a classifier must be used between a determiner and a noun only addresses the underuse, not the correct use, of classifiers. In the case of the perfective *–le*, because of its complex form-meaning mapping and because of the fact that it can appear in multiple positions of a sentence, the metalinguistic comment that a *–le* should be used does not guarantee the correct use of the structure.

The second reason behind the inclusion of the correct form plus metalinguistic explanation in the feedback type in question is that a major goal of this study is to explore the effects of implicit and explicit feedback and their constraining factors. Combining explicit correction and metalinguistic explanation, two explicit feedback types, would make the resultant feedback more explicit, hence increasing the contrast in the implicit-explicit dichotomy. One might argue that the addition of more information in the feedback or explicit feedback proper is likely to interrupt the flow of communication and that explicit feedback may only result in the development of explicit knowledge. However, as Ellis et al. (2006) pointed out, "the metalinguistic time-outs from

communicating afforded by explicit correction constitute a perfect context for melding the conscious and unconscious processes involved in learning" (p.343). Ellis et al. also demonstrated that explicit feedback in the form of metalinguistic feedback led to the learning of implicit knowledge.

Episode 3 shows how explicit feedback was provided to a learner's misuse of a classifier. As shown, when saying that there is a cigarette on the table, the learner misused the classifier for cigarettes. The native speaker reformulated the noun phrase with the classifier, followed by the metalinguistic information. It should be noted that the term "measure word" is used to refer to "classifier" in pedagogical Chinese grammar and in Chinese classes, although, as previously discussed, measure words are different from classifiers. To be consistent with the learners' classroom language, in this study the term "measure word", rather than "classifier" was used in providing metalinguistic correction on classifier use. It should also be noted that the explicit feedback was provided in English to make it accessible to the learners and to prevent the possibility of the non-incorporation of the feedback as a result of the learners' failure to comprehend the information the feedback contains. This is especially true of the explicit feedback for the perfective *–le*, which will be discussed in the next section.

Episode 3

NNS: * zài zhuōzi shàng yī gè yān

在 桌子 上 一 个 烟。

On table-*Prep* one-*CL* cigarette.

On the table there is a cigarette.

NS: yī zhī yān. The measure word is *zhī*.

一 支　　烟。 Provided feedback in English.

One-CL cigarette.

A cigarette.

NNS: yī　zhī　yān　　xièxie

一　支　烟。　谢谢。

One-CL cigarette.　Thanks.

A cigarette. Thanks.

While it was relatively easy to provide metalinguistic comments on the use of classifiers, phrasing the metalinguistic information for the usage of the perfective (verbal) *–le* posed a challenge. As discussed in the literature section, verbal *–le* is used in completed, bounded situations. There are two ways to delimit a situation: One is through the use of a number to atelic verbs (as in *sleep for two hours*, *eat three apples*) and the other is through the use of telic verbs that encode instantaneity and that have a natural endpoint (such as *die*, *drop*, *fall*, etc.). Based on Chinese pedagogical grammar (Li & Thompson, 1981), the metalinguistic feedback for the verbal *–le* was provided in two ways: one, for atelic verbs, is to inform the learner that *–le* is used with a number; the other, for telic verbs, is to inform the learner that *–le* is used with instantaneous verbs. Once again, in either situation, the metalinguistic information was provided in English to make it accessible to the learner. The following two episodes illustrate the two situations where metalinguistic correction was provided in response to the learners' wrong use of *–le*.

Episode 4

*NNS: nóngfū　zài　zhāi　lí,　　yī　gè　　lí　　diào

农夫　在　摘　梨。一 个　梨　掉。
Farmer *Prog*-pick pear. one-CL pear drop-[missing *Perf*]
A farmer was picking pears. A pear dropped.

NS: diào le. You need to use a *–le* here because it is completed and the verb *diào* is instantaneous.
掉了。Followed by feedback provided in English.
Drop-*Perf*.
Dropped.

Episode 5

*NNS: qùnián wǒ zài nàlǐ gōngzuò sān gè yuè.
去年　我 在 那里　工作　三 个　月。
Last year I at there work-[missing *Perf*] three-*CL* month.
Last year I worked there for three months.

NS: gōngzuò le sān gè yuè. You should use a *le* because it is completed and there is a number here.
工作　了 三 个　月。Followed by feedback in English.
Work-*Perf* three-*CL* month.

NNS: duì, *le*
对，了。
Yeah, le.
Yeah, (I should've used a) *le*.

In episode 4, the learner did not use *–le* with *drop*, a telic verb. The native speaker corrected the mistake by adding *–le*, followed by the provision of the metalinguistic explanation. In episode 5, the learner failed to use *–le* with an atelic situation that was bounded through duration of time. The native speaker added the aspect marker in his correction and the learner acknowledged the correction before the metalinguistic information was supplied.

Target Structures

The two target structures are Chinese classifiers and the Chinese perfective *–le*. The choice of these two structures is because they are different, they emerge early in learners'

interlanguage, and they pose challenges for learners at all stages of their study. As outlined in the literature section, the two structures differ on various dimensions: redundancy, saliency, form-meaning mapping, and explicit knowledge. These differences are summarized in Table 4. Another consideration in target structure selection is the amount of previous knowledge the learner has about the structure. It is speculated that feedback works best for structures the learner already has some knowledge about but has not yet fully mastered (Han, 2002; Mackey & Philp, 1998). Both Chinese classifiers and the perfective *–le* appear in the textbooks during the learners' first semester of study in the two programs where this study was conducted. The instructors were also consulted to make sure that the learners had some exposure to the two structures prior to the data collection. Previous research (Wen, 1995; Zhang, 2005) also demonstrated that the two structures appeared early in learner language.

Table 4. Chinese classifiers and the Chinese perfective *-le*

Dimensions	*Chinese Classifiers*	*Perfective -le*
Redundancy	• Not redundant—wrong classifier use likely causes communication breakdown	• Other features can be used to compensate for its absence
Saliency	• Salient	• Non-salient
Form-meaning mapping	• Transparent	• Opaque
Explicit knowledge	• Simple	• Complex
Learnability	• Relatively easy given sufficient input	• Difficult—advanced learners are not more accurate than beginners

Despite the early emergence of the two structures in L2 Chinese learners' interlanguage, they pose challenges for learners throughout their study. Wen (1995, 1997), Yang et al. (1999, 2000), and Li (2009) found that advanced learners did not outperform learners at beginning levels in their use and knowledge of the perfective *–le* or classifiers. This happened for different reasons: the usage of *–le* was complex and the input frequency of classifiers was low in textbooks for advanced learners. The fact that learners at both beginning and advanced stages have difficulty learning the two structures serves as another justification for selecting them for feedback treatment.

Recall that the perfective *–le* has two variants: the verbal *–le* and the sentence final *–le*. They appear in different positions and have different interpretations. Due to the sophisticated usage of this aspect marker and the limited amount of treatment each learner received, this study only focused on the effects of feedback on the learning of the verbal *–le*.

Tasks

Treatment Tasks for Classifiers

Two tasks were used where obligatory contexts for classifier use were provided[5]. The first task is called picture description in which the learner was asked to describe seven pictures that contained 15 cases of classifier use (see Appendix A for a sample picture). The pictures had different numbers of various objects (such as two trees, a river, three horses, etc.) so that the learner would have to use classifiers when they described the objects and reported how many of them there were. Distracter objects were included in addition to the objects related to the use of the selected classifiers. A vocabulary list (with Chinese characters, the Pinyin, and their English equivalents) was provided in each

picture that contained the nouns that accompany the classifiers the learner was expected to produce. Providing the vocabulary list also facilitated the flow of communication, especially for less advanced learners who did not have sufficient linguistic resources at their disposal. Also, the learner was allowed to ask the native speaker researcher vocabulary questions but not grammar questions. The sequence of the pictures was randomized so that each learner described them in a different order. The native speaker provided recasts or metalinguistic correction in response to the learner's wrong classifier use. The learners in the control group were asked to read a story about a Chinese idiom *shú néng shēng qiǎo* (Practice Makes Perfect) and retell the story by following some clues. A vocabulary list was provided to assist the story retelling, which did not require the use of the selected classifiers. No feedback was provided in the control condition.

At times, the native speaker must make conscious efforts to elicit the use of classifiers. In situations where the learner did not describe certain objects as desired such that the obligatory contexts for the use of the corresponding classifiers were not established, the native speaker would ask questions such as *zhàopiàn lǐ hái yǒu shénme*? (What else is in the picture?) to prompt the learner to talk about the objects related to the target structure. There were also cases where the learner only pointed out that there was something in a certain picture but did not state the quantity of the object, in which case there was no context for the use of the classifier and the learner might have done so to avoid using a classifier. The native speaker would then need to ask about the quantity so as to construct the context. The following example illustrates.

Episode 6

NS: hái yǒu shénme?
还 有 什么？

Still there be what?
What else is there?

NNS: mǎ
Horse.
马。
Horses.

NS: duōshǎo
多少？
How many?

*NNS: liǎng gè
两 个？
Two-*CL*?
Two?

NS: liǎng pǐ, liǎng pǐ mǎ.
两 匹， 两 匹 马。
Two-*CL*, two-*CL* horse.
Two, two horses.

The second task is called spot the difference, where there were three sets of pictures. Each set had two pictures that contained more or less the same items but the two pictures were different in a number of aspects (see Appendix B for a sample picture used in this task). The native speaker and the learner each held a picture, and the learner asked questions to find out what the differences were. Completion of the task required the use of the same 15 selected classifiers as appeared in task 1. As in task 1, the learner was provided a vocabulary list for each picture and was allowed to ask vocabulary related questions. The sequence of the three picture sets was randomized for each learner.

The selection of classifiers was based on the responses from 45 native speakers of Chinese to a survey on classifier use. The survey serves two purposes. One is to select appropriate “classifier + noun” combinations for treatment tasks. Although in most cases of classifier use there is a one-to-one correspondence between a classifier and the

accompanying noun, there are situations where more than one classifier is compatible with one object. For instance, there are two possible classifiers for dogs: *zhī* and *tiáo*. The other purpose of the survey is to make sure that the selected special classifiers can not be replaced by the general classifier *gè*. The general classifier can substitute for a special classifier in many situations, which confuses L2 Chinese learners and which partly explains why classifiers constitute a problematic structure.

The survey had 40 items, each providing a context for classifier use. For each item, the respondent was asked to fill in the missing classifier and then decide whether the classifier could be replaced by the general classifier. The surveyed classifiers were mostly selected from the textbooks used in the Chinese programs contributing participants for this study (Chou, Eagar, & Chiang, 1999; Zhang et al., 2002; Liu, Yao, Bi, Ge, & Shi, 2009); some were from other commercial Chinese textbooks (Zhao, Li, & Lin, 1999; Huang & Ao, 2002; Wu,, Yu, Zhang, & Tian, 2007) used in North America; others were from Erbaugh's list of core classifiers (1986) and the Chinese grammar book by Li and Thompson (1981).The example below shows a sample item in the survey.

Example

房间里有四_____椅子。(There are four chairs in the room)

该量词可否用"个"代替？(Can the measure word be replaced by *gè*?)

A. 是　B. 否 (A. Yes　B. No)

The respondents were 45 Mandarin native speakers studying or working in the local community where this study was conducted. Among them, there were three

undergraduate students, 20 graduate students, and 12 working at local companies or government agencies. 20 of them had a bachelor's degree, 19 had a master's degree, and 6 had a doctoral degree. Their specializations were varied, including humanities, science, and engineering. The average age was 32.08.

Altogether 15 cases of classifier use were selected out of the 40 surveyed items. In order to be eligible to be included in the study, a classifier must reach an agreement rate of 80% or higher among the respondents regarding the collocation of the classifier with the accompanying noun and the insubstitudability of the general classifier for the special one.

Treatment Tasks for the Perfective –le

Two tasks were used to elicit the production of the perfective *–le*: video narrative and interview. In the video narrative task, the learner was asked to watch a 7-minute video clip and tell what happened in the story. The video clip (with sound effects but no words), which is called *The Pear Film*, was created by Chafe (1980) to elicit narrative language samples. It started by showing a farmer picking pears. Then a boy came on a bike and stole a basket of pears. He went through some adventures before the farmer realized that the pears were missing (Erbaugh, 2001). The video is full of background and foreground events and has been used in numerous previous studies to elicit Chinese narratives (Christensen, 1994; Duff & Li, 2002; Yang, 2002) and investigate Chinese aspectual marking.

The learner was required to follow some provided clues when they retold the story. The clues are provided in English in the form of sentence fragments that contain obligatory contexts for the use of the perfective *–le*. The Chinese equivalents (with

Pinyin) of some key words in the clues are also provided to minimize the difficulty the learner is likely to encounter in finding the right vocabulary in the narrative. The learner was asked to speak Chinese but was allowed to ask vocabulary questions in English. The provision of clues serves two purposes. One is to free up the learner's cognitive demands from processing meaning so that linguistic forms can be attended to. VanPatten's claim (2002) that learners cannot process form and meaning simultaneously affords theoretical support for this practice. Another rationale is that learners are likely to avoid producing certain linguistic features if their knowledge about the features is incomplete and/or if they are unable to accurately use the features in communication (Gass & Selinker, 2008). Thus the requirement that the learner follow some clues that contain the target structure prevents the potential problem of avoidance.

To establish the obligatory contexts for the use of the perfective *–le*, the scripts of the oral narratives of the Pear Story from 40 native speakers of Chinese from Erbaugh, (1986) and Christensen (1994) were examined. Identification of obligatory contexts was also based on Li and Duff's detailed description of the use of *–le* by 9 nonnative speakers and 9 native speakers of Chinese in their narratives of the Pear Story (2002). After the obligatory contexts were established, they were matched with the corresponding contexts in the English scripts of the oral narratives of 20 native speakers of English from Erbaugh's study. The English clues are therefore from the speech data of native speakers of English.

In Task 2, which is called Interview, the learner was asked to answer 16 questions related to his/her recent experiences. The questions were written on flash cards in English and the Chinese translations of one or two potential new words were provided for each

question. Asking the questions in English prevents the modeling of the target structure as would happen if the questions were asked in the target language. The task was created to increase the number of tokens and types of the target structure. Recall that the Chinese perfective (verbal) *–le* is used in post-verbal positions in bounded situations; boundedness is encoded by either the inherent features of verbs (e.g., achievements) or, in the case of atelic verbs such as activity and accomplishment verbs, the addition of some external devices such as expressions of duration. Examination of native speakers' narratives of the Pear Story showed that the obligatory contexts for the use of the perfective *–le* were unevenly distributed among different verb types, a large number of which being verbs of achievements (such as "fall", "appear", "spill", etc.). Task 2 was therefore intended to supply more contexts for verb types other than achievement verbs such as activity and accomplishment verbs.

While performing the two tasks, learners in the experimental groups were provided with either explicit feedback in the form of metalinguistic correction or implicit feedback in the form of recasts in response to their wrong use of the perfective *-le*. While the target structure of these two tasks is the perfective *–le*, feedback was at times directed toward errors related to other structures. In Task 2 (see the sample interview question below), each interview question has at least two parts, one of which asks the learner about information other than that involving the use of the target structure. These moves were performed to minimize the learner's awareness of the target structure of the study. Learners in the control group were asked to answer some questions about their everyday life such as what type of food they like, whether they have a pet and why, and so on. Answers to these questions do not involve the use of the perfective *–le* and no feedback

was provided on any error. Learners in all groups were allowed to ask vocabulary related questions at any time in performing the tasks.

Example Interview Question in Task 2

How long do you sleep everyday? How long did you sleep last night? sleep 睡 shuì last night 昨天晚上 zuó tiān wǎn shàng

Testing

Table 5 provides the information on the different measures used in the study and the related descriptive statistics including the number of items, possible points, mean, standard deviation, range, and reliability coefficient. These measures include a proficiency test, tests of treatment effects (grammaticality judgment and elicited imitation), a language analytic ability test (the Words in Sentences subtest of the MLAT), and a working memory test. Details on these measures are elaborated on in the following sub-sections.

Table 5. Measures and descriptive statistics[a]

Measure		*Items*	*Points*	*Mean*	*SD*	*Range*	**Reliability**[f]
Proficiency—HSK		60	60	29.99	8.39	36.00	N/A[g]
Grammaticality judgment[b]	Perfective -*le*	15	15	5.05	2.09	9.00	0.63
	Classifiers	15	15	5.78	1.26	6.00	0.74
Elicited imitation	Perfective -*le*	15	15	3.84	3.50	14.00	0.87
	Classifiers	15	15	3.24	2.22	10.00	0.68
Language analytic ability		45	45	24.25	6.37	27.00	0.81
Working memory[c]	Reaction time	72	Ave[d]	3769.53[e]	523.63	2701.17	0.98
	Plausibility judgment	72	72	63.64	5.27	30.00	0.80
	Recall	72	72	50.79	9.84	43.00	0.89

Note. a. The results are based on the data contributed by all participants (n = 78).

b. Descriptive statistics related to the measures of treatment effects are based on all participants' pretest scores, and the information regarding different groups and their respective performances at different time points is presented in the results section.

c. The working memory score for each participant is the average of the z scores related to the three components of the test. The standard deviation of z scores is 1 and the mean is 0. Therefore, the descriptive statistics are computed for each component instead of the global working memory score.

d. Reaction time is the average of the reaction times related to all 72 items.

e. Reaction time was recorded in millisecond.

f. Cronbach's α is used as the reliability coefficient.

g. The test is maintained by Beijing Language and Culture University and reliability for the sample was not available.

Proficiency Test

In light of the fact that learners' proficiency is an independent variable and that the participants were recruited from different levels at two different institutions, a proficiency test was used to measure their linguistic competence. The test is a revised version of the HSK, a standardized test of Chinese as a foreign language sponsored by Beijing Languages and Cultures University and recognized by the People's Republic of China and numerous countries worldwide. It has three sections: listening, reading, and grammar. The test has three versions, which are for beginners, intermediate learners, and advanced learners respectively. In this study, the HSK basic test is used, which is for learners with 100-800 hours of classroom instruction and an accumulated vocabulary of 400-3,000 characters. The revised HSK basic test used in this study consists of 60 items: 30, 20, and 10 for listening, grammar, and reading respectively. Each item is assigned 1 point, the total score being 60. More weight was given to listening comprehension and grammar than to reading comprehension to match the format of the interventional treatment, where feedback was provided orally to errors in oral production. Learners were required to mark all answers on an answer sheet. Table 6 illustrates the items in each part of the test.

Table 6. An illustration of the HSK test

Part		*Item Description*	*No. of Items*	
			Part	*Total*
Listening	Choose a picture	(1) Listen to a statement describing a scenario (2) Choose one from the four given pictures that matches the scenario	8	30
	Choose a response	(1) Listen to a question (2) Choose an appropriate response to the question	7	
	Choose an answer	(1) Listen to a dialog or passage (2) Listen to a question about the dialog (3) Choose the right answer	15	
Grammar	Choose the right sentence	(1) Read four structurally similar sentences (2) Choose the one that is grammatically correct	10	20
	Choose the right word	(1) One word of each sentence is left out (2) Choose the one that can complete the sentence	10	
Reading	Choose the right answer	(1) Read a short passage and a long passage (2) Choose the right answer to each question	10	10

Tests for Treatment Effects

Tests of implicit and explicit knowledge. In order to measure the effects of feedback, two tests were used: elicited imitation (EI) and grammaticality judgment test (GJT). Previous research demonstrated that the two tests tapped into different types of knowledge: implicit knowledge and explicit knowledge (Erlam, 2006; Ellis, 2004, 2005, 2006; Ellis et al., 2006; Ellis, Loewen, Elder, Erlam, Philp, & Reinders, 2009). Implicit knowledge is unconscious, easily accessible, procedural, and intuitive; explicit knowledge is conscious, accessible through controlled processing, declarative, and verbalizable. Ellis et al. (2006) summarized how tests of implicit and explicit knowledge should be operationalized:

> Tests of implicit knowledge need to elicit use of language where the learners operate by feel, are pressured to perform in real time, are focused on meaning, and have little need to draw on metalinguistic knowledge. In contrast, tests of explicit knowledge need to elicit a test performance in which the learners are encouraged to apply rules, are under no time pressure, are consciously focused on form, and have a need to apply metalinguistic knowledge. (p.354)

When measuring the effects of interventional treatment, SLA researchers have the tendency to use tests that bias toward explicit knowledge and therefore fail to provide a complete picture of learners' improvement, or lack thereof, as a result of instruction. Research syntheses on the effectiveness of (different types of) L2 instruction also

revealed that different test formats yielded different results with regard to magnitude of effect (Li, 2010; Lyster & Saito, 2010; Norris & Ortega, 2000), which explains why there has been a call to include tests that measure both implicit and explicit knowledge in empirical research (Ellis et al., 2009).

Previous research validated elicited imitation as a test that measures implicit knowledge (Erlam, 2006). In an elicited imitation test, the learner is asked to listen to some statements on a range of topics. In each item, after listening to the statement, the learner decides whether the statement is true or not true for him/her and whether he/she is not sure[6]. Following the decision, the learner is asked to repeat the sentence correctly regardless of whether the statement is true or whether the learner is not sure. Erlam argued that such a test taps into learners' implicit knowledge because of the following characteristics:

(1) The primary focus of the test is on meaning rather than form. The test is described as a "survey questionnaire" which asks test takers for their opinion on statements relating to their everyday life.

(2) Learners' production is reconstructive in nature. Test takers are asked to repeat the sentence in a correct way. The repetition is reconstructive rather than rote imitation because there is a delay (distracter) between the presentation of the target stimulus and the reproduction of the sentence. Also, there is no significant correlation between length of the stimuli and success rate of repetition.

(3) Learners do not rely on explicit knowledge. Because of the spontaneity of learners' production, they are unlikely to draw on their metalinguistic knowledge about the target structure. It has been demonstrated (Ellis, 2005) that learners' performance

on an EI test and their performance on spontaneous oral tests are highly correlated, indicating that they tap into the same construct.

(4) Learners' reproduction is an indication of internalization. If learners successfully repeat (in the case of grammatical sentences) or repair a stimulus (in the case of ungrammatical sentences), it is evidence that they have internalized the target structure.

Grammaticality judgment tests (GJT) are generally believed to measure explicit knowledge. However, it may not necessarily be so. As Ellis pointed out (2004, 2005) and demonstrated, whether or not there is time constraint is an important factor in determining what type of knowledge a GJT measures. Whereas untimed GJTs tap into explicit knowledge, timed GJTs measure implicit knowledge. Under pressure, learners tend to rely on their hunch when making a judgment about whether a sentence is grammatical or not. Another issue is what type of GJT is used. Learners may be asked to simply make a decision regarding the grammaticality of a sentence, to identify the error, to correct the error, to state the rule, to indicate the degree of certainty regarding their judgment, or any combination of them (Ellis, 2004). In this study, an untimed GJT is used, which asked the learner to make a grammaticality judgment and locate and correct the error (details are provided below).

