EFFECTS OF FEEDBACK TIMING AND TYPE ON LEARNING ESL GRAMMAR RULES

By

Elizabeth H. P. Lavolette

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2014

ABSTRACT

EFFECTS OF FEEDBACK TIMING AND TYPE ON LEARNING ESL GRAMMAR RULES

By

Elizabeth H. P. Lavolette

The optimal timing of feedback on formative assessments is an open question, with the cognitive processing window theory (Doughty, 2001) underlying the interaction approach suggesting that immediate feedback may be most beneficial for language acquisition (e.g., Gass, 2010; Polio, 2012) and two educational psychology hypotheses conversely suggesting that delayed feedback may be superior for error correction (dual-trace hypothesis, Kulik & Kulik, 1988; interference-perseveration hypothesis, Kulhavy & Anderson, 1972).

To explore the effects of varied feedback timing on both item learning and rule generalization, 118 intermediate ESL students were randomly assigned to item-by-item or end-of-test computerized feedback conditions. Within each timing group, half of the students received feedback that indicated the correct answer and whether they had answered correctly or incorrectly (without metalinguistic feedback). The other students received additional feedback that stated a rule that applied to the item (metalinguistic feedback). A pretest, two treatments, a 5-minute-delayed posttest, and a 1-week-delayed posttest were administered. Each treatment contained 17 multiple-choice items that were followed by item-by-item or end-of-test feedback. The pretest and both posttests included all items from the treatment (to test item learning) plus 10 new multiple-choice items to test generalization of rules. The data were analyzed using mixed-design ANOVAs.

The item-by-item metalinguistic feedback group had higher gain scores than the other feedback groups on the treatment items on both posttests, although no significant main effects were found for either feedback timing or type. This suggests that item-by-item metalinguistic

feedback is better for item learning. On the items that did not appear on the treatment, the item-by-item groups outperformed the end-of-test groups, with a marginally significant main effect of feedback timing, $F(1, 108) = 3.61$, $p = .06$, $\eta^2_{part} = .032$. This suggests that item-by-item feedback may be better for learning to generalize. In addition, the groups that received item-by-item feedback spent significantly less time reading the feedback than did the groups who received end-of-test feedback, $F(1, 108) = 4.14$, $p = .044$, $\eta^2_{part} = .037$. These combined results suggest that item-by-item metalinguistic feedback may be more effective and efficient for language learners for both item learning and learning to generalize, although the small effects sizes indicate that providing this type and timing of feedback should be only one of many interventions to improve instruction. In addition, these results lend support to the cognitive processing window theory and attention-based theory underlying the interaction approach.

I dedicate this work to my husband, R. Jess Lavolette. This work was only possible with your love and support.

ACKNOWLEDGEMENTS

Many people contributed to this study. First, I thank all of the members of my committee, Dr. Susan Gass and Dr. Paula Winke, who have provided helpful guidance and feedback throughout the process. In particular, my co-chairs, Dr. Senta Goertler and Dr. Charlene Polio, have spent countless hours advising me and challenging me to improve my work.

I am grateful to numerous others who contributed to this work. Thank you to Rod Ellis, who provided helpful feedback on the study design. Thank you to the friends and family who anonymously responded to norming survey for the test items. Thank you to Brian Adams in the MSU College of Arts and Letters who installed the Concerto testing platform on a server, which allowed me to collect all of the data. Thank you to the students in Dr. Goertler's CALL class who gave me helpful feedback on the pilot study proposal and report drafts. Thank you to Dr. Daniel Reed, who kindly granted permission for me to conduct my research in ELC classes and provided data on mean TOEFL scores. Thank you to the teachers in the MSU English Language Center who allowed me to use their class time and access their students for my pilot and main studies: Leah Addis, Collin Blair, Janet Colson, Carmella Gillette, Ashley Hewlett, Peter Hoffman, David Krise, Ann Letson, Alicia Norgrove, Laura Ramm, Stacy Sabraw, Peter Sakura, Carlee Salas, and Cristen Vernon. Thank you to Mike Kramizeh, who was unendingly patient in scheduling the computer labs, and thank you to the lab assistants were very helpful in getting the labs prepared for my study. Last but not least, thank you to the ELC students who participated in the study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Purpose and Rationale of the Study