While it would be ideal to develop tests that measure distinctly different types of knowledge, it is not easy to do so. For instance, learners may still access explicit knowledge when performing online tasks under time pressure. By the same token, learners may recourse to their "feel of the language" when performing untimed offline tasks. For instance, native speakers or learners in naturalistic settings may not have

access to explicit knowledge. Learners' general proficiency may also affect the validity of a test. Less advanced learners, for instance, may possess less implicit knowledge than advanced learners regardless of how they are tested. The interface between knowledge type and test type is therefore complicated and may not be as clear-cut as speculated. However, although more research should be done to validate existing tests of the two types of knowledge, it may at least be safe to argue that in general elicited imitation tests tap into more implicit knowledge than explicit knowledge; and untimed GJTs are more likely to measure explicit knowledge than timed GJTs.

Elicited imitation test. An EI test was used in both the experiment for classifiers and the experiment for the perfective *–le* to measure learners' implicit knowledge as a result of the provision of feedback. The test is called "Survey Questionnaire" to suggest that the objective is to elicit for learners' opinion rather than measure their linguistic competence. During the test, learners were asked to listen to some statements related to their everyday life or their personal experience. The stimuli, which were read at normal speed by the researcher and were recorded on an audio disc, were presented manually by the researcher using a disc player. After learners heard each statement, the disc was paused to allow them to decide whether it was true, not true, or whether they were not sure. Learners were then asked to repeat each statement in correct Chinese. To prevent the likelihood that learners (especially low-level learners) fail to understand and repeat a sentence because of their lack of knowledge about the vocabulary rather than about the target structures, annotation was provided for some key words in each statement. The annotation includes the character(s), Pinyin transcript, and English explanation. Since the purpose of the test was not to measure learners' vocabulary knowledge but to measure

their ability to use the target structures, it was considered appropriate and necessary to supply some vocabulary explanation. Below are two example EI test items:

Example EI item for classifier use

Script: 我家附近有一个小河。(Near my house, there is a river)

A. True B. Not True C. Not Sure
(fùjìn 附近 nearby; hé 河 river)

Example EI item for the perfective *–le*

Script: 我昨天晚上睡 7 个小时。(I slept for seven hours last night)

A. True B. Not True C. Not Sure
(shuì 睡 sleep)

The EI test has three versions: pretest, immediate posttest, and delayed posttest. The three versions contain the same target items but different distracter items. The test has 15 target items, 7 of which are grammatical and 8 are ungrammatical. Both the pretest and the immediate posttest have 8 distracter items; 4 of them are grammatical and 4 are ungrammatical. Thus, the pretest and the immediate posttest each has a total of 23 items, 11 being grammatical and 12 ungrammatical. Among the 23 items, 15 are target items and 8 are distracters. The delayed posttests for classifiers and the perfective *–le* were administered in the same session and were therefore combined. The combined test has a total of 40 items, of which 15 relate to classifiers, 15 to the perfective *–le*, and 10 are distracters. The target items in the three versions of the EI test were randomized, so the order in which the items appeared was different in each version. The item randomization, combined with the fact that different distracter items were included in each version, was

intended to prevent the realization on the learner's part that the target items were the same across tests. During the exit interview (details of which will be provided later) after the whole project was finished, learners indicated that some items in different tests were similar but they were not exactly the same.

The 15 target items in the test for classifier use measure the 15 classifier uses involved in the treatment tasks. The obligatory contexts for classifier use in the treatment tasks (picture description and spot the differences) are the same as those in the test items. In other words, the same classifiers and their accompanying nouns were targeted in the treatment tasks and the test. Ungrammatical sentence stimuli were created by deleting a classifier or substituting the general classifier *gè* or a wrong classifier for a correct classifier. As to the 15 target items in the EI test for the perfective *–le*, the verb in each item also appeared in the treatment tasks (video narrative and interview). Care was taken to make sure that the verb types (activity, achievement, and accomplishment) used with *–le* were evenly distributed among the 15 items (5 for each of the three verb types)[7]. Ungrammatical sentences were created by omitting *–le* where it should have been used.

Grammaticality judgment test. Grammaticality judgment tests were used to measure learners' explicit knowledge about the target structures. Unlike in some previous studies where learners were only asked to judge whether a certain item was grammatical or ungrammatical, in this study learners were asked to judge whether a sentence was grammatical or ungrammatical or whether they were not sure. In cases where learners judged a sentence to be ungrammatical, they were asked to locate the error and correct it. Adding the choice of "not sure" or avoiding a binary choice minimizes the chances for random guesses on the learner's part and increases test validity. The addition of the

choice indicating the learner's uncertainty proved to be necessary because in coding the data, it was noticed that all the learners chose the option of "not sure" at least once on the GJTs. The decision to ask the learner to locate and correct the error when a sentence is ungrammatical is motivated by the speculation that the learner might judge the sentence to be ungrammatical without knowing precisely what the error is or even if the error is identified, how to correct it (Mackey & Gass, 2005). The data coders found that this was indeed the case—it was very common that learners made the right judgment in indicating that a sentence was ungrammatical when it was ungrammatical; but they either corrected a non-target structure or failed to provide the correct form for the wrong target structure.

As with the EI test, the GJT has three versions, a pretest, an immediate posttest, and a delayed posttest. The test has 15 target items and varying numbers of distracting items depending on the timing of the test. Among the 15 target items, 8 are ungrammatical and 7 are grammatical. The sentence stimuli in the GJT are different from those in the EI test except for the obligatory contexts for the use of the target structures, which involve classifiers and their accompanying nouns and the target verbs extracted from the treatment tasks for the perfective *–le*. The pretests for classifiers and the perfective *–le* were combined and were taken in the same session as the proficiency test. The combined test contains 15 target items for each target structure and 5 distracters, totaling 35 items. In the immediate posttest for each target structure, there are 23 items, out of which 8 are distracters. The delayed posttests for both target structures were merged and the resultant test had a total of 40 items, out of which 30 were target items and 10 were distracters. Different tests included a different set of distracters, which concern structures other than the target structures such as the *ba*- structure, word order, and so on. Vocabulary

annotation was provided and learners were allowed to ask vocabulary related questions. For each item, Pinyin is provided for each character. To avoid providing hints on how the characters in the sentence should be clustered or combined to form words, which is especially in the favor of less advanced learners, all characters and their corresponding Pinyin representations are equally spaced instead of being arranged in word units. Each test has 6 practice items modeling ways to correct mistakes in ungrammatical sentences, such as addition, deletion, replacement, and relocation. Since the study concerns the effectiveness of oral feedback and the test is not intended to measure character writing, corrections in Pinyin (Romanized orthographic system of Chinese characters) are acceptable. There was no time limit for the GJT. Here are two example GJT items:

Example GJT Item for -*le*

wǒ zài lù shang zǒu de shí hòu， kàn dào yǒu yī gè rén de qián bāo diào

我 在 路 上 走 的 时 候， 看 到 有 一 个 人 的 钱 包 掉。

[While I was walking, I saw someone's wallet drop (The translation was not provided in actual testing)]

(钱包 qiánbāo wallet; 掉 diào fall)

A. Grammatical B. Ungrammatical C. Not sure

Example GJT Item for Classifiers

jīn tiān wǒ de yóu xiāng lǐ yǒu sān fēng xìn

今 天 我 的 邮 箱 里 有 三 封 信。（邮箱 yóuxiāng mailbox）

[Today my mailbox had three letters]

A. Grammatical B. Ungrammatical C. Not sure

Validity and reliability of the GJT and EI tests. In order to ensure test validity, that is, the tests measure what they are intended to measure, all target items included in the GJT and EI tests items for both target structures were piloted with native speakers of Mandarin Chinese. To select 60 test items (15 for two target structures and two test formats), a pool of 150 piloting sentences were created, most of which were revised from sentences in the learners' textbooks and other commercial Chinese learning materials. The sentences involved the two target structures as well as other structures which are considered to be problematic to L2 Chinese learners such as the *ba* (把) structure, negation, sentence final question particles, and so on. The items were given to 15 native speakers who were told to judge whether a certain item was correct and correct the error if it was incorrect. The 15 L1 Chinese speakers were from different professions (graduate student, journalist, teacher of Chinese as a foreign language, and civil servant) in the U.S. and China, and they held at least a bachelor's degree. In order to be eligible to be included in the tests, an item must receive unanimous judgment in terms of its grammaticality and if it is ungrammatical, how it should be corrected.

Test of Language Analytic Ability

Learners' language analytic ability was tested using the Words and Sentences subtest of the MLAT (Carroll & Sapon, 1959, 2002), a most widely used aptitude test in SLA research. The subtest is used to measure language learners' sensitivity to grammatical structures or the "ability to handle the grammatical aspects of a foreign language" (Carroll & Sapon, 2002, p.3). In each item, learners are provided with a key sentence where a certain part is underlined and one or more comparison sentences with five underlined parts. Learners choose the one part in the sentence(s) that matches the

function of the designated part in the key sentence. The test has 45 items and learners are required to complete it within 15 minutes. One point is assigned for each item, so the total score of this test is 45.

Test of Working Memory

Learners' working memory capacity is measured using a listening span test. The rationale behind the decision to use a listening span rather than a reading span test is that the interventional treatment of this study involves oral feedback, which does not draw on learners' ability to store and process visual stimuli. The test was created by using the stimuli from Waters and Caplan (1996). There are 72 sentences divided into 4 sets of sentences at span sizes 3, 4, 5, and 6. The sentence stimuli have the following structures:

It was the woman that ate the apple. (cleft subject: CS)
It was the damaged car that the mechanic fixed. (cleft object: CO)
The police arrested the man that punched his dog. (object-subject: OS)
The story that the man told amused the audience. (subject-object: SO)

These sentences differ in number of propositions and syntactic complexity. CS and CO sentences have one proposition, but OS and SO sentences have two. CS and OS sentences involve canonical assignment of thematic roles (Agent + Theme) and are therefore easier to process than CO and SO sentences. Half of the sentences have verbs that require animate subjects and half have verbs that require inanimate subjects. Half of the sentences are plausible and half are implausible. Implausible sentences are constructed by "inverting the animacy of the subject and object noun phrases" (Waters & Caplan, p.55)

(e.g., “It was the dissertation proposal that defended the man”). All four included sentence types (CS, CO, OS, and SO) and two plausibility possibilities (“Good” or “Bad”) are evenly distributed among the test stimuli. In each set, there are a mixture of sentences with different structures and plausibility possibilities. The sequence in which sentence sets of different span sizes is presented is randomized. All stimuli are read by a native speaker of English who holds a master’s degree in education. The test is created by using DMDX, free software used in psycholinguistic studies to measure reaction time when visual and auditory stimuli are responded to.

During the test, the learner listens to each sentence in a certain set and decides whether it is plausible, that is, whether it is about something that could happen in the real world. When the whole set is finished, there is a pause; the learner recalls the final word of each sentence in that set and writes down the words on a blank sheet before starting the next set. Recalling is not subject to time constraint and the sequence in which words are recalled is not taken into consideration in scoring (and learners were so informed). Before responding to test stimuli, the learner is exposed to eight practice items. Reaction time, plausibility judgment, and recall accuracy scores are all recorded and the learner is informed that all three components are equally important. Unlike some previous studies that only include recall scores, this study also includes reaction time and plausibility scores because WM capacity should involve both the processing and storage functions and because previous studies (Waters and Caplan, 1996; Leeser, 2007) showed that learners traded off between different components, that is, they sacrifice one component for a better performance in another (such as when learners process slower to achieve more accuracy in word recall).

Procedure

The study has four sessions on four separate days. In session 1, the learner filled out a background questionnaire and took a proficiency test (HSK), which was followed by a GJT pretest with items targeting both classifier use and items that involve the use of the perfective *–le*. Students' performance scores on the proficiency test were used for group assignment. The combined GJT pretest was used to provide baseline data to detect treatment effects and for screening purposes: students who scored over 75% on items related to a target structure were considered overqualified (based on the speculation that there would be ceiling effects or that there would not be sufficient room for improvement) and were excluded from the study thereafter. The proficiency test lasted 50 minutes, and there was no time limit for the GJT. The time each learner took to complete the GJT varied from 20 to 30 minutes. At the end of session 1, the participant was asked to schedule the remaining three sessions in such a way that sessions 2 and 3 happen on two consecutive days and session 4 happen one week after session 3.

In sessions 2 and 3, the learner received feedback (implicit or explicit) on their erroneous use of the target structure (classifiers or perfective *–le*). A learner that was assigned to a certain feedback condition received that type of feedback in both sessions 2 and 3 on two consecutive days. For instance, a learner in the implicit group received recasts on his/her non-target-like use of classifiers and the perfective *–le* respectively in the two sessions. The same principle applied to learners in the explicit condition. The order in which the two tretment tasks (for either classifier use or the perfective *–le*) were completed was randomized. Prior to the treatment tasks, an elicited imitation (EI) test was administered, which served as a pretest. The EI test, which has 23 items, took around

10 minutes to complete. The treatment tasks lasted around 40 minutes. After the instructional treatment, the learner took the EI test and the GJT, which served as immediate posttests. It must be pointed out that the EI test always preceded the GJT to minimize the potential modeling effect of the written test on the oral test. The GJT, which has 23 items, took about 15 minutes. Each of the treatment sessions lasted approximately 80-90 minutes. It is to be noted that the order in which a learner participated in the two treatment sessions was randomized, that is, half of the learners participated the classifier session before participating the session on the perfective *–le* and half attended the session on *–le* before the session on classifiers.

During the final session (seven days after session 3), the learner took a delayed EI test (about 20 minutes) and GJT (both tests containing items for both target structures) (20 minutes), the test of language analytic ability (Part IV of the MLAT), and the working memory test (15 minutes). Finally, the learner participated in a semi-structured exit interview asking about how she/he felt about the study and whether she/he recognized the objectives of the study. Table 7 illustrates the procedure of the study.

Table 7. Procedure of the study

Session 1		*Session 2 (classifier)*[*]		*Session 3 (-le)*		*Session 4*	
Tasks	Duration	Tasks	Duration	Tasks	Duration	Tasks	Duration
• Proficiency test	50 min	• EI pretest	10 min	• EI pretest	10 min	• EI posttest 2	20 min
• GJT pretest	25 min	• Treatment tasks ▪ Picture description ▪ Spot the difference	40 min	• Treatment tasks ▪ Video narrative ▪ Interview	40 min	• GJT posttest 2	25 min
		• EI posttest 1	10 min	• EI posttest 1	10 min	• Aptitude test	15 min
		• GJT posttest 1	15 min	• GJT pposttest 1	15 min	• WM test	15 min
						• Exit interview	5 min

[*] *Note.* The order in which the leaner participated in the two treatment sessions was randomized, and so was the sequence for treatment tasks.

Scoring and Coding

GJTs and EI Tests

GJTs. The GJTs used in this study asked the learner to judge whether a sentence is grammatical or ungrammatical or whether he/she is not sure and then to correct the error if it is ungrammatical. The availability of multiple options and the obligation to correct errors in the case of an ungrammatical sentence led to a variety of possibilities in responding to each test item. A complete list of possible responses is shown in Table 8. However, some further elaboration is in order regarding the scoring scheme:

- If a sentence is grammatical but was judged to be ungrammatical, the answer received 1 point if a non-target structure was changed. But the answer received a zero if a change was made to the target structure such that the sentence became ungrammatical. For instance, a correct classifier was replaced by a wrong one or deleting the perfective *–le*.
- If a sentence is ungrammatical, it was judged to be ungrammatical, and the error was corrected, the answer received 1 point. However, if judgment was correct but a change was made to a non-target structure, no credit was given. In cases where the learner recognized the error (such as by marking the error) but failed to correct it, the answer received no point.
- One might question whether it is reasonable to give credit in cases where a grammatical sentence was considered ungrammatical but a change was made to a non-target structure and to give no credit in cases where an ungrammatical sentence was regarded as being ungrammatical but a change was made to a non-target structure (See Mackey & Gass, 2005 for further discussion on the scoring of GJT tests). During the

exit interview, all learners indicated that when they performed a correction about a certain part of a sentence, they believed that the rest of the sentence was correct. In other words, for a grammatical sentence where the target structure is correctly used, if a learner changed a part other than the target structure, it should be assumed that the learner did not believe that the use of the target structure was problematic. Therefore it can be concluded that he/she had the knowledge about how the target structure is correctly used. By the same token, for an ungrammatical sentence where the target structure is wrongly used, if the learner corrected a part other than the target structure, it means that she/he did not have the knowledge about the structure even if she/he made the correct judgment (ungrammatical judged as ungrammatical).

In addition to a generic coding and scoring scheme, additional criteria were established for the test data related to each of the target structures because of their idiosyncratic linguistic features. Recall that the perfective *–le* is used in bounded situations and boundedness is encoded through numerical expressions (indicating temporality or quantity) with atelic verbs (describing actions without a natural endpoint such as "study") or through the inherent instantaneous nature of telic verbs (describing actions with a natural endpoint such as "fall"). Therefore, in the case of atelic verbs, the correct use of this structure can be illustrated in this formula:

(1) V_{atelic} + *le* + Numeric Expression [+ Object (if the verb is transitive)]

In the case of telic verbs, a numeric expression is unnecessary because boundedness is expressed by the verb itself. Hence this formula:

(2) V_{telic} + *le*

Table 8. Coding and scoring of GJTs

Stimuli	*Learner's Judgment*	*Learner's Correction*	*Score*
Grammatical	Grammatical	No correction	1
	Ungrammatical	Corrected a non-target structure	1
	Ungrammatical	Replaced the correct structure with a wrong one	0
	Not sure	No correction	0
Ungrammatical	Ungrammatical	Corrected the target structure	1
	Ungrammatical	Corrected a non-target structure	0
	Grammatical	No correction	0
	Ungrammatical	Marked the error and/or made a wrong correction	0
	Not sure	No correction or corrected a non-target structure	0

In cases involving formula (2), correctness can be easily determined by the absence or presence of *–le*; cases involving formula (1) are more complex. When performing corrections for these sentences, some learners showed the patterns as listed in Table 9, making it challenging to score. The patterns are based on the following sentence:

zuótiān wǒ xué le sān gè xiǎoshí zhōngwén
昨天 我 学 了 三 个 小时 中文。
Yesterday I study-Asp three-*CL* hours Chinese
Yesterday I studied Chinese for three hours.

These corrected sentences are still problematic after learners' modifications, but half a point was assigned based on the following rationale. Although the sentences are ungrammatical, but as far as the target structure is concerned, the obligatory contexts were established and the wrong modifications do not seem to result from a lack of knowledge about the target structure. Problems in these cases mostly seem to relate to sentence order, which is of no surprise given the cross-linguistic difference between English and Chinese in this regard, especially in the location of temporal expressions. Another problem lies with the use of an additional *–le* (3 and 4), which might result from learners' knowledge that there are two *–le*s in Chinese: a verbal *–le* and a sentence-final *–le*. The additional *–le* (presumably sentence-final *–le*) in (3) might have been accidentally placed before the final word of the sentence. Regardless, it was considered appropriate to assign partial credit to these cases because the obligatory contexts were created, the morpheme was supplied, and the problems were not caused by the lack of knowledge about the target structure.

Table 9. Additional criteria regarding GJT data on *-le*

Sentences after Correction	*Problem*
1. *zuótiān wǒ xuéle zhōngwén sāngèxiǎoshí Yesterday I study-*Asp* Chinese three hours Yesterday I studied Chinese for three hours	The temporal expression "three hours" does not follow *–le.*
2. *zuótiān wǒ sāngèxiǎoshí xuéle zhōngwén Yesterday I three hours study-*Asp* Chinese Yesterday I studied Chinese for three hours	The temporal expression precedes the verb "study".
3. *zuótiān wǒ xuéle sāngèxiǎoshíle zhōngwén Yesterday I study-Asp three hours-*Asp* Chinese Yesterday I studied Chinese for three hours	An additional *–le* is used, which is after the temporal expression "three hours".
4. *zuótiān wǒ xuéle zhōngwén sāngèxiǎoshíle Yesterday I study-Asp Chinese three hours-*Asp* Yesterday I studied Chinese for three hours	An additional *–le* is used at the end of the sentence.
5. *zuótiān wǒ zhōngwén xuéle sāngèxiǎoshí Yesterday I Chinese study-*Asp* three hours Yesterday I studied Chinese for three hours	The object "Chinese" precedes the verb "study".
6. *zuótiān wǒ xuézhōngwénle sāngèxiǎoshí Yesterday I study Chinese-*Asp* three hours Yesterday I studied Chinese for three hours	*-le* does not follow the verb.

Whereas the difficulty in scoring the GJT data concerning the perfective *–le* is attributable to the complexity of the rules governing the use of the morpheme, a different set of problems arose in scoring the test data on classifier use. The problems mainly lie in learners' difficulty in spelling out classifiers using the Romanized Pinyin system (including tones) and in writing characters. Additional scoring criteria were created to code with these problems. Among the following listed cases, the first four received half a point and the last one received a full point:

(1) The correct Pinyin for a classifier is provided but with a wrong tone (e.g., "zhì" for "zhī").

(2) The correct Pinyin for a classifier is provided but without a tone (e.g. "tou" for "tóu");

(3) A Pinyin is provided that differs from the correct Pinyin by one sound but that is close enough to the correct pronunciation to allow the reader to pinpoint the corresponding character (such as replacing a sound with one that involves the same place of articulation ("chī" for "zhī") or adding a nasal ("bǎn" for "bǎ").

(4) Providing a character with the same Pinyin as the right classifier but with a different tone (e.g., "跳 tiào" for "条 tiáo");

(5) Providing a character that is a homophone of the required classifier ("风 fēng" for "封 fēng").

EI tests. Unlike GJTs, which are written, visual, and without time constraint, EI tests are aural, involve oral reconstruction, and are taken under time pressure. The scoring of EI tests therefore involves different criteria. Full credit was given to cases where the target structure was supplied in obligatory contexts. This would mean that no credit was given if the target structure was supplied but the context for the use of the structure was

not established; it also means that scoring only focused on the use of the target structure and the rest of a reproduced sentence was ignored. Also, the purpose of an EI test is to measure learner's implicit knowledge, which is unconscious and automatic. Therefore cases containing self-correction, which shows the learner's conscious processing of the target structures, did not receive credit. These generic rules, of course, did not suffice to account for all the varied responses to the provided stimuli. The examples shown in Table 9 are representative of special cases in the EI test data.

In example (1) of Table 10, the noun phrase *qúnzi* was mispronounced as *kūnzi* (probably because of the absence of the consonant /q/ in English), but the error was ignored in scoring because the correct classifier was produced and the error was committed on a non-target structure. Examples (2), (3), (4), and (5) all involve self-correction, but in (2) and (3), the first attempts are erroneous but the second attempts are correct; in (4) and (5), the learners were correct at first but then they changed the targetlike uses into nontargetlike uses. In either case, no credit was given. In examples (6) and (7), although the produced sentences sound awkward, the perfective *–le* was used and the obligatory contexts were established containing activity verbs and bounding devices—temporal expressions. Therefore partial credit was given. In contrast, (8) and (9) did not receive any credit even though the target structures were produced because the obligatory contexts were not established. In (8), the noun that the classifier accompanies was not provided; in (9), there is no bounding device (temporal expression) to necessitate the use of *–le*.

Table 10. Scoring of EI Data

Category	*Example*	*Score*
Problem with non-target structure	(1) liǎng tiáo kūn [the correct pronunciation is *qún*] zi 两 条 昆 [裙]子。 two-*CL* skirt Two skirts.	1
Self-correction	***From wrong to correct:***	0
	(2) wǒ měi gè yuè xiě yī gè yī zhāng zhīpiào jiāo fāngzū 我 每 个 月 写 一 个 一 张 支票 交 房租。 I every month write one-*CL* one-*CL* check pay rent Every month, I write a check to pay my rent.	
	(3) zuótiān wǎnshang wǒ shuì shuìle qī gè xiǎoshí 昨天 晚上 我 睡， 睡了 七 个 小时。 Yesterday night I sleep, sleep-*Perf* seven-CL hour. Last night I slept seven hours.	
	From correct to wrong:	0
	(4) qùnián xuéle liǎng gè yuè zhōngwén xué liǎng gè yuè zhōngwén 去年 学了 两 个 月 中文， 学 两 个 月 中文 Last year study-*Asp* two-*CL* month Chinese, study two-*CL* month Chinese. Last year [I] studied Chinese for two months.	
	(5) wǒjiā qiánmiàn yǒu liǎngkē liǎngshù 我家 前面 有 两棵， 两树 my home in front of there be two-*CL* two trees.	

Table 10 (cont'd)

	In front of my home, there are two trees.	
Context established	(6) zhù le wǒde péngyou jiā yī gè xīngqī 住 了 我的 朋友 家 一 个 星期。 Live-*Perf* my friend home one-*CL* week. [I] lived in my friend's house for a week.	.5
	(7) xué zhōngwén le liǎng gè yuè 学 中文 了 两 个 月。 study Chinese-*Asp* two-*CL* months. [I] studied Chinese for two months.	
Context not established	(8) wǒ měitiān wǔ zhī 我 每天 五 支。 I everyday five-*CL* Everyday I [smoke] five [cigarettes].	0
	(9) zài běijīng xuéle zhōngwén 在 北京 学了 中文。 in Beijing study-*Asp* Chinese. [I] Studied Chinese in Beijing [for two months].	

Inter-coder reliability. All data were coded by two native speakers of Mandarin Chinese: the researcher and an experienced instructor of Chinese as a heritage language. At the time of data collection, the researcher has a master's degree in linguistics and is an ABD in second language acquisition. The Chinese instructor has a bachelor's degree in ESL. Both coders have many years of experience teaching ESL and Chinese.

Altogether four rounds of coding were performed. Initially the two coders coded 10% of the test data and created a coding scheme after extensive and intensive discussion. The data subjected to initial coding include data related to pretests, immediate posttests, and delayed posttests. Following the scheme both coders agreed upon, the two coders coded all test data, which involved 7,020 codes for the GJT and EI tests on each of the target structure. The two coders then checked all the codes once again to make sure that their coding was accurate and consistent. The agreement rate for GJT codes is 98.3%, and for EI codes it is 97.6%. In the final round of coding, the two coders carefully examined the codes they had disparity on and resolved the differences after detailed discussion. For the EI data, the two coders transcribed all the responses verbatim, compared their transcripts, and resolved the differences prior to scoring.

The Working Memory Test

During the working memory test, learners were asked to listen to 72 sentence stimuli divided into 4 span sizes (3, 4, 5, and 6) and three sets at each span size, decide whether each sentence makes sense, and recall the last word of each sentence after listening to all stimuli in a certain set. Half of the 72 sentences are plausible and half are implausible. A WM score for each learner has three components: plausibility judgment, reaction time, and recall accuracy. The raw score for plausibility judgment is 72, with 1 point assigned

for each correct judgment. Reaction time was only calculated for correctly judged items. The full score for recall accuracy is also 72, with each accurately recalled sentence-final word receiving one point. There was no penalty for errors related to inflectional morphemes (such as "worked" recalled as "work") when recalled words were scored.

Analysis

This study investigates whether the effects of implicit and explicit feedback are constrained by learners' proficiency, the choice of target structure, and learners' individual differences in language analytic ability and working memory capacity. To answer the question of whether the two types of feedback work differently for high and low learners in the learning of the perfective *–le*, mixed design repeated measure ANOVAs were performed separately for data generated by the GJT and EI tests. The within-group variable is the timing of tests (pretest, posttest 1, and posttest 2), and the between-group variables are feedback type (implicit, explicit, and control) and proficiency (high and low). Subsequently, one-way ANOVAs and post hoc contrasts were conducted on gain scores to detect group differences. The same analytic procedures were repeated for the data on classifiers.

Prior to the statistical analyses, different tests were conducted to investigate the assumptions of parametric statistics. Table 11 displays the results of Shapiro-Wilk's tests of normality regarding the performance scores on both the GJT and EI tests of each group as defined by feedback type, proficiency, and timing of posttests. As shown, among the 72 group scores, 63 are normally distributed. The Mauchly's test was performed for each repeated measure analysis and the results showed that the assumption of sphericity was not violated. Levene statistic was examined for follow-up ANOVAs, and it was found

that the assumption of homogeneity of variances was met.

Table 11. Tests of normality

Test	*Proficiency*	*Group*	*Perfective -le*			*Classifiers*		
			Pretest	Post 1	Post 2	Pretest	Post 1	*Post 2*
GJT	Low	Implicit	.92	.91	.97	.91	.89	*.91*
		Explicit	.94	.96	.93	.87	.96	*.85*[*]
		Control	*.77*[*]	.92	.88	*.79*[*]	.92	*.75*[*]
	High	Implicit	.96	.90	.90	.91	.95	*.93*
		Explicit	.91	.89	.92	.87	.92	*.90*
		Control	.83[*]	.94	.98	.91	.95	*.91*
EI	Low	Implicit	.94	.88	.91	.87[*]	.92	*.89*
		Explicit	.88	.94	.95	.88	.90	*.85*[*]
		Control	.71[*]	.88	.90	.92	.98	*.93*
	High	Implicit	.89	.88	.92	.95	.94	*.94*
		Explicit	.85[*]	.91	.87	.90	.94	*.89*
		Control	*.89*	*.96*	*.91*	*.99*	*.91*	*.96*

Note. [*] The significance value is below .05, which means that the scores related to the condition are not normally distributed.

In addition to using *p* values to determine whether group differences were significant, effect sizes (Cohen's *d*) were calculated to explore if the effects of feedback were different across different test formats and target structures. While a *p* is useful in deciding whether to reject or accept a null hypothesis, it provides no information on the magnitude of an effect or relationship. The effect size, in contrast, indicates "the

magnitude of an observed difference between two groups in standard deviation units" (Norris & Ortega, p. 442). Cohen's *d*, one of the most commonly used effect size indexes for group differences, is calculated through dividing mean difference by pooled standard deviation (which takes into account sample sizes and standard deviations of both groups involved). An effect size of 0.2 is small, 0.5 is a medium effect, and 0.8 suggests a large effect. Examining effect sizes makes it possible to examine the effect of a certain instructional intervention across different conditions such as the effects of feedback on the learning of different target structures.

Pearson's correlation analyses were used to probe into the relationship between feedback type and learners' individual differences in language analytic ability and working memory capacity. Instead of data from all participants in the study, included in the correlation analyses were only the data from learners who were in their 4th semester of study. Recall that the participants were at different stages of their study at the time of data collection. Among the 78 recruited participants, 41 were in their 4th semester of study and 37 were in their 6th and 8th semesters of study. Performing correlation analyses on all participants would be less ideal because of the heterogeneity among them in terms of the amount of prior instruction the learners received, which might to some extent mask the relationship between aptitude components and treatment effects. This is because any relationship between aptitude and learning has to be interpreted as follows: given the same amount of instruction, learners with higher aptitude (or higher ability in a certain aptitude component) achieve more or progress at a faster rate. Therefore, to have a clearer picture of the role of aptitude in learning, the more dimensions learners are comparable

on, the more reliable the results are. As far as this study is concerned, it would be ideal to conduct separate correlation analyses on learners at different levels and with similar amount of prior instruction. However, the number of learners from higher level classes (n = 20 and n = 17 including those assigned to the control groups) is too small for correlation analyses, hence the decision to conduct the analyses only with the learners in their 4^{th} semester of study.

CHAPTER 4 RESULTS

Chapter 3 detailed the methodology of the study including participant information, feedback operationalization, the target structures, tasks, testing, procedure, scoring, coding, and analytic procedures. This chapter presents the results and summarizes the results by answering the research questions advanced at the end of the 2nd chapter.

Analyses regarding treatment effects are conducted by target structure (perfective *-le* and classifiers) and test (GJT and EI) and results will be presented accordingly. For the GJT and EI data on each target structure, descriptive statistics will be presented regarding the means and standard deviations of the three involved groups: implicit, explicit, and control. These are followed by the results from the repeated measure analyses on how the effects of feedback type are mediated by proficiency. Mean gain scores and standard deviations of each feedback group at each proficiency level will be calculated, and post hoc group contrasts will be conducted on the gain scores. Results from the two test formats will be compared using effect sizes to explore whether feedback contributes more to the acquisition of implicit knowledge or explicit knowledge in the learning of each target structure. Effect sizes will also be used to determine whether feedback affects the learning of the two target structures differently. As to the results pertaining to the two aptitude components, separate correlation analyses are conducted on the gain scores of each feedback group and learners' performance scores on the MLAT subtest and the working memory test. Descriptive statistics will also be presented.

Results on the Perfective *-le*

GJT Results

Table 12 shows the descriptive statistics of the GJT scores on the perfective *-le*

including means and standard deviations of pretest and posttest scores of each group. The changing patterns of the pretest and posttest scores of the three groups are shown in Figure 4. A one-way ANOVA analysis was performed on the pretest scores of the three groups and no significant difference existed between the three groups, F (2, 77) = 1.66, p = 0.2. The mean scores of all three groups increased over time, and the scores of the explicit group appeared to have dropped the most from time 2 to time 3.

Table 12. Perfective *–le*: Descriptive statistics on GJT scores

Condition	n	Pretest		Posttest 1		Posttest 2	
		M	SD	M	SD	M	SD
Implicit	28	6.52	1.68	8.75	2.27	9.75	2.91
Explicit	29	5.56	2.03	11.65	1.74	9.94	2.99
Control	21	5.74	2.58	6.98	2.93	7.62	2.75

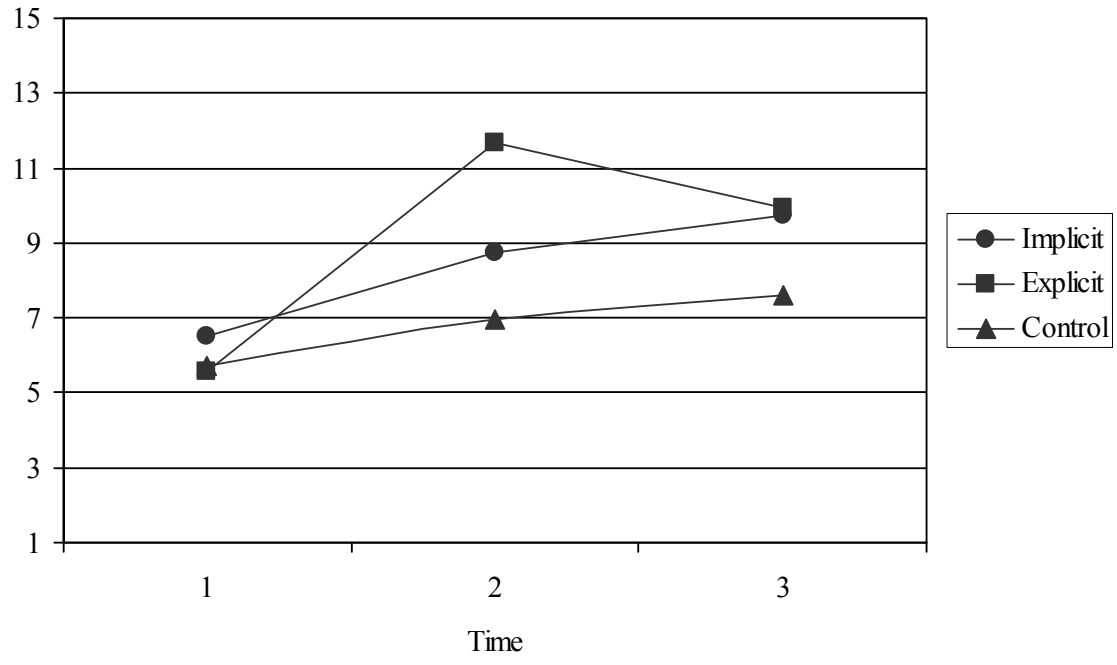


Figure 4. Perfective *–le*: GJT score changes

A 3 × 2 × 3 mixed design repeated measure ANOVA was conducted to obtain a general picture of how GJT score variation was impacted by feedback type (3 groups: implicit, explicit, and control) and proficiency (2 levels: low and high) and the timing of testing (3 time points: pretest, immediate posttest/posttest 1, and delayed posttest/posttest 2). As shown in Table 13, significant main effects were found for time, F (2, 142) = 92.15, $p < .05$, feedback, F (2, 72) = 14.45, $p < .05$, and proficiency, F (1, 72) = 92.15, $p < .05$. An interaction effect was found between time and feedback, F (4, 142) = 15.54, $p < .05$. The three-way interaction between time, feedback, and proficiency and the two-way interaction between feedback and proficiency approached significance.

In order to determine the sources of difference, the gain scores of each feedback group at each proficiency level at posttest 1 and posttest 2 were subjected to post hoc analyses. Gain scores were obtained by subtracting pretest scores from posttest scores. The descriptive statistics of the gain scores appear in Table 14 (see Appendix C for the descriptive statistics related with the raw scores of different proficiency levels). Results pertaining to group contrasts and the corresponding effect sizes (Cohen's *d*) are shown in Table 15. As shown, at the low proficiency level, the explicit group outperformed the control group and the implicit group at both posttests; the implicit group did not show significant improvement at either posttest. At the high proficiency level, the implicit group did not perform significantly better than the control group at the time of posttest 1, but they did at posttest 2; learners benefited more from explicit feedback than implicit feedback at posttest 1 but the difference between the two feedback groups did not significantly differ at posttest 2. Examination of effect sizes showed that the effects of explicit feedback dropped substantially from posttest 1 to posttest 2 at both proficiency

levels and that the effects of implicit feedback increased over time at the high proficiency level.

Table 13. Perfective *–le*: ANOVA results related to GJT scores

Source	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*
Within-Group Results					
Time	492.16	2	246.08	92.15	.00
Time × Feedback	165.96	4	41.49	15.54	.00
Time × Proficiency	15.02	2	7.51	2.81	.06
Time × Feedback × Proficiency	7.12	4	1.78	.67	.62
Between-Group Results					
Feedback	70.68	2	35.34	14.45	.00
Proficiency	127	1	127	51.91	.00
Feedback × Proficiency	14.45	2	7.23	2.95	.059

Table 14. Perfective *–le*: Descriptive statistics on GJT gain scores

Proficiency	*Group*	*n*	*Gains at Posttest 1*		*Gains at Posttest 2*	
			Mean	SD	Mean	SD
Low	Implicit	14	2.00	2.16	2.14	2.07
	Explicit	15	6.29	1.93	4.18	2.91
	Control	10	1.40	1.76	1.55	.93
High	Implicit	14	2.46	2.53	4.32	2.15
	Explicit	14	5.69	1.96	4.42	2.89
	Control	11	1.09	2.77	2.18	1.77

Table 15. Perfective *–le*: Post hoc contrasts related to GJT scores

Low Proficiency				*High Proficiency*			
Posttest 1		Posttest 2		Posttest 1		Posttest 2	
Contrasts	ES	Contrasts	ES	Contrasts	ES	Contrasts	ES
I—C	.30	I—C	.29	I—C	.52	I—C*	1.03
E—C*	2.62	E—C*	1.03	E—C*	1.94	E—C*	.92
E—I*	2.10	E—I*	.81	E—I*	1.42	E—I	.04

Note. ES = effect size; I = implicit; E = explicit; C = control

* $p < .05$

EI Test Results

The descriptive statistics regarding learners' performance on the elicited imitation tests appear in Table 16. Overall, EI pretest scores are lower than GJT pretest scores, indicating that learners had less implicit knowledge than explicit knowledge about the target structure prior to the treatment. The standard deviations of EI scores are in general larger than those of GJT scores, suggesting that learners were more homogeneous in their explicit knowledge about the target structure. Figure 5 shows the development patterns of the three groups over time. Evidently the two experiment groups improved substantially after treatment but the control group did not undergo substantial change. As with the GJT results, the explicit group seemed to have dropped the most from posttest 1 to posttest 2. One-way ANOVA conducted on the pretest scores of the three groups showed that there was no significant difference between them before treatment, F (2, 77) = .56, p = .57.

Table 16. Perfective *–le*: Descriptive statistics on EI test scores

Condition	n	Pretest		Posttest 1		Posttest 2	
		M	SD	M	SD	M	SD
Implicit	28	4.14	3.59	8.71	3.47	7.61	4.12
Explicit	29	3.26	3.36	9.98	3.19	7.69	4.00
Control	21	4.67	3.63	5.60	4.26	5.12	3.12

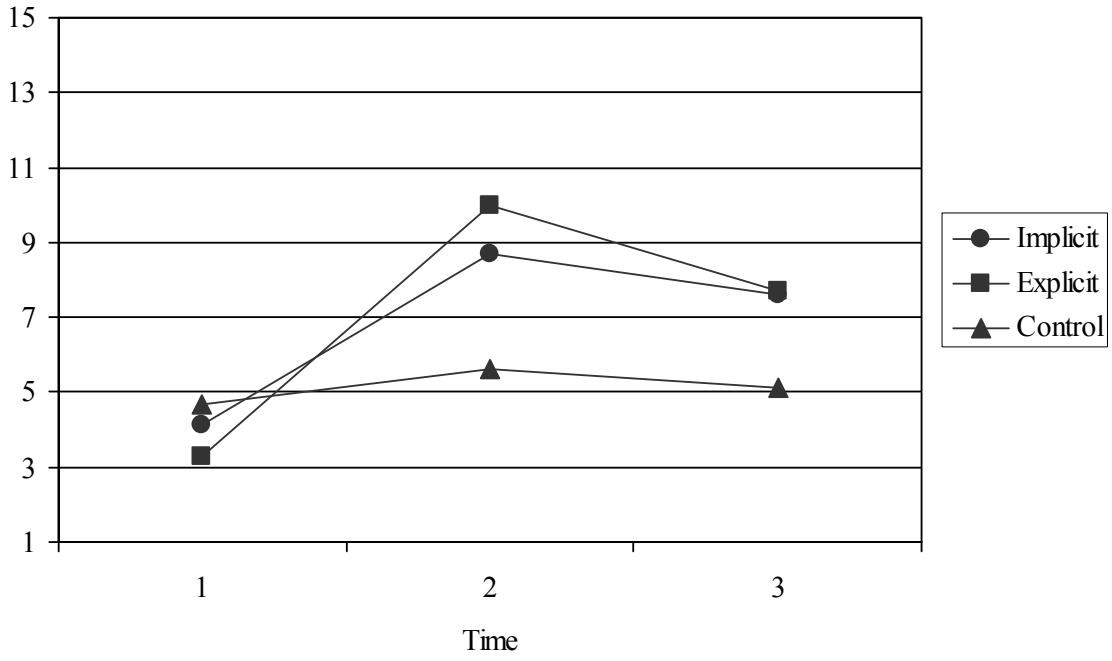


Figure 5. Perfective *–le*: EI score changes

The EI test scores pertaining to the perfective *–le* were subjected to a mixed design repeated measure ANOVA, with time as the within-group variable and feedback and proficiency as between-group variables. Results (Table 17) revealed that there is a significant effect for time, F (2, 142) = 113.48, $p < .05$, for time* feedback interaction, F (4, 142) = 13.98, $p < .05$, for feedback, F (2, 72) = 5.96, $p < .05$, and for proficiency, F (1, 72) = 74.66, $p < .05$. In order to identify group differences, gain scores were calculated for the six groups that were formed by feedback and proficiency and the results are displayed in Table 18 (see Appendix C for the descriptive statistics related with the raw scores of different proficiency levels).

Table 17. Perfective *–le*: ANOVA results related to EI test scores

Source	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*
Within-Group Results					
Time	704.89	2	352.44	113.48	.00
Time × Feedback	173.68	4	43.12	13.98	.00
Time × Proficiency	3.67	2	1.83	.59	.55
Time × Feedback × Proficiency	14.75	4	3.69	1.19	.32
Between-Group Results					
Feedback	66.69	2	33.34	5.96	.00
Proficiency	417.76	1	417.76	74.66	.00
Feedback × Proficiency	4.68	2	2.34	.42	.66

Table 18. Perfective *–le*: Descriptive statistics related to gain scores on EI tests

			Gains at Posttest 1		*Gains at Posttest 2*	
Proficiency	*Group*	*n*	Mean	SD	Mean	SD
Low	Implicit	14	4.39	2.58	2.68	2.58
	Explicit	15	6.86	2.78	3.93	2.58
	Control	10	1.15	2.07	1.35	1.53
High	Implicit	14	4.75	2.38	4.25	2.56
	Explicit	14	6.57	3.30	4.96	3.12
	Control	11	1.68	1.79	0.59	2.04

Results of post hoc group comparisons (Table 19) showed that at the lower proficiency level, learners benefited from both explicit feedback and implicit feedback as reflected on posttest 1, but the difference between the implicit condition and the control group was not significant on posttest 2. Effect sizes for both feedback types underwent remarkable decrease. At the higher proficiency level, learners benefited from both feedback types at both posttests. Also, it appeared that the effects of both feedback types were well maintained at the higher proficiency level, with a slight decrease for explicit feedback and an increase for implicit feedback. As to implicit-explicit contrasts, the only significant difference between the two feedback groups was found for low-level learners on the immediate posttest.

Table 19. Perfective *–le*: Post hoc contrasts related to EI test scores

Low Proficiency				*High Proficiency*			
Posttest 1		Posttest 2		Posttest 1		Posttest 2	
Contrasts	ES	Contrasts	ES	Contrasts	ES	Contrasts	ES
I—C*	1.36	I—C	.60	I—C*	1.43	I—C*	1.56
E—C*	2.27	E—C*	1.17	E—C*	1.80	E—C*	1.63
E—I*	.92	E—I	.49	E—I	.64	E—I	.25

Note. ES = effect size; I = implicit; E = explicit; C = control

* $p < .05$

Summary of the Results on the Perfective –le

Based on the results reported above, we turn to research question 1, that is, whether the two types of feedback have differential effects on learners of different proficiency levels in the learning of the Chinese perfective *–le*, the answer is affirmative. More specifically, the following results were obtained:

(1) The effect of implicit feedback on low proficiency learners was limited. The GJT results showed that the implicit group did not perform significantly better than the control group on either posttest. Although there were significant effects for implicit feedback on the immediate EI test, but the effects did not sustain.

(2) Implicit feedback was effective for high proficiency learners. On the EI test, implicit feedback showed large effects on both posttests; on the GJT test, although the implicit group did not perform substantially better than the control group on the immediate posttest, but it did on the delayed posttest. Also, the effects of implicit feedback increased over time: On both the GJT and EI tests, the effect sizes associated with the delayed effects are larger than those associated with the immediate effects.

(3) Explicit feedback was beneficial to learners at both proficiency levels as reflected on both test formats.

(3) The superiority of a certain feedback type seems to depend on proficiency level, test type, and timing of test. Out of the eight contrasts between the two feedback groups, four are significant in favor of explicit feedback. Out of the four significant contrasts, three pertain to low proficiency learners, GJTs, and immediate posttests. In other words, explicit feedback tended to be more effective than implicit feedback to less advanced learners when treatment effects were measured by using tests that favored explicit knowledge; the difference between the two feedback types tended to disappear over time.

Results on Classifiers

GJT Results

The descriptive statistics for the GJT results on classifier use, including group means and standard deviations, are displayed in Table 20. The group means are also plotted on the graph in Figure 6. As shown, both feedback groups outperformed the control groups on both posttests. Pretest scores were subjected to a one-way ANOVA, which showed that there was no significant difference between the three conditions, F (2, 78) = 2.1, p

= .13. This suggests that any difference between the two feedback groups and the control group did not result from their difference at the time of pretesting.

Table 20. Classifiers: Descriptive statistics on GJT scores

Condition	n	Pretest		Posttest 1		Posttest 2	
		M	SD	M	SD	M	SD
Implicit	28	6.00	1.12	9.23	2.39	9.20	2.51
Explicit	29	5.41	1.34	10.72	2.61	9.86	2.32
Control	21	6.02	1.22	6.57	1.72	6.29	1.82

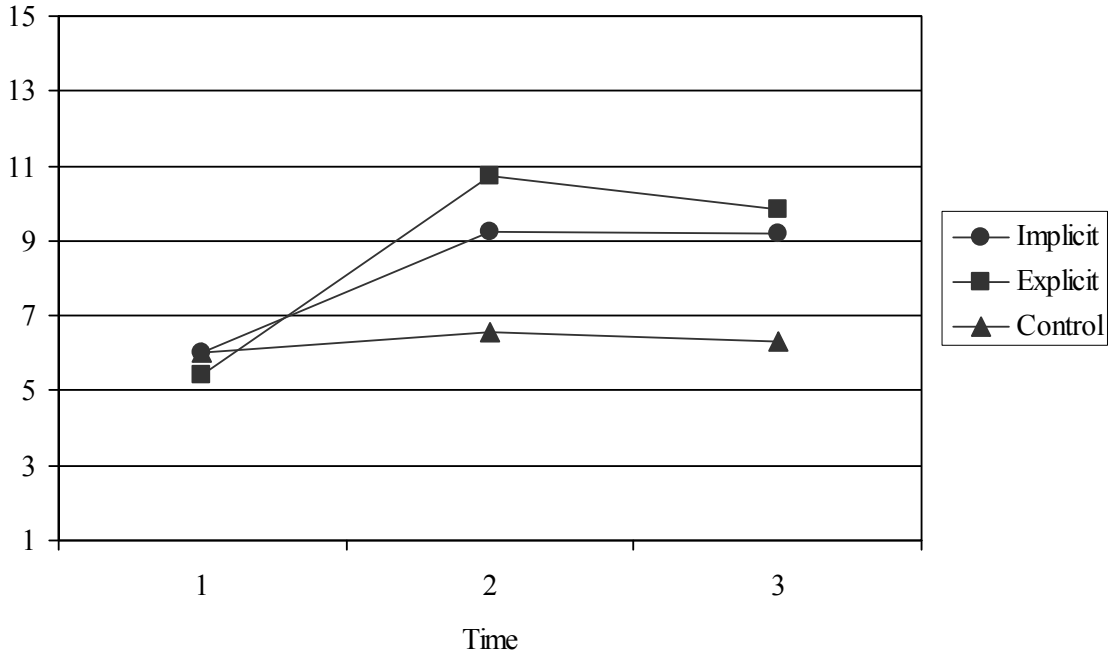


Figure 6. Classifiers: GJT score changes

A mixed design repeated measure ANOVA was conducted to determine if learners' performance scores on classifier use were mediated by feedback type, proficiency, and

timing of testing. Results as presented in Table 21 reveal that there was a significant effect for time, F (2, 143) = 21.57, $p < .05$, and for time* feedback interaction, F (4, 143) = 21.57, $p < .05$. The interaction between time and proficiency was marginally significant.

Between-group results showed a main effect for feedback, F (2, 72) = 20.83, $p < .05$, and for proficiency, F (1, 72) = 31.32, $p < .05$. To locate the source of the differences, post hoc pairwise comparisons were conducted on the gain scores of the subgroups formed by feedback type and proficiency. Descriptive statistics on the gain scores including means and standard deviations over time are displayed in Table 22 (see Appendix D for the descriptive statistics related with the raw scores of different proficiency levels).

As Table 23 shows, both feedback groups outperformed the control groups at both proficiency levels and the effects were well maintained: The effect sizes related to the delayed posttests did not appear to be substantially smaller than the effect sizes related to the immediate posttests. At the low proficiency level, there was a larger effect for explicit feedback than for implicit feedback, but at the high proficiency level, there was no significant difference between the two types of feedback.

EI Test Results

Descriptive statistics related to learners' performance on classifier use as reflected on the elicited imitation test, including means and standard deviations over time, are displayed in Table 24. The means of the three groups are also plotted graphically in Figure 7. It is evident that all three groups improved over time in their performance scores on the EI test and that the two experiment groups appeared to have made greater improvement than the control group.

Table 21. Classifiers: ANOVA results related to GJT scores

Source	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*
Within-Group Results					
Time	420.33	2	210.17	102.87	.00
Time × Feedback	176.25	4	44.06	21.57	.00
Time × Proficiency	11.27	2	5.64	21.76	.06
Time × Feedback × Proficiency	8.09	4	2.02	.99	.42
Between-Group Results					
Feedback	77.04	2	38.52	20.83	.00
Proficiency	57.92	1	57.92	31.32	.00
Feedback × Proficiency	.28	2	.14	.075	.93

Table 22. Classifiers: Descriptive statistics on GJT gain scores

Proficiency	*Group*	*n*	*Gains at Posttest 1*		*Gains at Posttest 2*	
			Mean	SD	Mean	SD
Low	Implicit	14	2.14	1.62	2.50	1.65
	Explicit	15	5.03	3.08	4.21	2.22
	Control	10	0.01	0.74	0.20	1.70
High	Implicit	14	4.32	2.15	3.89	2.61
	Explicit	14	5.61	2.59	4.71	2.62
	Control	11	0.96	1.94	0.32	1.49

Table 23. Classifiers: Post hoc contrasts related to GJT scores

Low Proficiency				*High Proficiency*			
Posttest 1		Posttest 2		Posttest 1		Posttest 2	
Contrasts	ES	Contrasts	ES	Contrasts	ES	Contrasts	ES
I—C*	1.53	I—C*	1.37	I—C*	1.63	I—C*	1.63
E—C*	2.01	E—C*	1.99	E—C*	1.99	E—C*	2.00
E—I*	1.16	E—I*	.88	E—I	0.50	E—I	0.31

Note. ES = effect size; I = implicit; E = explicit; C = control

* $p < .05$

Table 24. Classifiers: Descriptive statistics on EI test scores

Condition	n	Pretest		Posttest 1		Posttest 2	
		M	SD	M	SD	M	SD
Implicit	28	3.14	2.06	7.77	2.88	7.23	2.99
Explicit	29	2.90	2.20	9.33	3.13	8.50	3.50
Control	21	3.86	2.44	5.17	2.53	5.76	2.77

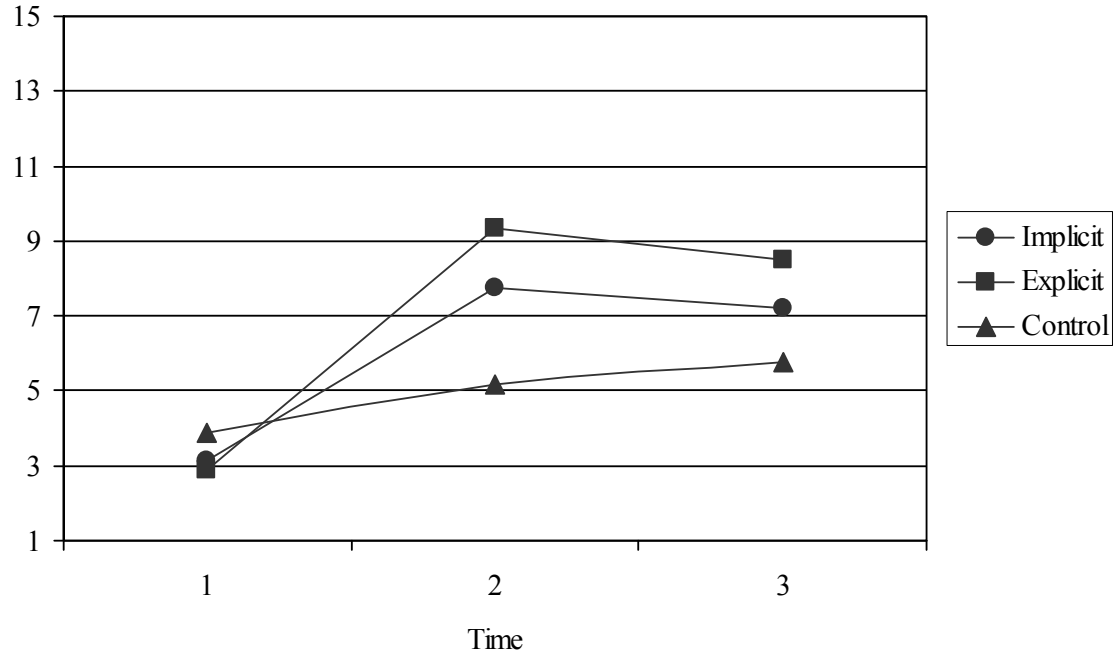


Figure 7. Classifiers: EI score changes

In order to determine if score variation is mediated by feedback type, proficiency, and time, a mixed design repeated measure analysis was performed. The within-group variable is time, and the two between-group variables are feedback type and proficiency. Before the mixed design ANOVA was conducted, a one-way ANOVA was performed on the pretest scores of the three groups, and no significant difference was found, F (2, 77) = 1.19, p = .31.

The mixed ANOVA (Table 25) showed that significant effects were found for time F (2, 143) = 151.02, $p < .05$, for time* feedback interaction, F (4, 143) = 15.89, $p < .05$, for feedback, F (2, 72) = 622.42, $p < .05$, and for proficiency, F (1, 72) = 6.34, $p < .05$. Post hoc group comparisons were conducted on the gain scores of the three involved groups at each proficiency level to locate the source of differences. The descriptive statistics related to the gain scores appear in Table 26 including pre-post change scores and standard deviations (see Appendix D for the descriptive statistics related with the raw scores of different proficiency levels). Table 27 displays the results generated by the post hoc analyses including group contrasts and the corresponding effect sizes. As shown, both the implicit and explicit groups performed significantly better than the control group on both posttests at the low proficiency level; at the more advanced level, the explicit group outperformed the control group on both posttests, but the implicit group only outperformed the control group on the immediate posttest. Explicit feedback worked better than implicit feedback for low-proficiency learners at the time of posttest 1 but the difference did not sustain. No difference was found between the two corrective moves at the high proficiency level.

Table 25. Classifiers: ANOVA results related to GJT scores

Source	*Sum of Squares*	*df*	*Mean Square*	*F*	*p*
Within-Group Results					
Time	803.52	2	401.76	151.02	.00
Time × Feedback	169.15	4	42.29	15.89	.00
Time × Proficiency	2.69	2	1.34	.51	.06
Time × Feedback × Proficiency	1.83	4	.46	.17	.95
Between-Group Results					
Feedback	54.80	2	2688.32	622.42	.00
Proficiency	129.06	1	129.06	6.34	.00
Feedback × Proficiency	2.51	2	1.26	.29	.75

Table 26. Classifiers: Descriptive statistics related to gain scores on EI tests

Proficiency	*Group*	*n*	*Gains at Posttest 1*		*Gains at Posttest 2*	
			Mean	SD	Mean	SD
Low	Implicit	14	4.61	2.00	4.11	2.36
	Explicit	15	6.57	2.37	5.29	3.15
	Control	10	1.55	1.30	1.70	1.21
High	Implicit	14	4.64	2.74	4.07	2.75
	Explicit	14	6.29	2.56	5.71	2.53
	Control	11	1.09	1.41	2.09	1.98

Table 27. Classifiers: Post hoc contrasts related to EI test scores

Low Proficiency				*High Proficiency*			
Posttest 1		Posttest 2		Posttest 1		Posttest 2	
Contrasts	ES	Contrasts	ES	Contrasts	ES	Contrasts	ES
I—C*	1.74	I—C*	1.22	I—C*	1.57	I—C	0.81
E—C*	2.48	E—C*	1.41	E—C*	2.43	E—C*	1.57
E—I*	0.89	E—I	0.42	E—I	0.62	E—I	0.62

Note. ES = effect size; I = implicit; E = explicit; C = control

* $p < .05$

Summary of the Results on Classifiers

The second research question asks whether implicit feedback and explicit feedback have different effects on learners at different proficiency levels in their learning of Chinese classifiers. ANOVAs and post hoc analyses generated the following findings:

(1) Both feedback types benefited the learning of the target structure at both proficiency levels and the effects were maintained. Despite the fact that the mean difference between the high-implicit group and the high-control group did not reach significance on the delayed EI posttest, the related effect size was large. Therefore, a claim can be made for the superiority of implicit feedback to no feedback.

(2) Neither feedback type seemed to have differential effects on learners of the two proficiency levels. Examination of the effect sizes related to individual feedback type across proficiency levels showed that neither feedback type impacted the two proficiency groups differently.

(3) As to which feedback type is more effective, it was found that overall explicit feedback showed larger effects than implicit feedback at the low proficiency level, but high-proficiency learners benefited equally from the two types of feedback.

The Perfective *–le* vs. Classifiers

The third research question concerns whether the choice of target structure mediates the effects of feedback. Because the present study includes two feedback types and two proficiency levels, this research question involves multiple interacting dimensions regarding the relationship between feedback type, proficiency, and structure type. First of all, in terms of the effectiveness of implicit feedback, low-proficiency learners did not benefit from this type of feedback in learning the aspect marker but they did in learning

classifiers. Implicit feedback also facilitated the learning of the aspect marker at the high proficiency level. Moreover, in the learning of the aspect marker, the effects of implicit feedback improved over time for high-proficiency learners, but the same pattern was not found for classifier learning. As to the effects of explicit feedback, it worked for both the perfective *–le* and classifiers and for learners of both proficiency levels. There was no evidence to show its superior effects for either structure or either proficiency level. In terms of which type of feedback is more facilitative of learners' interlanguage development, it was found that for both target structures, low-level learners benefited more from explicit feedback than implicit feedback; this was more so on immediate posttests and when treatment effects were measured by using the GJT test.

To determine whether the overall effects of feedback were different for the two target structures, effect sizes were calculated for the two feedback groups (as compared with the control group) on the GJT and EI test scores. The results, which appear in Table 28, show that the mean effect size (averaged across feedback types and test formats) associated with classifiers is larger than that associated with the perfective *–le*, indicating that overall, feedback tend to be more effective for the learning of the former.

To identify if the effects of feedback on the two target structures are reflected differently on GJTs and EI tests, mean effect sizes were calculated for the two feedback types for each structure on the results related to both posttests. The results, which are presented in Table 29, show that in general, the effect sizes associated with EI test results are larger than those associated with GJT results, regardless of target structure and timing of testing.

Table 28. Effect sizes associated with perfective *–le* and classifiers

		Perfective *-le*		Classifiers	
Test	Feedback	Posttest 1	Posttest 2	Posttest 1	Posttest 2
GJT	Implicit	0.43	0.67	1.39	1.50
	Explicit	2.27	1.03	2.01	1.00
EI	Implicit	1.41	1.14	1.67	2.03
	Explicit	2.06	1.42	2.50	1.48
Mean effect size		1.54	1.07	1.89	1.50

Table 29. Effects of feedback shown on different Tests

	-le		Classifier	
Test	Immediate	Delayed	Immediate	Delayed
EI	1.73	1.28	2.85	1.75
GJT	1.35	0.85	1.70	1.25

Results on Language Analytic Ability and Working Memory

To answer research questions 4 and 5, which ask about the relationship between the two aptitude components, feedback type, and the choice of target structure, correlation analyses were performed on the data contributed by the learners in their 4th semester of study. More specifically, it is the data produced by the 30 learners assigned to the two experiment groups (15 for each) that were analyzed because the purpose was to examine the extent to which the gains under the two feedback conditions correlated with the aptitude components. The data related to the control group are therefore irrelevant. Also,

as previously mentioned, only the data from learners in their 4th semester of study were analyzed to ensure homogeneity of learners' background in terms of the amount of prior instruction.

Prior to the correlation analyses, descriptive statistics were calculated on the scores of language analytic ability and working memory and on the gain scores of the two feedback groups. The full score of the language analytic test is 45. Learners in the implicit condition scored an average of 26.43 and those in the explicit condition averaged 23.20 (Table 30). Learners' working memory score consists of three components: plausibility judgment, reaction time, and recall (of sentence final words). The raw mean scores for the three components and the standard deviations are presented in Table 31. Following Leeser (2007), the raw scores were transformed into *z* scores. The composite WM score for each participant is the average of the *z* scores associated with the three components. Table 32 shows the GJT and EI gain scores and standard deviations by feedback type and target structure at the time of posttest 1 and posttest 2.

Table 30. Scores of language analytic ability

Feedback	n	Mean	SD
Implicit	15	26.43	7.96
Explicit	15	23.20	4.52

Table 31. Raw scores of working memory

Feedback	Plausibility Judgment		Reaction Time (ms.)		Recall Accuracy	
	Mean	SD	Mean	SD	Mean	SD
Implicit	64.00	3.78	3683.23	417.38	53.29	10.69
Explicit	63.47	4.12	3816.92	650.62	50.00	9.61

Pearson's correlation analyses were performed on the scores of language analytic ability, working memory, and the gain scores of the feedback groups. The results are displayed in Table 33, which show that significant correlations existed between language analytic ability and the delayed effects of explicit feedback in the learning of the perfective *–le*, $r = 0.67$, $p = .01$ (GJT); $r = 0.55$, $p = .04$ (EI). Working memory was not significantly related to any gain score as far as the learning of the aspect marker is concerned.

Table 32. Descriptive statistics for 4th semester learners: Gain scores

Perfective -le					
		Posttest 1		Posttest 2	
Feedback	Test	Mean	SD	Mean	SD
Implicit	GJT	2.14	2.56	3.32	2.33
	EI	4.50	2.62	3.25	2.77
Explicit	GJT	6.29	1.79	3.82	3.10
	EI	6.86	2.50	4.14	2.27

Table 32 (Cont'd)

		Classifiers			
Implicit	GJT	3.32	1.99	2.96	1.65
	EI	5.00	2.71	3.64	2.55
Explicit	GJT	5.31	2.91	4.57	2.05
	EI	6.28	2.54	5.63	2.86

In terms of classifier learning, there was a significant correlation between language analytic ability and the delayed effects of implicit feedback (GJT scores), $r = 0.69$, $p = .01$; a significant correlation was also found between working memory and the delayed effects of explicit feedback (GJT), $r = 0.57$, $p = .03$. In the implicit condition, language analytic ability was found to be significantly related to working memory, $r = 0.61$, $p = .02$, raising the question of whether there was an overlap between the two constructs and how much variance language analytic ability accounts for in the delayed effects of implicit feedback in the learning of classifiers. Partial correlation analyses were therefore conducted to explore the unique contribution of either construct when one of them was held constant. It was found that when working memory was controlled for, language analytic ability continued to be significantly correlated with the gain scores of the implicit group on the GJT test at the time of posttest 2, $r = 0.67$, $p = .01$. The result suggests that language analytic ability was solely responsible for a significant portion of the variance in the treatment effects after working memory was partialed out. Working memory, however, did not significantly correlate with the effects of implicit feedback in the learning of classifiers when language analytic ability was held constant.

Table 33. Feedback, aptitude, and the target structure: Correlation results

Feedback	Test Type	Posttest	LAA (*r*)	*p*	WM (*r*)	*p*
			Perfective -le			
Implicit	GJT	1	0.39	0.17	0.19	0.50
		2	0.24	0.41	0.10	0.73
	EI	1	0.31	0.28	0.18	0.54
		2	0.20	0.49	-0.14	0.63
Explicit	GJT	1	0.37	0.19	0.15	0.60
		2	0.67	0.01*	-0.51	0.07
	EI	1	0.14	0.63	0.41	0.15
		2	0.55	0.04*	-0.47	0.09
			Classifiers			
Implicit	GJT	1	0.38	0.19	-0.01	0.98
		2	0.69	0.01*	0.29	0.32
	EI	1	0.31	0.28	0.30	0.29
		2	0.39	0.17	0.20	0.49
Explicit	GJT	1	-0.12	0.66	0.15	0.59
		2	-0.09	0.75	0.57	0.03*
	EI	1	-0.32	0.25	0.12	0.68
		2	-0.10	0.73	0.37	0.17

Note. * $p < .05$; GJT = grammaticality judgment test; EI = elicited imitation; LAA = language analytic ability; WM = working memory

CHAPTER 5 DISCUSSION

This study sought to ascertain the impact of some learner-internal and learner-external factors on the effectiveness of corrective feedback in second language acquisition. These factors include feedback type, learners' proficiency, the choice of target structure, and learners' individual differences in language analytic ability and working memory. The two feedback types under investigation are implicit feedback in the form of recasts and explicit feedback operationalized as metalinguistic correction. 78 L2 Chinese learners were recruited and divided into two proficiency levels. At each level, learners were assigned to three conditions: implicit, explicit, and control. They were tested on their use of the Chinese perfective *–le* and Chinese classifiers before and after instructional treatment by means of two tests: grammaticality judgment and elicited imitation. Learners' language analytic ability and working memory were tested by using the Words in Sentences subtest of the MLAT and a listening span test respectively.

The following results were obtained. First, there was an interaction between feedback type, proficiency, and the choice of target structure. In the learning of the perfective *–le*, there was a limited effect for implicit feedback at the low proficiency level, but this type of feedback showed large effects high-level learners. Also, implicit feedback showed larger delayed effects in relation to immediate effects at the high proficiency level, indicating that the effects improved over time. Explicit feedback worked for all learners, irrespective of their proficiency. In the learning of classifiers, both feedback types benefited learners at both levels of proficiency. Unlike the results on the perfective *–le*, proficiency was not a mediating factor, that is, neither feedback type had differential effects on learners of the two different proficiency levels. As to which type of feedback is

more effective, it was found that explicit feedback demonstrated superior effects as compared with implicit feedback at the lower proficiency level and it was more so for immediate effects; at the higher proficiency level, there were no significant differences between the two corrective moves. The result was obtained for both target structures, indicating the relative robustness of this finding.

Second, with regard to the interaction between the two aptitude components, feedback, and the choice of target structure, it was found that language analytic ability correlated with the effects of implicit feedback in the learning of classifiers and with the effects of explicit feedback in the learning of the perfective *–le*. Working memory only correlated with the effects of explicit feedback in the learning of classifiers. All significant correlations were related with the delayed effects of feedback but not with immediate effects.

Previous research has obtained valuable findings on the effects of corrective feedback. However, researchers have mostly either examined the effectiveness of one feedback type (e.g., recasts) or compared different types of feedback (e.g., recasts versus metalinguistic feedback or prompts) in terms of their effectiveness per se without investigating factors constraining their effectiveness. The inclusion of multiple feedback types and independent variables makes it possible to take an integrated approach to the efficacy of feedback, as shown in Figure 1 ("An integrated model of corrective feedback"), for which the obtained findings of this study provided empirical evidence. In what follows, the results will be discussed by exploring the joint and separate contribution of different factors to the obtained findings.

The Perfective -le

One of the most striking findings is that implicit feedback in the form of recasts did not have substantial effects for low-proficiency learners in the learning of the perfective –*le*. This finding is attributable to several factors. First, it is partly attributable to the implicit nature of this corrective move. Recasts are implicit and are intended not to overtly draw learners' attention to linguistic forms. Those who have argued for the benefits of recasts (Doughty, 2001; Long, 1996, 2007) argued that this type of feedback constitutes an ideal strategy of focus on form just because of its implicitness: It is non-intrusive, and at the same time, it juxtaposes nontargetlike production with the correct form and affords opportunities for cognitive comparison. However, a number of studies (Ellis et al., 2006; Lyster, 2004; Sheen, 2007a, 2010; Yang & Lyster, 2010) found that recasts did not work in classroom settings where feedback was directed toward multiple learners because they were implicit and because learners might have interpreted recasts as confirmation of the content without realizing the corrective force subsumed in the feedback. This argument is backed up by the fact that studies conducted in laboratory settings, where recasts became more salient, have consistently demonstrated that recasts were effective (e.g., Han, 2002; Lyster & Izquierdo, 2009; Mackey & Philp, 1998). The same trend was also obtained in research syntheses on feedback research (Li, 2010; Mackey & Goo, 2007) which showed that lab-based studies produced larger effects than classroom studies. Instructional setting does not contribute to the implicitness of recasts in this study because it was conducted in the laboratory. However, the implicit nature of recasts might be partly responsible because explicit feedback was effective in the same

condition.

The implicit nature of recasts cannot account for the whole picture because this type of feedback was effective for learners at the same level of proficiency (low) in the learning of the other linguistic structure, Chinese classifiers. An interpretation therefore has to be sought with recourse to the nature of the linguistic target. The perfective *–le*, as previously discussed, is redundant, and non-salient, which may have made it difficult for the learner to notice and benefit from recasts, a type of feedback which is implicit by nature.

A link can be established between the above claim and the mechanism of input processing and learners' strategies in semantic encoding. According to VanPatten's Input Processing Theory (2007), learners are more likely to process non-redundant structures than redundant structures because the former carry more communication load than the latter. By the same token, learners give priority to lexicon as compared with morphosyntax in processing input because lexicon is more meaning-loaded and facilitates or impedes comprehension more than morphosyntax does. Researchers adopting functional approaches (the concept-oriented approach in this case) to SLA (Bardovi-Harlig, 2007) hold that learners make use of a range of linguistic options to express a concept such as contextual information, lexical devices, and grammatical morphemes. Beginners tend to rely more on context and lexicon than on morphosyntax in their L2 production. The non-salient and redundant nature of this structure is also likely to affect the learners' test performance: Learners failed to notice and correct the errors related to the use of *–le* during the GJT test, which is essentially a comprehension-based test; in the production-based EI test, they perceived it less necessary to use it than linguistic features

that were more meaningful.

It follows from the above discussion that when recasts do not work, it may not be entirely because of its implicitness and the instructional setting; one important factor to consider is the nature of the linguistic structure to be learned. It is to be reminded that recasts were not effective in Ellis et al. (2006), Lyster (2004), Sheen (2007a, 2010), or Yang and Lyster (2010)[8]. Indeed, what is common about these studies is that all were conducted in the classroom, but what they also have in common is that the target structures in these studies—English past *–ed*, French gender agreement, and English articles(a/an and the)—are either non-salient or redundant. Another piece of evidence for the role of the target structure in moderating the effects of recasts comes from Ammar & Spada (2006), which was also classroom-based but which, unlike the above mentioned studies, demonstrated that recasts were effective. The effects of recasts in Ammar and Spada's study may be attributable to, among other factors, the fact that the target structure is the English possessive determiners *his/her,* which is meaning distinctive and salient.

The next question to be tackled is why implicit feedback oprationalized as recasts did not work for low-level learners but did for high-level learners in the perfective *–le*. If the implicit nature of recasts and the non-salient, redundant nature of the linguistic feature jointly account for the poor performance on the part of the less proficient learners, it appears difficult to understand why more advanced learners benefited from the same feedback when learning the same structure. It would appear that the only explanation is proficiency. It is possible that compared with low-proficiency learners, high-proficiency learners have more cognitive resources at their discretion such that they were more able

to notice the corrective force of recasts despite their implicitness, and they were more likely to process and produce the target structure despite its non-saliency and redundancy. Philp (2003) comprehensively summarized the factors constraining noticing of recasts, which include, among others, developmental readiness, saliency of linguistic structure, and type of instruction. She found that learners tended not to notice linguistic features that were beyond their level of acquisition. Experimental studies (Ammar & Spada, 2006; Mackey & Philp, 1998) demonstrated that recasts were indeed more beneficial to learners who were more developmentally ready.

It should be noted that developmental readiness is often operationalized as learners' prior knowledge about a target structure in previous research (Ammar & Spada, 2006; Mackey & Philp, 1998). In this study, however, learners' level of linguistic competence refers to their general proficiency, which is measured through a standardized proficiency test. Developmental readiness is most probably consistent with general proficiency but may not always be so. The former involves the learning of a particular linguistic structure, whereas the latter concerns learners' overall linguistic development. Learners with higher proficiency, because their load in processing other competing linguistic stimuli is reduced, likely have more cognitive space freed up and are more cognitively involved in processing the corrective information contained in recasts on the target structure.

The fact that recasts facilitated the learning of the perfective *–le* provides further empirical evidence for the claim that learners, or rather advanced learners, are able to induce grammatical rules through repeated exposure to input (Ellis, 2009; Williams, 1999, 2005). This is encouraging because the target structure is complex and the usage of which requires detailed metalinguistic explanation. Ellis speculated that in most so-called

implicit conditions, learning may not be implicit because repeated practice may prompt learners to engage in "meta-awareness in the form of hypothesis-testing and conscious rule-formation" (p. 8). One might argue that learners may not have engaged in rule-inducing in this case; the resultant learning may be exemplar-based: Learners may have merely memorized target-related cases as separate items. This may not be true because correlation analyses revealed that only one of four gain scores (from two tests at two time points) associated with the high-implicit group—the immediate GJT scores—significantly correlated with working memory. The relationship no longer existed on the delayed GJT test. It would seem that learners may have only used the memory traces associated with the structure immediately after the treatment when taking the GJT test, but it is the induced rule that learners used over a longer term.

Classifiers

Let us now turn to the effects of recasts on classifier learning. The question to be answered is why this type of feedback did not work for low-proficiency learners in learning the perfective *–le* but it did in learning classifiers, and why it has differential effects on high- and low-proficiency learners in learning the perfective *–le* but it is equally effective for both levels of learners when it comes to classifiers. The answer lies in the different attributes of these two structures. As previously discussed, classifiers and the perfective *–le* differ in level of redundancy, degree of saliency, form-meaning mapping, and learnability. Classifiers are obligatory, salient, transparent in form-meaning mapping. The perfective *–le* is to the opposite: redundant, non-salient, and opaque in form-meaning mapping. The perfective *–le* is a post-verbal morpheme, which involves long, complex movement. The interpretation of *–le* has to be made at the sentential level.

The classifier is situated in the DP (determiner phrase) and does not involve complex movement. The interpretation of a classifier is made locally in relation to the noun phrase it is attached to, and in most cases, there is a one-to-one correspondence between classifiers and a category of objects. The use of classifiers is therefore more semantically than syntactically driven. Therefore, to a certain degree, the distinction between classifiers and the aspect marker *–le* also parallels one between morphosyntax and lexicon.

Classifiers are salient and therefore are easy to notice; they involve simple form-meaning mapping and hence are amenable to minimal instruction (DeKeyser, 2005). So despite the implicit nature of recasts and the small amount of information that it contains as compared with metalinguistic information, this type of instruction proved to be effective for even low-level learners; and unlike the findings with the perfective *–le*, high-level learners did not benefit more from recasts than their low-level counterparts in classifier learning. Interaction studies focusing on learners' perception showed that learners' noticing of feedback was indeed constrained by the nature of linguistic targets: learners were more likely to recognize lexical recasts than morphosyntactic recasts (Carpenter, Jeon, Macgregor, & Mackey, 2007; Mackey, Gass, & McDonough, 2000). These studies corroborate the superior effects of recasts on classifiers, a structure which is more of a lexical than a morphosyntactic nature.

One interesting finding pertaining to the interaction between recasts and linguistic targets is that the effects of recasts improved over time in the learning of the perfective *–le* at the high proficiency level, but the same trend was not found for classifier learning: The effect sizes associated with the delayed effects of recasts are not higher than the

effect sizes associated with the immediate gains at the high proficiency level. It is speculated that structures involving complex meaning-form mapping engage deeper cognitive processing and the effects of instruction are therefore more endurable. The improved effects of recasts might also have to do with the implicit nature of the feedback because the effects of explicit feedback did not increase over time. Learning under the explicit condition is less endurable probably because the availability of external assistance reduces the need for deep cognitive processing and the obtained knowledge is less likely to be proceduralized. Following this line of thinking, the following hypotheses might be formulated regarding the sustainability of learning outcome:

(1) Effects related to the learning of complex structures are more persistent than effects related to simple structures.

(2) Effects obtained under implicit conditions are more persistent than effects under explicit conditions.

Certainly, these claims are only true of high-proficiency learners in this study and are at best hypothetical. Li's meta-analysis (2010) also found a larger long-term effect for implicit feedback. However, his study did not take into consideration the impact of proficiency and the choice of target structure due to the lack of research on the two variables. There might exist a complicated relationship between proficiency, learning conditions, and the complexity of the linguistic structure as far as the sustainability of treatment effects is concerned, and the jury is still out. SLA researchers rarely address the distinction between immediate and delayed effects when investigating or discussing a certain type of instruction or interventional treatment. Even in cases where several posttests are included, the related results are only reported rather than interpreted. It is

hoped that the discussion in this regard prompts feedback or SLA researchers to address or investigate the sustainability of treatment effects; after all, proceduralized L2 knowledge as reflected on delayed measures is the ultimate goal of SLA.

Explicit-Implicit Comparison

Following Sheen (2007a, 2010), explicit feedback in this study was operationalized as metalinguistic correction, that is, the provision of the correct form followed by metalinguistic explanation. Thus, the explicit feedback contains positive evidence (the provision of the correct form) as well as negative evidence in the form of rule explanation. The explicit feedback benefited all the learners in this study, irrespective of proficiency and target structure. Also, the magnitude of its effects did not seem to vary across the two proficiency levels and target structures. The working principle of explicit feedback seems obvious: It is salient and increases learners' awareness of the target structure. According to Sheen (2007a), metalinguistic feedback is especially facilitative of L2 acquisition because it develops learners' awareness at the level of both noticing and understanding. The provision of rule explanation, especially in the case of the aspect marker, moved the learner's level of awareness from mere noticing to rule awareness, which is "strongly facilitative of subsequent learning" (Robinson, 2002, p. 226; also see Schmidt, 2001). It should be noted that metalinguistic information takes different forms: It can be brief and only serves a metalinguistic alert (e.g., "Think about your question again" [Loewen & Nabei, 2007]); and it can also be detailed and contain rule explanation (e.g., "The fox. You should use the definite article 'the' because you have already mentioned 'fox'" (Sheen, 2007a)). Exactly how much metalinguistic information should be provided is not known and may depend on the complexity of the target and the amount of metalinguistic

knowledge the learner has about the target prior to the treatment. In this study, the metalinguistic information on the perfective *–le* is detailed and that on classifiers is brief (because classifier use does not involve complex rule explanation). It is speculated that learners may have especially benefited from the detailed rule explanation in addition to the provision of the correct form while learning the perfective *–le*, a complex and non-salient structure.

The finding that explicit but not implicit feedback was effective for the perfective *–le* for low-proficiency learners has two implications. First and foremost, it shows that complex structures are amenable to explicit instruction. This finding deviates from the claims of Krashen (1981, 1994) and Reber (1989) that conscious learning is only effective for some easy and semantically transparent structures, and that complex, semantically opaque structures can only be learned implicitly or unconsciously. In the meantime, it is in line with Hulstijn and de Graaff's argument for the advantage of explicit instruction in the learning of complex linguistic features (1994). Second, it testifies to the importance of selective (focused) attention (Gass, 1997, 2003). Low-level learners have heavy processing load and are faced with a large amount of competing stimuli, the explicit information afforded through metalinguistic correction helps focus their attention on a semantically opaque structure that was very difficult to notice under implicit learning conditions. Gass, Svetics, and Lemelin (2003) argued that in learning complex grammatical rules, "internal devices are insufficient for learning, and focused attention…may be a necessary crutch" (p. 528).

As to which feedback type is more facilitative of L2 development, this study showed that the superiority of explicit feedback to implicit feedback is subject to several caveats,

the discussion of which will have useful pedagogical implications. It was found that while it is true that explicit feedback seemed to be effective regardless of proficiency and target structure, it was not always more effective than implicit feedback (also see DeKeyser, 1993; Loewen & Nabei, 2007). The advantage of explicit feedback was more evident for low-level learners on immediate posttests and the differences were greater on GJTs than on EI tests. These results are somewhat different from what was found in previous research. SLA literature abounds in studies that promote the utility of explicit instruction. Both empirical studies (e.g., Carroll & Swain, 1993; Ellis et al., 2006; Sheen, 2007a, 2010) and research syntheses (Li, 2010; Lyster & Saito, 2010; Norris & Ortega, 2000; Spada & Tomita, 2010) directly and indirectly showed that explicit feedback is unequivocally more effective than implicit feedback.

However, what these studies did not examine is the role of proficiency in moderating the effects of feedback. Ammar and Spada (2006) and Li (2009) are probably the only studies that included proficiency as an independent variable when investigating the differential effects of different types of feedback and they reported similar results. In a classroom setting, Ammar and Spada compared prompts, which included metalinguistic feedback, and recasts, and they found that prompts were more effective than recasts for low-level learners but the two feedback types worked equally well for high-level learners. Li's lab-based study showed that the pre-post gains of the explicit group were larger than the gains of the implicit group. To be noted is the fact that proficiency in Ammar and Spada's study was operationalized as learners' prior knowledge about the target structure, and as enrollment status in Li's study.

To emphasize the importance of considering learners' proficiency in opting for

explicit or implicit feedback is not to favor one instruction type over the other. The point here is to view them in perspective. Given the superior effects of explicit feedback for low-level learners, it seems advisable to employ instruction types that facilitate their awareness of the linguistic structure and that provide them with more external resources. However, where explicit feedback does not lead to more learning than implicit feedback or where they are equally effective, as is the case for high-proficiency learners, implicit feedback would be a better choice. This recommendation is defendable both theoretically and pedagogically.

From a theoretical point of view, according to the Socio-Cultural Theory (Aljaafreh & Lantolf, 1994; Lantolf, 2009; Ohta, 2009), learning occurs through mediation (which is, in this case, interaction) between a novice and an expert in the zone of proximal development (ZPD), which refers to the difference between what a learner can achieve independently and what he/she can achieve with external assistance. Learning should evolve from object-regulation to other-regulation and finally to self-regulation. The purpose of instruction is to provide assistance to the effect that the learner's reliance on assistance is progressively reduced and ultimately the learner becomes autonomous. The ideal condition for learning to happen is one where the learner is offered "just enough assistance…[and] assume[s] increased responsibility for arriving at the appropriate performance" (Aljaafreh & Lantolf, 1994, p.469). Providing too little or too much assistance is both beyond what the ZPD requires for optimal learning outcome. Thus, low-proficiency learners are under-assisted if they are provided with implicit feedback, which contains insufficient information; high-proficiency learners are over-assisted if they are provided with explicit feedback, which is imbued with superfluous information.

Too much assistance impedes the development of the learner's ability to work independently and autonomously (Ohta, 2009, p. 52).

From a pedagogical perspective, implicit feedback in the form of recasts is ideal for form-focused instruction. The essence of form-focused instruction is that learning is optimal in situations where the primary focus is on meaning but linguistic forms are attended to in the meantime. In their seminal synthesis of the literature on recasts, Ellis and Sheen (2006) precisely summarized the interactionist views (Long, 1996, 2007) on the nature of this corrective strategy and its match with form-focused instruction:

> They [recasts] induce a joint focus on form and meaning, thereby encouraging form-function mapping in the context of, and without disturbing, the communicative flow of the interaction. Furthermore, they allow for cognitive comparison of erroneous and target language forms in a context in which the learner is primed to notice the difference. In contrast, explicit forms of correction involve treating language as an object and interrupting the flow of communication and, thus, will not assist form-function mapping. (p. 578)

Lyster and Mori (2006) pointed out that recasts constitute an especially favored feedback type in immersion classes because in addition to serving language learning purposes, they help learners focus on content and communicate about subject matter that is beyond their current linguistic competence. Moreover, one attribute of recasts as an implicit type of feedback is that its effects might be longer-lasting than explicit feedback, at least for high-proficiency learners in the learning of complex structures whose form-meaning mapping is not transparent. As previously discussed, the finding that implicit feedback is better at retaining instructional effects than explicit feedback is likely not anecdotal

because the same pattern was also obtained in Li's meta-analysis on previous empirical research on corrective feedback (2010).

Again, it would be arbitrary and misleading to take an absolute rather than dialectical position when it comes to determining which type of feedback is a better option because both, like any other type of instruction, have their respective advantages and limitations. In light of the findings of this study and the above discussion on the related theoretical underpinnings and pedagogical implications, it is recommended, albeit tentatively, that explicit feedback be provided to low-proficiency learners and implicit feedback to more advanced learners.

The Effects of Target Structure and Testing

In addition to the interactions between feedback type, nature of the target structure, and proficiency, one question this study attempts to answer is whether there is a main effect for the target structure, that is, whether feedback in general has differential impact on the two target structures. It was found that the average effect size associated with the learning of classifiers, regardless of proficiency and feedback type, was greater than that associated with the perfective *–le.* This is not surprising because classifiers are perceptually more salient, syntactically simpler, and semantically more transparent than the perfective *–le*. The information contained in the implicit feedback targeting classifiers is more likely to be perceived and incorporated than that targeting the perfective *–le*. In the case of explicit feedback, because the linguistic derivation of the perfective *–le* is complex and the metalinguistic information is difficult to comprehend, the information is less likely to be internalized even if the information is delivered in an obvious, straightforward manner.

The following excerpt from the exit interview with a participant (from the high-implicit group) indicates how salience might have contributed to the larger effects for classifiers. When asked to comment on the two target structures, she said:

> You can still get your point across without saying -*le*. Maybe that's the reason [why I omitted –*le* in some cases]. It's not saying that it's not important, but that the -*le* structure is something you can easily drop [and] still get your point across. I think measure words are more important. When in China, and you are having a conversation, 'cause as a foreigner, I don't think many people understand you. So giving a measure word is one more clue to what you're talking about. Dropping a measure causes more misunderstanding than dropping -*le*.

The superior effect of feedback for classifiers in relation to the aspect marker is to some extent consistent with what was found in previous literature. Mackey and Goo (2007) found that interaction, an important component of which is feedback, had a larger effect on lexical items than on morphosyntactic items; Yang & Lyster (2010) found that recasts were more effective for irregular past forms than the regular past form. Lexicon and irregular past forms are necessarily more salient and semantically more transparent than morphosyntax and regular past forms. Spada and Tomita (2010) showed that both implicit and explicit instructions showed larger immediate effects for simple structures than for complex structures. Certainly, complexity, salience, and difficulty are not the same and the relationship between them is far from clear and is controversial in current SLA literature. This topic is beyond the scope of the current study. Regardless, one thing seems obvious: the perfective –*le* is a more complex structure than classifiers.

The calculation of mean effect sizes related to the two test formats showed that EI

tests showed larger effects than GJTs. EI tests were intended to measure implicit knowledge as a result of instructional treatment, and GJTs measure learners' explicit knowledge. Previous feedback research also showed a larger effect of feedback as measured by tests tapping into learners' online spontaneous performance than tests reflecting learners' offline, declarative knowledge of the target structure (Ellis et al., 2006; Li, 2010; Loewen & Nabei, 2007; Lyster & Saito, 2010). This would suggest, as Ellis et al. pointed out, that instruction, which in this case is realized through the provision of corrective feedback, does contribute to the development of learners' implicit knowledge. It confirms DeKeyser's (1998; 2007) and N. Ellis's (2008) claim that explicit and implicit knowledge stands in a continuum, and that through practice, conscious, declarative knowledge can be converted into unconscious, procedural knowledge. It counters Krashen's (1981, 1994) and Hultijin's (2002) argument that explicit knowledge is unlikely to become proceduralized.

While it appears true that feedback leads to the acquisition of implicit knowledge, questions need to be answered regarding why gains on EI tests are larger than on GJTs. Does it mean that feedback leads to more implicit knowledge than explicit knowledge? What caused the difference between the two test formats? Answers to these questions are critical to understanding feedback research and the effectiveness of feedback, for the same study might show different, or even conflicting, results depending on how the effects are tested. Without appropriate testing, it would be misleading to discuss the effects of any instructional intervention. The difference between the two test formats might be accounted for from a combination of three perspectives: transfer appropriate processing, amount of previous explicit knowledge, and the typological features of

Mandarin Chinese.

Transfer appropriate processing (TAP) is a cognitive approach of information processing advanced by Morris, Bransford, and Franks (1977) (see Lightbown (2008) for how the TAP applies to SLA). Morris et al., based on empirical evidence, contended that "assumptions about the value of particular types of acquisition activities must be defined relative to the type of activities to be performed at the time of test" (p.531). The basic tenet of this approach is that information or input is better retrieved in tasks that resemble the conditions where the information is received. Thus, if learning occurs in meaning-focused activities where attention is also paid to language, the learning outcome will be best demonstrated on tests that are similar to the learning conditions, that is, taxing the learner's ability to process meaning and form simultaneously. Likewise, if learning occurs in activities that involve discrete item practice where language is learned as an object, the learner will perform better in item-based tasks that primarily involve the processing of linguistic forms. In the case of this study, oral feedback was provided in meaning-based interaction. The EI tests are also oral and the cognitive construct underlying the EI tests is the same as the treatment tasks: on-line, simultaneous processing of meaning and form. The GJTs are written and the primary focus of the tests is on linguistic forms. Therefore, the TAP may partly explain the larger effects on the EI tests than on the GJTs.

The difference between the two tests in terms of treatment effects may also have to do with the greater amount of explicit knowledge learners had at the outset of the study. Re-examining the descriptive statistics displayed in tables 9 and 13 confirmed this speculation: the GJT scores of all three groups (implicit, explicit, and control) are larger

than their EI test scores on the pretests. Therefore, there was a certain ceiling effect for the learners' level of explicit knowledge[9].

The third potential factor relates to the typological features of the target language. Unlike alphabetic languages where phonological systems are linked with orthographic representations, that is, the pronunciation of a word is somewhat predictable from its spelling, the Chinese language has separate writing and speech systems. The association between a character or word and its pronunciation is arbitrary. The GJTs, unlike the EI tests, are written, so learners might have difficulty recognizing the vocabulary words involved in the obligatory contexts for the target structures in test items because these words were presented orally in the treatment tasks. Although Pinyin, the Romanized phonetic representation of Chinese characters, was provided for each single character in each test item, more than 83% of the learners indicated during the exit interview that they either entirely ignored the Pinyin or only occasionally referred to it when they felt the need to. In fact, a few of them even reported that Pinyin was distracting and affected their reading speed. According to one of the instructors of the classes from which the participants were recruited, students were encouraged to use Pinyin as a crutch during the first year of study, but it was discouraged in the second year and thereafter because the Chinese language, after all, is not an alphabetic language and students must learn to decipher written materials without Pinyin. In fact, non-teaching Chinese materials are never accompanied by Pinyin.

One might also argue that learners might have exercised more control over their output during the EI tests than the GJTs, calling into question the validity of the test as a measure of implicit knowledge. However, this is unlikely to be true because when asked

on which test they used more grammar knowledge, 85% percent of the participants indicated that they did so more on the GJT and that they relied more on "hunch" or "feel" during the EI test. One participant, who is from the low-explicit group, commented on his use (or non-use) of grammar rules during the two tests: "I did it more by intuition when I was speaking and I tried to remember rules we learned in class while I was doing the written one. I thought more about rules when I did the written part." Interestingly, one participant (high-implicit) pointed out that he did give more thought to the correctness of a sentence when taking the GJTs, but the extra efforts had a negative effect on his test performance:

> When I was listening, I don't think I was thinking about it as much. Just my instinct said that I need to add -*le* here. I didn't actually process that sentence in my mind. And then when I am reading it, I kind of thought about it more. I second-guessed myself and I wasn't sure, so I didn't do it [make the correction]. So that extra time actually hurt me.

To conclude the discussion on the interface between the effects of feedback and testing, some further comments are in order. First, if the TAP is accepted as one criterion for test validity, it stands to reason to call for the need to match treatment conditions with testing conditions. In other words, what is measured should be what is learned/taught. Thus, it would be appropriate to use tests that involve oral production such as elicited imitation to measure the effects of oral feedback in communicative activities. If the purpose of a study is to investigate the effects of feedback provided in discrete item practice, similar tasks should be used in testing. Second, to accommodate the dissociation between orthographic and phonological systems of non-alphabetic languages, items in

written tests might be presented both orally and visually if the instructional treatment mainly involves oral production and if a written test must be used.

Third, the implicit-explicit distinction in terms of knowledge type might be relative rather than dichotomous and might be subject to multiple factors. Proficiency, for instance, is likely to affect how much implicit knowledge learners have in their discretion. Beginners might have to access their explicit knowledge about the target language most of the time. Therefore, the implicit-explicit distinction may be more applicable to learners at higher stages of their language development. One remedy to this problem might be to include the component of reaction time in EI tests. Obviously, the amount of time a learner takes to respond to linguistic stimuli is an indicator of the automaticity of L2 knowledge: The faster the learner responds, the higher the likelihood that the related knowledge is proceduralized. Another factor that affects the validity of a test of implicit or explicit knowledge is the characteristics or dynamics of the instructional setting. For instance, learners from intensive language programs may have more explicit knowledge than learners from an immersion context where language is taught or learned through the subject matter and where priority is given to fluency rather than on accuracy of language use. Also, how learners acquire their first language literacy may also be relevant. Learners who are accustomed to a meaning-oriented approach to linguistic materials in their first language may transfer the corresponding learning strategy into their second language learning. If that is the case, they are likely to engage in more semantic than syntactic processing, hence affecting their performance on a GJT.

Feedback, Linguistic Structure, and Aptitude Components

A major goal of this study is to determine whether the effects of feedback are related

to learners' individual differences in two aptitude components: language analytic ability as measured by the Words in Sentences subtest of the MLAT and working memory that is assessed through a listening span test. It was found that there was an interaction between feedback type, the choice of target structure, and the two aptitude components. More specifically, language analytic ability correlated significantly with the effects of implicit feedback in the learning of classifiers and with the effects of explicit feedback in the learning of the perfective *–le*; learners' working memory capacity was related only with the effects of explicit feedback in the learning of classifiers. All the significant correlations were found for the delayed effects of feedback.

These findings underscore the importance of exploring aptitude-treatment interaction (Snow, 1987, 1991) and provide further justification for the necessity of taking a componential rather than monolithic approach to aptitude research (Dörnyei & Skehan, 2003; Robinson, 1997, 2002, 2005; Skehan, 2002). Clearly, the idiosyncratic characteristics of each learning condition, defined jointly by the type of feedback and the nature of the linguistic target, set different processing demands on learners' cognitive abilities, hence the resultant dynamic relationships between the two aptitude components and the two feedback types. A related point is that unlike previous feedback research that focused on the two-way feedback-aptitude interaction (e.g., Sagarra, 2007), this study revealed the impact of a third variable, the target structure, on the sensitivity of individual difference factors to the effects of corrective feedback.

Language Analytic Ability

The significant correlation between language analytic ability and the effects of implicit feedback in the learning of classifiers is seemingly surprising because no

metalinguistic information or rule explanation is available in this learning condition. A plausible explanation seems to be that when grammatical explanation is not available, learners with high analytic ability achieved more. They were better-versed than learners with lower analytic ability in extracting and generalizing the syntactic regularities related to classifier use based on the positive and/or negative evidence contained in the provided recasts. However, two questions arise regarding this interpretation. One is whether learners engage in syntactic processing in implicit learning conditions; the other is whether language analytic ability is drawn upon given the fact that classifiers constitute an easy structure and do not involve complicated rule explanation.

To answer the first question, it is necessary to draw on Robinson's work on aptitude-treatment interaction (1997). Robinson found that in the implicit condition, where they were told to simply memorize some examples without being provided with any rule explanation, learners with high aptitude claimed to have actively looked for and were able to verbalize rules. Therefore, learners with high analytic ability are more likely to be aware of linguistic problems and engage in hypothesis testing about the target structure. To answer the second question, it is useful to refer to Polio's work on the acquisition of Chinese classifiers (1994). Polio investigated how speakers of L1 Japanese (a classifier language) and of L1 English (a non-classifier language) used Chinese classifiers. Despite Polio's claim that the nonnative speakers did not seem to have difficulty using classifiers in obligatory contexts, her data revealed that all the classifier omission errors were committed by L1 English speakers and none was made by the Japanese speakers. Thus, it can be inferred that although the classifier is not a complicated structure, it does pose problems for speakers of languages where this structure is absent. Therefore, it is

reasonable to speculate that language analytic ability did play a role in the implicit condition in the learning of classifiers because the structure was a challenge, especially when the metalinguistic information was unavailable.

While language analytic ability was related to the learning of classifiers in the implicit condition, it was not the case with the learning of the aspect marker in the same condition. Lack of awareness and the nature of the target structure might be jointly accountable for this result. On one hand, the implicitness of the learning condition and the non-saliency of the perfective *–le* may have made the feedback and the target structure difficult to notice. Learners in this condition may therefore not have engaged in syntactic processing by utilizing their analytic ability. On the other hand and more importantly, because of the complexity and difficulty involved in learning the perfective *–le*, learners were not able to make inductions or generalizations on the usage of the structure through mere reliance on their internal resources (in this case, language analytic ability), even if they were aware of the problems in their L2 production. In other words, it is beyond all learners' cognitive capacity to conduct syntactic processing of this linguistic structure with recourse only to the limited external assistance in the form of implicit feedback.

Consideration of the nature of the linguistic target also helps explain the conflicting findings in previous studies (Sheen, 2007a; Trofimovich et al., 2007). Whereas Sheen failed to find a significant correlation between the effects of recasts and language analytic ability, Trofimovich et al. did. In Sheen's study, the target structure is the English articles (*the* and *a/an*), a non-salient, complex structure; in the study by Trofimivoch et al., the linguistic structure is the English possessive determiners (*his/her*), a salient, simple

structure. Clearly it was difficult for learners to notice the English articles in an implicit condition, but even if there was a high level of noticing, learners were likely unable to extract rules about the articles by using their analytic ability. However, by taking advantage of their analytic ability and the input from recasts, it is possible for learners to solve problems related to the possessive determiners, a relatively easy structure; hence the significant correlation in Trofimovich et al.'s study. Another related study, which is cited by Robinson (2005), is by Robinson and Yamaguchi (1999), who found that "there were nonsignificant correlations of learning of relative clauses [a complex structure] during task-based interaction (supplemented by targeted recasts) and the grammatical sensitivity aptitude subtest" (p.56). Taken together, these studies point to the possibility that non-significant correlations between language analytic ability and the learning gains under implicit conditions are attributable to the complex, difficult nature of the linguistic structure, which may be beyond learners' processing capacity.

Language analytic ability has been consistently found to correlate with learning under explicit conditions where rule explanation or metalinguistic information is available (Robinson, 1997; Erlam, 2005; Sheen, 2007a). This study is no exception: It significantly correlated with the effects of explicit feedback in the learning of the perfective *–le*. The mechanism through which this aptitude component works under this condition seems simple: Learners with superior analytic ability are better at discovering patterns in the input or processing and applying the knowledge "assimilated from external sources" (Roehr & Ganem-Gutierrez, 2009, p. 167). In the case of the perfective *–le*, the explicit feedback enhances learners' awareness of the linguistic structure and provides metalinguistic information about a very complex linguistic structure; learners likely

engaged in active processing of the information by applying their previous knowledge and language analytic skills.

But why is the learning of classifiers not related to learners' language analytic ability in the explicit condition? Once again, this potentially related with the nature of the target structure. Since the classifier is a relatively transparent structure and the provided metalinguistic information is easy to process and internalize, learners' individual differences in language analytic ability was likely not drawn upon and therefore did not impact learning in this condition. Therefore, in the case of classifiers, it is, as it were, the availability of the metalinguistic information that led to the marginization of the role of language analytic ability.

Working Memory

Variation in learners' working memory capacity was found to be related to the effects of explicit feedback in the learning of classifiers. Surrounding this finding, two important questions need to be answered: (a) What is the mechanism through which working memory functions in relation to the treatment effects? (b) Why is this aptitude component not sensitive to the other learning conditions formed by feedback type and choice of linguistic structure?

To answer the questions, it is necessary to revisit the architecture of working memory and the functions of its components. As previously discussed, working memory is characterized by simultaneous storage and manipulation of information. According to Baddeley's componential model (Baddeley, 1986, 2000, 2006, 2007; Baddeley & Logie, 1999), working memory is composed of a central executive and three slave systems: a phonological loop, a visuospatial sketchpad, and an episodic buffer. The central executive

is responsible for information processing and integration and coordination between the three subcomponents. The phonological loop encodes, temporarily stores, and rehearses incoming auditory stimuli; the visuospatial sketchpad is a short-term store of visual and spatial information; and the episodic buffer activates information in long-term memory, constructs integrated representations, and encodes them in long-term memory as schema.

The processing demands of classifier learning through external assistance in the form of metalinguistic correction seem a perfect match to the mechanism of working memory. When the learner's attention was brought to the target structure through the provided feedback, the learner encoded the auditory stimuli (sound representations about a classifier as well as the metalinguistic information) in the phonological loop, matching the phonological codes with existing codes (e.g., sounds and tones the learner previously learned) archived in long-term memory. This was followed by subvocal rehearsal of the stored information. The central executive maintained the information in focal attention and processed it for storage in long-term memory through the episodic buffer. The cognitive processing may have taken place by matching a certain classifier with a noun and analyzing the metalinguistic information; it may also have involved the inhibition of other classifiers in the repertoire, which likely competed for the limited capacity of working memory. Evidently, classifier learning in the explicit condition draws heavily on the learner's ability to store and process the provided information, which led to the significant correlation between working memory and the treatment effects.

The significant correlation found in the explicit condition between working memory and classifier learning has to do with consciousness, a defining feature of explicit learning conditions. Almost all models of working memory, such as the Multiple-

Component Model (Baddeley & Logie, 1999), and the Executive Attention Model (Engle, 2002), the Embedded-Process Model (Cowan, 1999), acknowledge the role of consciousness and attention control (also see Dehn, 2008; Paradis, 2009). Baddeley pointed out that "as has become increasingly obvious over the years, conscious awareness appears to be closely related to the executive control, and hence to the operation of working memory" (2007, p. 302). Cowan argued that awareness "increases the number of features encoded, and…allows new episodic representations to be available for explicit recall" (1999, p. 62). Engle even stated that working memory is not about short-term span; rather, it is about the ability to focus attention on relevant information and inhibit irrelevant information. Indeed, in this study, learners' ability to focus their attention on the information contained in feedback and at the same time resist distracting information may be critical to the development of their knowledge about classifier use.

The association between consciousness and working memory also explains why this cognitive construct was not related with the effects of implicit feedback (in learning either target structure). Implicit feedback was not intended to make the learner conscious/aware of the target structures, and thus working memory may not have been implicated in this learning condition. Schmidt (1990) stated that implicit/unconscious processes are not susceptible to working memory capacity. Similarly, Ellis (2009) pointed out that implicit learning does not implicate central attentional resources; explicit learning, in contrast, relies heavily on working memory because it involves conscious memorization of facts.

There is discrepancy in previous research with regard to the link between working memory and the effects of implicit feedback. Trofimovich et al. (2007) did not find any

association between the effectiveness of recasts and measures of working and phonological memory. Sagarra (2007), however, found such association. It is not clear what caused such a discrepancy, but some speculations can be made based on the research methods of these two studies. In both studies, recasts were provided via the computer in discrete item practice, which made "the corrective nature of recasts more salient". The learning conditions are therefore explicit, which likely contributed to learners' consciousness of the learning tasks, and hence the significant correlation in Sagarra's study. The nonsignificant correlation in Trofimovich et al.'s study might be due to two factors: (a) The test items were identical to treatment items, all of which involved picture description, and (b) both posttests were immediate (which is also pointed out by Sagarra). The identicalness between testing and treatment, the additional assistance from picture cues, and immediacy of posttests (which will be discussed below) may have minimized the role of individual differences in learners' working memory.

A final explanation needs to be explored related to the absence of a connection between working memory and the effectiveness of explicit feedback in the learning of the perfective *–le*. It would seem that unlike classifier learning, which involves simultaneous storage and processing of information, the perfective *–le* is purely rule-based. While rules must be memorized for learning to occur, the greater or lesser capacity of the short-term store is not as influential as it is in learning more exemplar-based structures such as the classifier. Therefore, the learning of this structure likely does not tax the storage function of working memory. Certainly, working memory is also responsible for information processing, which is controlled by the central executive. However, the central executive is often considered a hub of attention control (such as selective attention, attention

switching, resource allocation, etc.) and is rarely credited with syntactic processing. Of course, the central executive must be responsible for a certain amount of syntactic processing such as, in the case of classifier learning, quickly encoding and decoding a permutation of classifier use (numeral + classifier + noun) when the related feedback is provided. Nevertheless, the learning of complex morphosyntax might rely more on language analytic ability than working memory, a construct that indexes the efficiency in temporarily holding and processing information; this explains why it is the former, not the latter, that correlates with the learning of the aspect marker in the explicit condition. Detailed explanation on the relationship between these two aptitude components is provided in the next section.

Language Analytic Ability vs. Working Memory

One interesting finding of this study is that in the implicit condition, language analytic ability correlated with working memory but such a correlation was not found for the explicit condition. Recall that the analyses related to the two cognitive factors were based on the data contributed by learners' in their 4th semester of study to control for variation in the amount of instruction. A correlation analysis performed on the whole dataset that involved all participants' scores on the two variables indicated that there was a significant correlation between them, $r = 0.3$, $p = .01$. This suggests that the correlation between these constructs found in the implicit condition did not happen by chance. It also suggests that there is a certain overlap between the two, and yet they are separate constructs as the correlation is small. But how are they different and related? How do they contribute differently to L2 learning?

The two aptitude components differ along the following dimensions. First and

foremost, they are measures of different cognitive abilities. Language analytic ability refers to the learner's sensitivity to linguistic regularities and ability to identify linguistic patterns. Working memory is an indicator of the learner's capacity in information storage and online cognitive processing. Therefore, the former is more likely to be drawn upon in tasks that place heavy demands on syntactic processing (either online or offline) as in the learning of complex linguistic rules such as the Chinese perfective -*le*; the latter is most useful in the learning of more data-driven, exemplar-based linguistic items. Second, related to the first point, they interact differently with different learning conditions. For instance, as this study demonstrates, the role of working memory is probably more obvious in explicit learning conditions; language analytic ability is useful in implicit conditions in the learning of simple rules and in explicit conditions in the learning of complex rules. Third, language analytic ability is domain-specific whereas working memory is domain-general. It is obvious that language analytic ability is only implicated in language acquisition, or rather, adult second language acquisition (because first language acquisition and child second language acquisition are not dependent upon language analytic ability (DeKeyser, 2000; Harley & Hart, 1997; Sasaki, 1996)). Working memory, however, is "a general-purpose system that can perform multiple functions" (Dehn, 2008, p. 41) and has been found to be related with many academic skills such as math reasoning, science, and so on (Gathercole & Aloway, 2008; McGrew & Woodcock, 2001).

The relationship between language analytic ability and working memory is definitely not one between oranges and apples—they work in concord in facilitating second language development. The finding that working memory correlated with language

analytic ability, a core aptitude component of the MLAT, provides further evidence that working memory is an aptitude component. Previous research also found working memory to be related with the L2 learners' overall performance in aptitude and aptitude components as measured by the MLAT or similar test batteries (Robinson, 2002; Safar & Kormos, 2008)[10]. In fact, working memory, a measure of the ability to handle the juggling of information storage and processing is considered an ideal substitute for the Paired Associates subtest of the MLAT, which measures learners' rote memory ability, especially in form-focused instruction where linguistic forms are attended to during meaningful communication.

Sawyer and Ranta pointed out that working memory may "serve as an arena in which the effects of other components of aptitude are integrated" (2001, p. 342). For example, phonetic coding, an aptitude component measured by the MLAT, is critical to the functioning of the phonological loop of working memory, which relies heavily on the phonemic awareness and efficiency in phonological encoding and decoding. Language analytic ability affects the speed and efficiency of processing of the central executive such that high analytic ability frees up space for the storage subsystems, and deficits in analytic ability slows down processing and causes working memory overload. The storage component of working memory, on the other hand, might impact the learner's capacity in rule identification and application. Temporary maintenance of information affects comprehension of input which in turn influences the accuracy of language analysis. Another bond between language analytic ability and working memory is noticing. Robinson (1997) found that in the implicit condition of his study, there was a strong relationship between grammatical sensitivity (language analytic ability) and

awareness. In Mackey et al.'s study (2002), learners with high working memory reported more noticing of the target structure, confirming the link between working memory and consciousness. And noticing, or conscious awareness, is a defining attribute of the central executive of working memory.

Aptitude and Testing

It is interesting that the feedback-aptitude interactions found in this study are subject to the timing of testing and type of measure: all correlations are related to delayed effects and are demonstrated on different test formats. The correlation of aptitude measures with the delayed effects of feedback is consistent with the findings of previous research (Ando et al., 1992; Mackey et al., 2002; Trofimovich et al., 2007). Even in Sheen's study (2007a) where language analytic ability correlated significantly with both the immediate and delayed effects of feedback, the correlations appeared stronger on delayed posttests. It is not clear why it is so, but researchers have made some reasonable speculations. Robinson (2002), noting that "immediate posttest performance shows very little relationship to measures of IDs [individual differences]", speculated that "learning continued as a consequence of the immediate delayed transfer test experiences…[and] IDs in relevant abilities contributed to the capacity to build on initial exposure during training, and continue to learn during the posttests" (p. 204). Trofimovich et al. explained that the role of language analytic ability may be greater when corrective feedback is no longer available. Mackey et al. hypothesized that learners with greater working memory capacity may have "gleaned more data to process and consolidated this over time", in comparison with learners with smaller working capacity who "could not 'hold on' to data with great accuracy" (p. 204).

In essence, these speculations come down to two themes. One is that the immediacy of testing may have leveled out the impact of variation in aptitude. The other is that during the interval between instructional treatments and delayed tests, learners conduct off-line processing of the data they obtained during the treatments. In addition to the two possible explanations, another factor that comes into play might be the research setting. This study, as well as the ones by Mackey et al. and Trofimovich et al., is conducted in the laboratory and targets only one linguistic structure, which makes the treatments well implemented. The lab setting may have partly made the immediate effects of individual difference variables less obvious. This claim may find indirect support from Sheen's studies (2007a; 2007b) where correlations between language analytic ability and the effectiveness of feedback were found at the time of immediate posting. Both Sheen's studies were conducted in classroom settings where there is more distraction than in laboratory settings and where the effect of individual differences is more likely to surface immediately. However, this speculation is tentative and is subject to further empirical verification.

Another interesting finding related to testing effects on aptitude-feedback correlations is that the correlations are constrained by test formats. First, with respect to classifier learning, language analytic ability correlated with learners' performance on the grammaticality judgment test but not with elicited imitation test scores (the result relates to implicit feedback). However, the results related to the perfective *–le* showed that language analytic ability correlated with the gain scores on both tests (the result relates to the explicit feedback). It would appear easy to understand the correlation between language analytic ability with GJT scores because it was measured with a written test that

taps into metalinguistic knowledge, which was obviously utilized in the GJT tests. Somewhat unexpected is the fact that language analytic ability was also drawn upon in the EI test for the aspect marker. It is possible that learners had proceduralized the knowledge obtained through the application of their analytic ability under the assistance of the metalinguistic information afforded in the explicit feedback. The proceduralized knowledge became accessible during the EI test.

Working memory scores were correlated with the GJT scores that are related to classifier learning in the explicit condition but not with the EI test scores. The non-significant correlation between working memory and EI test results has two implications. First, it provides evidence that the EI test is not a measure of memory. An EI test item involves comprehending a verbally presented sentence stimulus, judging the semantic applicability to the test taker per se, and repeating it in a correct way. It is tempting to consider the test as being related with the test taker's memory capacity. The fact that the test results were not associated with working memory scores further demonstrates the validity of the test as a measure of the learner's knowledge about the target structure in question; working memory, necessary as it may be in completing the test, did not have substantial impact on the test scores.

Second, accuracy in online comprehension and oral production of the target structure was not reliant on working memory. This is because the EI test measures implicit, automatized knowledge, which is accessed through long-term memory instead of working or short-term memory. Working memory involves conscious processing, so if skills or knowledge are fully automatized, performance or activation does not need much support from working memory (Conway & Engle, 1994; Gathercole & Baddeley, 1993;

Montgomery, 1996; Schmidt, 1990). Dehn (2008) elaborated that comprehension of spoken language happens immediately when related information is directly retrieved from long-term memory; "[t]his activated long-term information automatically facilitates comprehension without the necessity of creating a working memory representation" (p. 98). And there is empirical evidence to back up the argument. For instance, Walters and Caplan (2004) found that differences in online syntactic processing did not relate to working memory capacity. Also, results from the WJ III Tests of Cognitive Ability revealed that among all the academic skills (reading, writing, math, etc.), oral expression showed the weakest correlation with working memory, $r = 0.38$, all other coefficients being greater than 0.50 (McGrew & Woodcock, 2001).

The association between working memory and the GJT results is subject to two possibilities. First, it is possible that during the test, the learner was able to engage in conscious retrieval of the information encoded and registered thorough working memory during the treatment. It is also possible that the information available for use during the test was the information that, after initial encoding and registration through working memory during the treatment, was stored as explicit knowledge in declarative memory (Paradis, 2009). Second, whereas working memory does not seem to be required in oral production that involves instant access to automatized knowledge, research has consistently shown that it is implicated in reading comprehension, both in L1 and L2 learning (e.g., Deneman & Carpenter, 1980; Leeser, 2007). The GJT is a written test and accuracy in learners' performance clearly involves comprehension of the sentence stimuli.

CHAPTER 6 CONCOUSION

The main objective of this study is to investigate the extent to which the effectiveness of corrective feedback is mediated by learner-external factors such as explicitness of feedback and the nature of the linguistic target, as well as learner-internal factors such as proficiency and individual differences in language analytic ability and working memory. It attempts to adopt a holistic, integrated approach to the efficacy of feedback, overcoming the limitation of a monolithic, isolated approach. Noteworthy is the fact that this study is the first attempt to address the three-way interaction between feedback type, learners' proficiency, and the choice of target structure. It is also the first attempt to tackle the complicated relationship between aptitude components, feedback type, and the nature of the linguistic target.

Previous research has demonstrated that recasts did not fare well in classroom settings (e.g., Lyster, 2004; Sheen, 2010) but always seemed to be beneficial in laboratory settings (e.g., Han, 2002; Lyster & Izquierdo, 2009) where the feedback became salient. However, this study showed that the effects of recasts were also determined by the nature of the target structure and learners' overall L2 proficiency: The feedback benefited less advanced learners in the learning of the easy, simple structure (Chinese classifiers), but not the hard, complex structure (the Chinese perfective -*le*); whereas less advanced learners did not benefit from the feedback in the perfective –*le*, more proficient learners did. Also, while proficiency was a mediator in the learning of the perfective –*le*, it was not in the learning of classifiers. The results were interpreted from multiple perspectives: L2 learners' input processing strategies, saliency of the linguistic structure, noticing of feedback, and amount of available cognitive resources. Furthermore,

the delayed effects of recasts were larger than the immediate effects in the learning of the perfective *–le* at the higher proficiency level. This indicates that in the learning of a complex structure, the effects of recasts are well maintained and may even increase over time—a finding that is significant and needs further investigation.

Previous research has revealed an almost unequivocal advantage of explicit feedback over implicit feedback (Carroll & Swain, 1992; Ellis et al., 2006), and similarly, research syntheses on the effectiveness of second language instruction in general showed a superior effect of explicit instruction over implicit instruction (Norris & Ortega, 2000; Spada & Tomita, 2010). This study showed that the "explicit-better-than-implicit" claim was only true for low-proficiency learners, particularly in the learning of the perfective *–le*, a complex structure, and that the difference did not obtain for high-proficiency learners. The result points to the importance of external assistance in prompting the learner to notice the linguistic target at the beginning stage of SLA. Beginners, therefore, should be provided with explicit feedback to achieve optimal learning outcome. The researcher argued for the utilization of implicit feedback for advanced learners because (1) it is as effective as explicit feedback, (2) its effects are better retained, (3) it is an ideal strategy for form-focused instruction, and (4) according to the Socio-Cultural Theory, it affords the appropriate amount of scaffolding in the learner's zone of proximal development and is conducive to making one a more autonomous learner. Certainly, more empirical evidence needs to be accumulated before any definitive claims can be made on the relationship between the explicitness of feedback (or instruction) and level of proficiency. No research syntheses in SLA have investigated proficiency as a mediating factor for instructional treatments because there is a lack of primary studies

that include proficiency as an independent variable (Li, 2010).

Ellis et al. (2006) noted that there is a tendency in feedback research to use tests of explicit knowledge, and consequently, the obtained results are biased toward explicit feedback. To minimize such a bias and have a comprehensive view of the effects of feedback, this study included an EI test as a measure of implicit knowledge and a GJT as a measure of explicit knowledge. The EI tests showed larger effects for feedback than the GJTs. Meta-analyses on the effectiveness of corrective feedback also showed larger effects for measures of implicit knowledge (Li, 2010; Saito & Lyster, 2010). While a claim can be made that feedback contributed to the acquisition of implicit knowledge, it was speculated that the larger effects shown by the EI tests result from several sources: congruence of treatments with tests, ceiling effects for the learners' explicit knowledge, and the typological features of the target language. It was argued that reaction time should be included as a component of an EI test because it is an indicator of the degree of automaticity. It was further argued that the implicit-explicit distinction is continuous and relative and might be constrained by multiple factors such as proficiency, instructional setting, and transfer of L1 processing strategy.

In response to the call from researchers of language aptitude for the investigation of aptitude components and of aptitude-treatment interaction (Robinson, 1997, 2002, 2005) and from researchers of corrective feedback for the investigation of individual difference variables (Ellis & Sheen, 2006), the study explored the interrelation between feedback type, the linguistic structure, and learners' variation in language analytic ability and working memory. It was found that language analytic ability was sensitive to the effects of explicit feedback in the learning of the perfective *–le* and to the effects of implicit

feedback in the learning of classifiers. It was argued that the provision or lack of metalinguistic information about the two target structures was accountable for the interaction between language analytic ability and the learning conditions. Working memory was found to be correlated only with the effects of the explicit condition in the learning of classifiers. Interpretations were sought through the mechanism of working memory, particularly the functions of the central executive.

The relationship between the two cognitive variables and learning conditions is mediated by the measures of feedback effects and the timing of testing. Language analytic ability correlated with both the GJT and EI test results pertaining to the perfective *–le* (in the explicit condition), but only with the GJT results related with classifiers (in the implicit condition). The speculation is that the acquired explicit knowledge about the perfective *–le* was proceduralized, and that it is the proceduralized knowledge that was retrieved during the EI test. The finding that working memory was related with the GJT results but not with the EI test results is probably attributable to the fact that the autonomized knowledge activated during the test was derived from long-term memory. The explicit knowledge measured in the GJT test was activated and processed through working memory. Finally, it was found that the two individual difference variables only correlated with the delayed effects but not the immediate effects of feedback. The immediacy of posting, the research setting, and possible offline processing on the part of the learners may all contribute to the finding.

This study has the following methodological strengths. First, the participants were recruited from two different instruction settings and were randomly assigned to different treatment conditions. The obtained results are likely to be more generalizable because the

participants represent a larger learner population. Second, the contexts for the obligatory use of the target structures were established based on native speakers' speech data extracted from previous studies (in the case of the perfective -*le*) or on native speakers' responses to questionnaire items regarding the usage of the target structure (in the case of classifiers). The validity of the obligatory contexts warranted the provision of feedback during the treatment tasks. Third, the tests used in this study have proven to be valid measures of the related constructs. The HSK, which was used to measure learners' proficiency, has been recognized across the world as an authoritative test for L2 Chinese learners. All test items in the GJT and EI tests were piloted among native speakers of the target language. The Words in Sentences subtest of the MLAT was used as a measure of language analytic ability, and the MLAT has proven to be a valid test of language aptitude by numerous studies. A listening span test rather than a reading span test was used to measure working memory to accommodate the fact that feedback was provided verbally and the encoding and decoding of auditory stimuli was pivotal for the internalization of feedback. All sentence stimuli in the working memory test are developed and validated by Walters and Caplan (1996) and have been used a number of SLA studies (Leeser, 2007; Sagarra, 2007). Also, a non-word repetition or digit span test was not used because they measure phonological short-term memory, which is passive in nature and does not involve information processing (Baddeley, 2003, 2007). Fourth, with respect to data analyses, effect sizes were calculated for all pairwise contrasts. Effect sizes complemented *p*-values in the interpretation of results and made it possible to examine the effects of feedback across target structures and test formats.

Last but not least, as with all studies, this study has limitations. First, the fact that the

study was carried out in a laboratory is a double-edged sword. On one hand, in laboratory settings experiments can be better implemented and variables can be better controlled than in classroom settings. Therefore, the results are less likely to result from latent, distracting variables and can be interpreted with more precision. The processes and principles underlying L2 learning can be clearly scrutinized and inspected. On the other hand, because of the different dynamics of laboratory and classroom settings, different results may have been obtained had the study been contextualized in the classroom setting. For instance, the role of individual differences in L2 learning in the laboratory may be less obvious than in the classroom as a result of the availability of more external assistance in the laboratory setting. Future research may investigate the interaction between feedback type, the linguistic target, and proficiency in a classroom setting, or how individual difference variables affect the effects of feedback differently in classroom and laboratory settings. Second, although the sample size is large (n = 78), the cell sizes are relatively small because of the number of groups (n = 6) the participants were assigned to. The sample sizes of interaction studies, characterized by dyadic interaction and multiple sessions, are typically smaller than other types of research such as psycholinguistic studies because of incurred logistic problems. Regardless, increasing the sample size may generate slightly different results; for instance, non-significant results are likely to turn significant. Third, the duration of treatment is short (less than1 hour), which might be partly responsible for the lack of effects for the low-implicit group in the learning of the perfective *–le*. More exposure to the target structure could have led to better performance by this group.

Fourth, Pearson's correlation analyses were performed to identify the relationship

between the two cognitive variables and the learning outcome of different learning conditions. It is a well-known truth that caution must be exercised where correlation coefficients are interpreted. As Field (2005) pointed out, there are two pitfalls with regard to bivariate correlations—the third variable problem and direction of causality. The former refers to the fact that there may be unmeasured variables that affected the results. The latter suggests that it is not know which of the two involved variables is the "cause" and which is the "effect". As far as this study is concerned, it would have been ideal to investigate working memory and language analytic ability as dichotomous rather than continuous variables, in which case it would have been possible to determine whether the difference between the learners along the two dimensions reacted differently to the interventional treatments. However, the relatively small cell sizes made it difficult to dichotomize the two individual difference variables. Increasing the sample size would be what faces future researchers to better examine the impact of cognitive factors on the effectiveness of different feedback conditions.

NOTES

[1]Neither the information about the number of subjects involved nor the biographic information about the data contributors was provided.

[2]See Sawyer and Ranta (2001) for a review.

[3]All three L1 Korean learners reported having stayed in the U.S. for more than five years, and all were enrolled in academic programs at the data-contributing universities. However, the extent to which they resemble L1 English speakers in terms of English proficiency is uncertain. This might cause a concern when these learners take the two aptitude tests—their L2 English background might make their test performances different from their L1 English peers. For this reason, they were placed in the two control groups, whose aptitude test scores were not used in the data analyses, which made the randomness of group assignment relative rather than absolute.

[4]Clearly, to a second language learner of Chinese whose native language (English) is a non-classifier language, there are two challenges associated with classifier learning. First, the learner must develop the awareness that a classifier must be used between a determiner and the following nominal phrase. Second, the learner must choose a proper classifier depending on the physical properties of the noun. As previously discussed, classifiers are semantically related with the physical characteristics of the objects they co-occur with. However, the connections between many classifiers and their accompanying nouns have become invisible and appear arbitrary as a result of the fact that the language has changed. From a pedagogical perspective, while it is possible to explain the rationale behind the choice of a classifier in some cases, it is not always realistic to do so because

of the seeming arbitrariness in terms of the connection between the classifier and the corresponding nominal phrase. To provide consistent instructional treatment, in this study, the explicit feedback has two components: informing the learner that a classifier is required and providing the correct classifier.

[5]The tasks for classifiers were revised based on Li (2009).

[6]Alternatively, the test taker is asked to decide whether he/she agrees or disagrees with the statement.

[7]States verbs (such as "like") were not included in the treatment or test. States verbs are atelic, so a delimiting device such as time duration must be added to warrant the use of *–le*. Though theoretically possible, using states verbs with a time period sounds odd and is therefore uncommon in Chinese as well as other languages.

[8]Recasts showed some effects on posttests as compared with pretest results, but did not show any effects when compared with the control group.

[9]Ceiling levels of explicit knowledge were also found by Ellis et al. (2006).

[10]Robinson (2002) reported a correlation of 0.35 between working memory and aptitude as measured by the MLAT; Safar & Kormos (2008) found that working memory correlated with inductive language learning ability at $r = 0.33$, and with the global aptitude scores at $r = 0.36$. These results, together with the correlation ($r = 0.3$) between working memory and language analytic ability consistently demonstrate that (1) working memory is moderately related to aptitude and components of aptitude measured by the

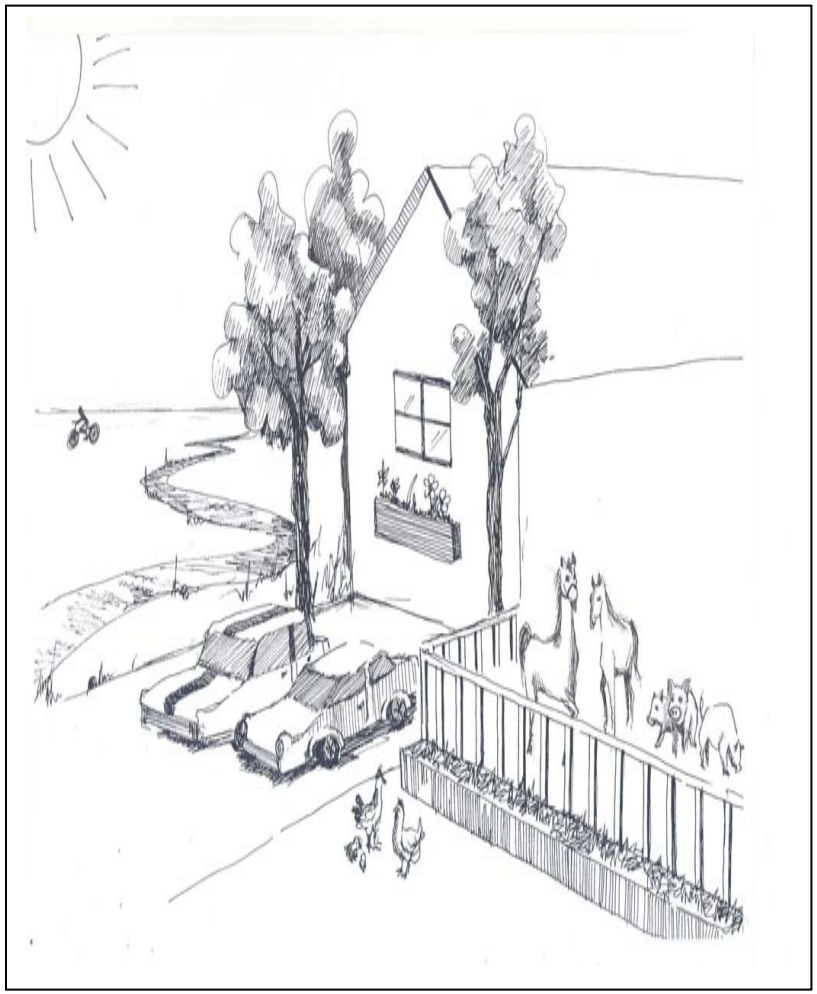
MLAT and yet it is a separate construct, and (2) it is justified to consider working memory an aptitude component.

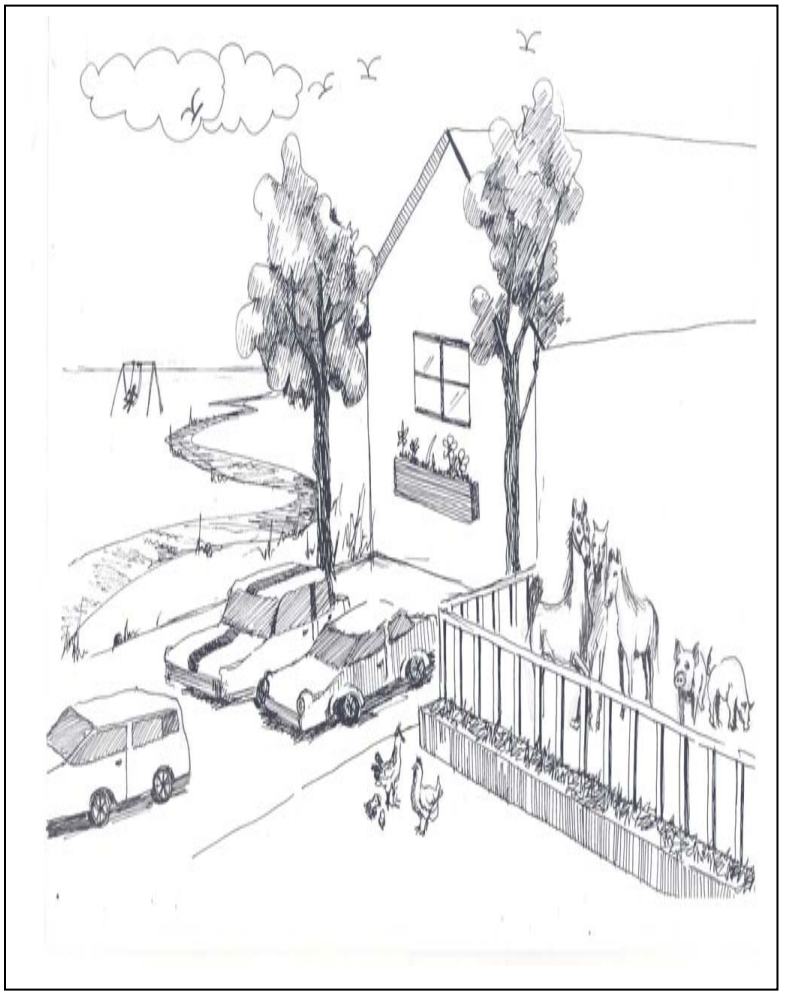
APPENDICES

APPENDIX A A Sample Card Used in Picture Description*

*For interpretation of the references n this and all other figures, the reader is referred to the electronic version of this dissertation.

APPENDIX B A Sample Picture Set Used in "Spot the Differences"





APPENDIX C Table C-1. Perfective *–le*: Descriptive Statistics Related to Raw Scores

Test	*Proficiency*	*Group*	*n*	*Pretest*		*Posttest 2*		*Posttest 3*	
				Mean	SD	Mean	SD	Mean	SD
GJT	Low	Implicit	14	5.75	1.01	7.75	1.99	7.89	2.31
		Explicit	15	4.70	1.81	11.11	1.91	9.00	2.45
		Control	10	3.80	0.95	5.20	1.77	5.35	1.53
	High	Implicit	14	7.29	1.88	9.75	2.14	11.61	2.19
		Explicit	14	6.57	1.86	12.23	1.36	1.096	3.27
		Control	11	7.50	2.30	8.59	2.88	9.68	1.78
EI	Low	Implicit	14	1.79	1.37	6.18	2.62	4.46	2.79
		Explicit	15	1.36	1.03	8.21	2.73	5.23	2.55
		Control	10	1.70	1.71	2.85	2.53	3.05	1.77
	High	Implicit	14	6.50	3.61	11.25	2.06	10.75	2.49
		Explicit	14	5.31	3.81	11.88	2.53	10.67	3.69
		Control	11	6.41	3.48	8.09	4.02	7.00	2.91

APPENDIX D Table D-1. Classifiers: Descriptive Statistics Related to Raw Scores

				Pretest		*Posttest 1*		*Posttest 2*	
Test	*Proficiency*	*Group*	*n*	Mean	SD	Mean	SD	Mean	SD
GJT	Low	Implicit	14	5.71	1.39	7.86	1.48	8.21	2.07
		Explicit	15	4.80	1.41	9.83	2.43	8.93	1.74
		Control	10	5.20	0.88	5.30	1.09	5.40	2.01
	High	Implicit	14	6.29	0.80	10.61	2.37	10.18	2.59
		Explicit	14	6.07	0.89	11.68	2.54	10.78	2.51
		Control	11	6.77	0.98	7.73	1.33	7.01	1.11
EI	Low	Implicit	14	2.07	1.66	6.68	2.38	6.18	2.78
		Explicit	15	1.57	1.10	8.13	2.91	6.96	3.55
		Control	10	2.25	1.79	3.80	2.36	3.95	2.27
	High	Implicit	14	4.21	1.88	8.86	3.00	8.29	2.91
		Explicit	14	4.32	2.21	10.61	2.93	10.04	2.79
		Control	11	5.32	2.02	6.41	2.05	7.41	2.11

REFERENCES

Ahrens, K. (1994). Classifier production in normals and aphasics. *Journal of Chinese Linguistics. 22,* 202-247.

Aljaafreh , A. , & Lantolf , J. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal*, *78*, 465–483 .

Ammar, A., & Spada, N. (2006). One size fits all? Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition, 28,* 543-574.

Anderson, R. (1989). *Unpublished lecture in the seminar on the acquisition of tense and aspect.* University of California, Los Angeles.

Avons, S., Wragg, C., Cupples, L., & Lovegrove, W. (1998). Measure of phonological short-term memory and their relationship to vocabulary development. *Applied Psycholinguistics, 19,* 583-601.

Baddeley, A. (1986). *Working memory.* Oxford: Oxford University Press.

Baddeley, A. (2000). The episodic buffer: A new component in working memory? *Trends in Cognitive Science, 4,* 417-423.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36,* 189-208.

Baddeley, A. (2006). Working memory: An overview. In S. Pickering (Ed.), *Working memory and education* (pp. 1-31). Burlington, MA: Academic Press.

Baddeley, A. (2007). Working memory, thought, and action. Oxford: Oxford University Press.

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Baddeley, A., & Hitch, G. (1994). Developments in the concept of working memory. *Neuropsychology, 8,* 485-493.

Baddeley, A., & Logie, R. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). Cambridge: Cambridge University Press.

Carpenter, H., Jeon, S., MacGregor, D., & Mackey, A. (2006). Learners' interpretations of recasts. *Studies in Second Language Acquisition, 28,* 209-236.

Carroll, J. (1962). The prediction of success in intensive foreign language training. In R.

Glaser (Ed.), *Training research and education* (pp. 87–136). Pittsburgh: University of Pittsburgh Press.

Carroll, J. (1973). Implications of aptitude test research and psychological theory for foreign language teaching. *International Journal of Psycholinguistics, 2,* 5-14.

Carroll, J. B. (1981). Twenty-five years of research on 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. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J., & Sapon, S. (1959). *Modern Language Aptitude Test*. New York: The Psychological Corporation/Harcourt Brace Jovanovich.

Carroll, J., & Sapon, S. (2002). *Manual for the MLAT*. N. Bethesda, Maryland: Second Language Testing, Inc.

Carroll, S., & Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition, 15,* 357-386.

Chang, W. (1986). *The particle le in Chinese narrative discourse: An interactive description.* Ph.D. dissertation. The University of Florida, Gainesville.

Chang, Hsianghua. (2002). *Child acquisition of the aspect marker –le in Mandarin Chinese*. Master's thesis. Michigan State University, East Lansing.

Chao, Y. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, H. (1996). *A study of the effect of corrective feedback on foreign language learning: American students learning Chinese classifiers.* Ph.D. dissertation. University of Pennsylvania, Philadelphia.

Chou, C., Eagar, J., & Chiang, J. (1999). *A new China: Intermediate reader of modern Chinese.* Princeton: Princeton University Press.

Comrie, B. (1976). *Aspect.* Cambridge: Cambridge University Press.

Conway, A., & Engle, R. (1994). Working memory and retrieval: A resource dependent resource inhibition model. *Journal of Experimental Psychology: General, 123,* 354-373.

Conway, A., Jarrold, C. Kane, M., Miyake, A., & Towse, J. (2007) (Eds.). *Variation in working memory.* Oxford: Oxford University Press.

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 62-101). Cambridge: Cambridge University Press.

Cook, V. (1996). *Second language learning and language teaching.* London: Arnold.

Christensen, M. (1994). *Variation in spoken and written Mandarin narrative discourse.* Ph.D. dissertation. Ohio State University, Columbus.

Craig, C. (1986). Introduction. In C. Craig (Ed.), *Noun classes and categorization* (pp. 1-11). Philadelphia: John Benjamins.

Cronbach, L., & Snow, R. (1977). *Aptitude and instructional methods: A handbook for research on interactions.* New York: Irvington.

Dehn, M. (2008). *Working memory and academic learning: Assessment and intervention.* Hoboken, NJ: John Wile & Sons, Inc.

DeKeyser, R. (1993). The effect of error correction on L2 grammar knowledge and oral proficiency. *The Modern Language Journal, 77,* 501-514.

DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition, 22,* 499-533.

DeKeyser, R. (2005). What makes learning second language grammar difficult? A review of issues. *Language Learning, 55,* 1-25.

DeKeyser, R. (2007). Skill acquisition theory. In B. VanPatten & J. Williams, *Theories in second language acquisition* (pp. 97-113). Mahwah, New Jersey: Lawrence Erlbaum Associates.

DeKeyser, R. (Ed.) (2008). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology.* Cambridge: Cambridge University Press.

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research, 20,* 445-464.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450-466.

Daneman, M., & Merikle, P. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 422-433.

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In Catherine D. and Michael L. (Eds.) *Handbook of Second Language Acquisition* (pp.589-630). Malden, MA: Blackwell Publishing Ltd

Doughty, C. (2001). The cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). Cambridge: Cambridge University Press.

Doughty, C., & Varela, E. (1998). Communicative focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp.114–138). New York: Cambridge University Press.

Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Duff, P., & Li, D. (2002). The acquisition and use of perfective aspect in Mandarin. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 417-452). Philadelphia: John Benjamins.

Egi , T. (2007). Recasts, learners' interpretations, and L2 development . In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 249–267). Oxford : Oxford University Press .

Ehrman, M., & Oxford, R. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal, 79,* 67-89.

Ellis, N. (2008). Implicit and explicit knowledge about language. In J. Cenoz and N. Hornberger (Eds.), *Encyclopedia of language and education* (pp. 119-131). New York: Springer.

Ellis, N., & Sinclair, S. (1996). Workign memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology, 49A,* 234-250.

Ellis, R. (2001). Investigating form-focused 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). Modeling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics, 27,* 431-463.

Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp.339-360). New York: Oxford University Press.

Ellis, R. (2008). *The study of second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. 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. 3-25). Tonawanda, NY: Multilingual Matters.

Ellis, R. (2010). Epilog: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32,* 335-349.

Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning, 51,* 281-318.

Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (Eds.) (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Tonawanda, NY: Multilingual Matters.

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

Ellis, R., & Sheen, Y. (2006). Reexamining the role of recasts in second language acquisition. *Studies in Second Language Acquisition, 28,* 575-600.

Engle, R. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19-23.

Erbaugh, M. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In C. Craig (Ed.), *Noun classes and categorization* (pp. 399-436). Philadelphia: John Benjamins.

Erbaugh, M. (2001). The *Chinese pear stories: Narratives across seven dialects.* Available at http://www.pearstories.org/

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research, 9,*147–171.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27,* 464-491.

Field, A. (2005). *Discovering statistics using SPSS.* Thousand Oaks, CA: SAGE Publications Inc.

French, L. (2006*). Phonological working memory and second language acquisition: Developmental study of Francophone children learning English in Quebec.* New York: The Edwin Mellen Press

Gardner, R. (1985). *Social psychology and second language learning: The role of attitudes and motivation.* London: Arnold.

Gass, S. (1997). *Input, interaction, and the second language learner.* Mahwah, NJ: Erlbaum.

Gass, S. (2003). Input and interaction. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 224–255). Malden, MA: Blackwell.

Gass, S. (2004). Conversation and input-interaction. *The Modern Language Journal, 88,* 579-616.

Gass, S., & Selinker, L. (2008). Second language acquisition: An introductory course. New York: Routledge.

Gass, S., Svetics, I., & Lemelin, S. (2003). Differential effects of attention. *Language Learning, 53,* 497-546.

Gass, S., & Varonis, E. (1994). Input, interaction and second language production. *Studies in Second Language Acquisition, 16,* 283-302.

Gathercole, S., & Alloway, T. (2008). *Working memory and learning: A practical guide for teachers.* London: Sage Publications.

Gathercole, S., & Baddeley, A. (1993). *Working memory and language*. East Sussex, UK: Lawrence Erlbaum.

Gathercole, S., Frankish, C., P:ickering, S., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 84-95.

Goldschneider, J., & DeKeyser, R. (2005). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 55* (Suppl.), 27-77.

Han, Z. (2002). A study of the impact of recasts on tense consistency in L2 output.

TESOL Quarterly, 36, 543–572.

Harley, B., & Hart, D. (1997). Language aptitude and second language proficiency in classroom learners of different starting ages. *Studies in Second Language Acquisition, 19,* 379-400.

Harley, B., & Hart, D. (2002). Age, aptitude, and second language learning on a bilingual exchange. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 301-330). Philadelphia: John Benjamins.

Harrington, M. (1991*). Individual differences in L2 reading: Processing capacity versus linguistic knowledge.* Paper presented at the Annual Meeting of the American Association of Applied Linguistics.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition, 14,* 25-38.

Havranek, G. & Cesnik, H. (2001). Factors affecting the success of corrective feedback. In S. Foster-Cohen & A. Nizegorodzew (Eds.). *EUROSLA Yearbook,* Volume 1. Amsterdam: John Benjamins.

Huang, W., & Ao, Q. (2002). *Chinese language and culture: An intermediate reader.* Hong Kong: The Chinese University Press.

Hulstijn, J. (2002). Towards a unified account of the representation, processing, and acquisition of second language knowledge. *Second Language Research, 18,* 193-223.

Hulstijn, J., & de Graaff, R. (1994). Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal. *AILA Review, 11,* 97-112.

Hummel, K. (2009). Aptitude, phonological memory, and second language proficiency in nonnovice adult learners. *Applied Psycholinguistics, 30,* 225-249.

Ishida, M. (2004). Effects of recasts on the acquisition of the aspectual form *–te i-(ru)* by learners of Japanese as a foreign language. *Language Learning, 54,* 311-394.

Iwashita, N. (2003). Positive and negative input in task-based interaction: Differential affects on L2 development. *Studies in Second Language Acquisition, 25,* 1-36 .

Juffs, A. (2004). Representation, processing, and working memory in a second language. *Transactions of the Philological Society, 102,* 199-225

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122-49.

Kim, H. R., & Mathes, G. (2001). Explicit vs. implicit corrective feedback. *The Korean TESOL Journal, 1,* 57-72.

Krashen, S. (1981). *Second language acquisition and second language learning.* Oxford: Pergamon.

Krashen, S. (1994). The input hypothesis and its rivals. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 45-78). London: Academic Press.

Krashen, S. (1995). *Principles and practice in second language acquisition.* Hertfordshire, England: Prentice Hall Europe.

Lantolf, J. (Ed.) (2009). *Sociocultural Theory and second language learning.* Oxford: Oxford University Press.

Leeman, J. (2003). Recasts and second language development: Beyond negative evidence. *Studies in Second Language Acquisition, 25,* 37–63.

Leeser, M. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning, 57,* 229-270.

Lehto, J. (1996). Are executive function tests dependent on working memory capacity? *The Quarterly Journal of Experimental Psychology, 94A*, 29-50

Li, C. & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. Los Angeles, CA: University of California Press.

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect.* Berlin: Mouton de Gruyter.

Li, S. (2009). The differential effects of implicit and explicit feedback on L2 learners of different proficiency levels. *Applied Language Learning*, *19,* 53-79.

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning, 60,* 309-365.

Li, W. (2000). Numeral-classifiers as a grounding mechanism in mandarin Chinese. *Journal of Chinese linguistics, 28,* 337-367.

Loewen, S. (2004). Uptake in incidental focus on form in meaning-based ESL lessons. *Language Learning, 54,* 153–188.

Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition, 27*, 361-386.

Loewen, S., & Nabei, T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 361-377). New York: Oxford University Press.

Loewen, S., & Philp, J. (2006). Recasts in the adult English L2 classroom: characteristics, explicitness, and effectiveness. *The Modern Language Journal, 90,* 536-556.

Lightbown, P. (2008). Transfer appropriate processing as a model for class second language acquisition. In Z. Han (Ed.), *Understanding second language process* (pp. 27-44). Clevedon, UK: Multilingual Matters.

Liu, X. (2001). *Explaining the grammatical meaning of the sentence-final le in modern Chinese.* Paper presented the 10th International Conference of Chinese Linguistics, in conjunction with the 13th North American Conference on Chinese Linguistics. University of California.

Liu, Y, Yao, T., Bi, N., Ge, L., & Shi, Y. (2009). *Integrated Chinese* (3rd ed.). Boston: Cheng & Tsui Company.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition.* Vol. 2: Second language acquisition (pp. 413-468). New York: Academic Press.

Long, M. H. (2007). *Problems in SLA*. Mahwah, NJ: Erlbaum.

Long, M., Inagaki, S., & Ortega, L. (1998). The role of negative feedback in SLA: Models and recasts in Japanese and Spanish. *The Modern Language Journal, 82,* 357-371.

Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning, 48,* 183–218.

Lyster, R. (2001). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning, 51* (Suppl. 1), 265-301.

Lyster, R. (2004). Different effects of prompts and effects in form-focused instruction. *Studies in Second Language Acquisition, 26*, 399–432.

Lyster, R., & Izquierdo, J. (2009). Prompts versus recasts in dyadic interaction. *Studies in Second Language Acquisition, 59,* 453-498.

Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance.

Studies in Second Language Acquisition, 28, 269–300.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition, 19,* 37-66.

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition, 32,* 265-302.

Mackey, A. (1999). Input, interaction and second language development. *Studies in Second Language Acquisition, 21,* 557-587.

Mackey, A. (Ed.) (2007). *Conversational interaction in SLA: A collection of empirical studies.* New York: Oxford University Press.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning, 60,* 501-533.

Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive international feedback? *Studies in Second Language Acquisition, 22,* 471-497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in SLA: A collection of empirical studies* (pp. 408–452). New York: Oxford University Press.

Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: recasts, responses, and red herrings? *The Modern Language Journal, 82,* 338-356.

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, *Individual differences and instructed language learning* (181-209). Philadelphia: John Benjamins.

McDonough, K. (2007). Interactional feedback and the emergence of simple past activity verbs in L2 English. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 323–338). New York: Oxford University Press.

McGrew, K., & Woodcock, R. (2001). *Woodcock-Johnson III technical manual.* Itasca, IL: Riverside Publishing.

Michas, I., & Henry, L. (1994). The link between phonological memory and vocabulary acquisition. *British Journal of Developmental Psychology, 12,* 147-164.

Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp.339-364). Mahwah, NJ: Erlbaum.

Montgomery, J. (1996). Sentence comprehension and working memory in children with specific language impairment. *Topics in Language Disorders, 17,* 19-32.

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

Nicholas, H., Lightbown, P., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning, 51,* 719–758.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50,* 417–528.

Ohta, A. (2009). Rethinking interaction in SLA: Developmentally appropriate assistance in the zone of proximal development and the acquisition of L2 grammar. In J. Lantolf, *Sociocultural theory and second language learning* (pp. 51-78). New York: Oxford University Press.

Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletion of the Psychonomic Society, 30,* 287-289.

Oxford, R. (1995). Gender differences in language learning styles: What do they mean? In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 34-46). Boston: Heinle and Heinle.

Pagagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language, 30,* 331-347.

Panova, I., & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly, 36,* 573-595.

Paradis, M. (2009). *Declarative and procedural determinants of second languages.* Philadelphia, PA: John Benjamins.

Petersen, C. & Al-Haik, A. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational and Psychological Measurement, 6,* 369-380.

Pica, T. (1988). Interlanguage adjustments as an outcome of NS-NNS negotiated

interaction. *Language Learning, 38,* 45-73.

Pienemann, M. (1998). *Language processing and second language development: Proccessability Theory*. Amsterdam: John Benjamins.

Philip, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition,* 25, 99-126.

Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. New York: The Psychological Corporation.

Polio, C. (1994). Non-native speakers' use of nominal classifiers in mandarin Chinese. *JCLTA, 29,* 51-66.

Polio, C. & Gass, S. (1998). The effect of interaction on the comprehension of nonnative speakers. *Modern Language Journal, 82,* 308-319.

Ranta, L. (2002). The role of language analytic ability in the communicative classroom. In P. Robinson (Ed.), *Individual differences and instructed language* learning (pp. 159-180). Philadelphia: John Benjamins.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219-235.

Reves, T. (1983). *What makes a good language learner? Personal characteristics contributing to successful language acquisition.* Ph.D. dissertation. Hebrew University, Israel.

Robinson, P. (1997). Individual differences and fundamental similarity of implicit and explicit adult second language learning. *Language Learning, 47,* 45-99.

Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA: A replication and extension of Reber, Walkenfield and Hernstadt (1991). In P. Robinson, *Individual differences and instructed language learning* (pp. 211-266). Philadelphia: John Benjamins.

Robinson, P. (2005). Aptitude and second language acquisition. *Annual Review of Applied Linguistics, 25,* 46-73.

Robinson, p., & Yamaguchi, Y. (1999). *Aptitude, task feedback and generalizability of focus on form: A classroom study.* Paper presented at the 12th AILA World Congress, Waseda University, Tokyo.

Roehr, K., & Ganem-Gutierrez, G. (2009). The status of metalinguistic knowledge in instructed adult L2 learning. *Language Awareness, 18,* 165-181.

Ross, S., Yoshinaga, N., & Sasaki, M. (2002). Aptitude-exposure interaction effects on Wh-movement violation detection by pre-and-post-critical period Japanese bilinguals. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 267-299). Philadelphia: John Benjamins

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for second language acquisition: A meta-analysis of the research. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 131–164). Amsterdam: Benjamins.

Safar, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *IRAL, 46,* 113-136.

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 229–248). New York: Oxford University Press.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence.* New York: Lang.

Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson, *Cognition and second language instruction* (pp. 319-353). Cambridge: Cambridge University Press.

Segalowitz, N. (1997). Individual differences in second language acquisition. In A.M.B. de Groot & J. F. Kroll (Eds.), *Language acquisition studies in generative grammar* (pp. 85-112). Killsdale, NJ: Erlbaum.

Service, E., & Kohonen, V. (1995). Is the relation between phonological memory and foreign-language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics, 16,* 155-172.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129-158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2003). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *The acquisition, processing, and use of formulaic sequences* (pp. 55–86). Amsterdam: John Benjamins.

Schwartz, B. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition, 15,* 147-163.

Sheen, Y. (2004). Corrective feedback and leaner uptake in communicative classrooms across instructional settings. *Language Teaching Research, 8,* 263–300.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research, 10,* 361–392.

Sheen, Y. (2007a). 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* (pp.301-322). New York: Oxford University Press.

Sheen, Y. (2007b). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly, 41,* 255-283.

Sheen, Y. (2008). Recasts, language anxiety, modified output, and L2 learning. *Language Learning, 58,* 835-874.

Sheen, Y. (2010). The role of oral and written corrective feedback in SLA: Introduction. *Studies in Second Language Acquisition, 32,* 169-179.

Sheen, Y. (2010). Differential effects of oral and written corrective feedback in the ESL classroom. *Studies in Second Language Acquisition, 32,* 203-234.

Shi, Z. (1990). Decomposition of perfectivity and inchoativity and the meaning of the particle –le in Mandarin Chinese. *Journal of Chinese Linguistics, 18,* 95-123.

Skehan, P. (1982). Memory and motivation in language aptitude testing. Ph.D. dissertation. University of London.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 70-94). Philadelphia: John Benjamins.

Smith, C. (1997). *The parameter of aspect.* Dordrecht: Kluwer.

Snow, R. (1987). Aptitude complexes. In R. Snow & M. Farr (Eds.), *Aptitude, learning, and instruction* (pp. 13-59). Hillsdale, NJ: Erlbaum.

Snow, R. (1991). Aptitude-treatment interaction as a framework for research on individual differences in psychotherapy. *Journal of Consulting and Clinical Psychology, 59,* 205-216

Snow, R. (1994). Abilities in academic tasks. In R. Sternberg & R. K. Wagner (Eds.), *Mind in context: Interactionnist perspectives on human intelligence* (pp. 3-37). New York: Cambridge University Press.

Spada, N. (1997). Form-focused instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching, 29,* 1-15.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60,* 263-308.

Sparks, R., Patton, J., Ganschow, L., & Humbach, N. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics, 30,* 725-755

Spolsky, B. (1989). *Conditions for second language learning.* Oxford: Oxford University Press.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in it development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235-252). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook and B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honor of H.G. Widdowson* (pp. 125-144).Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook on research in second language teaching and learning* (pp. 471-484). Mahwah, NJ: Lawrence Erlbaum.

Tai, J., & Wang, L. (1990). A semantic study of the classifier Tiao. *Journal of the Chinese Language Teachers Association, 25,* 35-56.

Thompson, C. (1968). *Aspects of the Chinese verb.* Linguistics, 38, 70-76.

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* (pp. 171–195). New York: Oxford University Press.

Van Den Berg, M. (1989). *Modern standard Chinese: Een functionele grammatical.* Muiderberg: Coutinho.

Van den Berg, M., & Wu, G. (2006). *The Chinese particle –le.* New York: Routledge.

VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten & J. Williams, *Theories in second language acquisition* (pp. 115-135). Mahwah, NJ: Lawrence Erlbaum Associates.

Vendler, Z. (1957). Verbs and times. *Philosophical Review, 66,* 143-160.

Waters, G., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology, 49A,* 51-79.

Waters, G., & Caplan, D. (2004). Verbal working memory and on-line syntactic processing: Evidence from self-paced listening. *The Quarterly Journal of Experimental Psychology, 57A,* 129-163

Wen, X. 1995. Second language acquisition of the Chinese particle *le*. *International Journal of Applied Linguistics, 5,* 45-62.

Wen, X. (1997). Acquisition of Chinese aspect: An analysis of the interlanguage of learners of Chinese as a second language. *ITL: Review of Applied Linguistics, 117/118,* 1-26.

Wesche, M. (1981). Language aptitude in 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.

Williams, J. (1999). Memory, attention and inductive learning. *Studies in Second Language Acquisition, 21,* 1-48.

Williams, J. (2005). Learning with awareness. *Studies in Second Language Acquisition, 27,* 269-304.

Williams, J., & Lovatt, P. (2003). Phonological memory and rule learning. *Language Learning, 53,* 67-121.

Wu, Y., & Bodomo, A. (2009). Classifiers ≠ determiners. *Linguistic Inquiry, 40,* 487-503

Wu, S., Yu, Y., Zhang, Y., & Tian, W. (2007). *Chinese link.* Upper Saddle River, NJ: Pearson Education, Inc.

Xiao, R., & McEnery, T. (2004). *Aspect in Mandarin Chinese*. Philadelphia: John Benjamins Publishing Company.

Yang, S. (1995). *The aspectual system of Chinese.* Ph.D. dissertation. University of Victoria, Canada.

Yang, J. (2002). *The acquisition of temporality by adult second language learners of Chinese.* Ph.D. dissertation. Tucson: The University of Arizona.

Yang, J. (2003). Back to the basic: The basic function of particle LE in modern Chinese. *Journal of the Chinese Language Teachers Association, 38,* 77-96.

Yang, S., Huang, Y., & Sun, D. (1999). Acquisition of Aspect in Chinese as a Second Language. *Journal of the Chinese Language Teachers Association 34*, 31-54.

Yang, S., Huang, Y., & Sun, D. (2000).Underuse of temporal markers in Chinese as a Second Language. *Journal of the Chinese Language Teachers Association* 35, 87-116.

Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition, 32,* 235-263.

Yao, T., Liu, Y., Bi, N., Hayden, J., & Wang, X. (2005). *Integrated Chinese*. Boston: Cheng & Tsui Company.

Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics, 16,* 43-59.

Zhang, K., Liu, S., Chen, X., Zuo, S., Shi, J., & Liu, X. (2002). *New practical Chinese reader.* Beijing: Beijing Languages University Press.

Zhang, Y. (2005). Processing and formal instruction in the L2 acquisition of five Chinese grammatical morphemes. In M. Pienemann, *Cross-linguistic aspects of processability theory* (pp. 155-177). Philadelphia: John Benjamins.

Zhao, X., Li, Y., & Lin, L. (1999). *An intensive reading course of intermediate Chinese.* Beijing: Beijing University Press