Computer-assisted language learning (CALL) applications can provide instant feedback to language learners on a wide variety of activities, from multiple-choice questions and cloze exercises to speaking and writing activities. Proponents of using technology in language teaching and assessment often explicitly claim that immediate feedback is superior to delayed feedback, without citing any evidence (e.g., Alderson, 2005; Brown, 1997; Chun & Brandl, 1992; García & Arias, 2010; Kane-lturrioz, 2008; Lan, Sung, & Chang, 2007). When these claims are not explicit, they are often implicit in the design of language-learning applications (e.g., Amaral & Meurers, 2011; Heift, 2010; Nagata & Swisher, 1995; Nagata, 1999). In addition, major commercial CALL applications like Rosetta Stone, Tell Me More, Duolingo, Open English, and Pimsleur all provide immediate feedback to learners on most of their activities. However, no evidence shows that the immediate feedback produced by computer applications is more useful to language learners than similarly produced delayed feedback. Related research in educational psychology has shown that immediate feedback on multiple-choice questions (provided by a computer or by using specially prepared paper forms) may help students retain material better than feedback that is delayed until the end of the activity or until a day later (e.g., Dihoff, Brosvic, Epstein, & Cook, 2004; Kulik & Kulik, 1988). Other researchers have found delayed feedback to be more effective (e.g., Guzmán-Muñoz & Johnson, 2008; Schooler & Anderson, 1990). However, little is known about how feedback timing affects second language learning. Given these conflicting results in educational psychology and the gap in the CALL and SLA literature, the purpose of this study is to provide evidence of how varied feedback timing affects the acquisition of English by adult second language learners. Specifically, in this study, I focus on measuring how differing timings

of feedback on multiple-choice questions affect learning to apply rules for using English articles. In addition, I address the question of how providing or not providing metalinguistic feedback affects learning the same rules and whether it interacts with feedback timing.

1.2 Definitions

Some concepts need to be defined before proceeding. First, a fundamental concept in this study is *feedback*. I use Cohen's (1985) simple definition from the context of computer-based instruction of "the message which follows the response made by the learner" (p. 33). This message may be delivered by an interlocutor during conversational interaction or during a test by a computer, and the message may follow more or less quickly after the response made by the learner. The types of feedback that will be discussed are defined by the amount of information provided. At the end of the spectrum that provides less information, *knowledge of results* feedback tells the learner only whether he or she answered correctly or incorrectly. Providing slightly more information, *knowledge of correct response* feedback implicitly or explicitly includes knowledge of results feedback, but also indicates the correct response. Finally, *informational feedback* includes knowledge of correct response feedback plus additional information, such as a rule that can be extended beyond the current question.

For timing, the terms *immediate* and *delayed* are imprecise. The terms are not used in a systematic way in the literature, either within second language acquisition (SLA) or in educational psychology, despite the publication 25 years ago of a taxonomy of such feedback in computer-based instruction (Dempsey & Wager, 1988). Drawing on this taxonomy and the work of Henshaw (2011), in the current study, I will focus on feedback on test items under two conditions: feedback provided on an item-by-item basis and provided at the end of a test (corresponding to Dempsey and Wager's (1988) "end-of-module" definition of feedback). When

considering the previous literature, I will use two further categories: delays shorter than end-of-test, such as a delay of 7 seconds after a response to a question before feedback is provided; and delays longer than end-of-test, such as a 24-hour delay. I will use these terms to better describe and compare the previous literature.

The next concept that needs to be defined is *learning*. Two types of learning, *system* and *item learning*, will be considered in this study (Schmidt, 1995). A system has been learned if the learner can correctly apply a rule to a previously unseen situation, such as a multiple-choice question or a writing task. Note that being able to state the rule (a type of verbal information) is not what is considered here. Rather, system learning is an intellectual skill (Gagné, 1985) in which the rule is applied. Compared to item learning, this type of learning is closer to what is generally investigated by SLA researchers. As an example of system learning, imagine that a learner knows an explicit rule, such as "Use ***the*** when the context makes a noun known to the reader/listener." Then, an item is presented such as "Its front tire was flat, so __ bicycle was unusable." The learner has to fill in the correct article. The learner has no memory of the correct answer, so he or she must apply the rule to respond correctly. The second type of learning is *item learning*, which is the type of learning generally studied in the educational psychology research on feedback timing. An item has been learned if the learner can correctly respond to it after a previous exposure to the item. If the learner has received feedback on the item that included the correct response, item learning may be the same as memorizing the correct response. This is a type of *verbal information learning* (Gagné, 1985). For example, imagine that that the same item, "Its front tire was flat, so __ bicycle was unusable," is presented, followed by the correct answer. If the *item* has been learned, the next time the learner sees this exact item, he or she may simply retrieve the correct answer from memory.

Finally, I will use the terms *error correction* and *reinforcement of a correct response* to differentiate between two possible feedback conditions. In both cases, feedback is provided to the learner on his or her response to a test item by providing the correct answer. *Error correction* is the case in which a learner has answered an item incorrectly. For example, if a learner is presented with the item "Its front tire was flat, so __ bicycle was unusable" and he or she chooses the response "a," the feedback would indicate that the correct answer is "the" and would be an instance of error correction. *Reinforcement of a correct response* is the case in which the learner has answered correctly. For example, if a learner is presented with the same item and chooses the correct answer, "the," then the same feedback (i.e., that the correct answer is "the") would be an instance of the reinforcement of a correct response. Both of these types of feedback are subtypes of knowledge of correct response feedback, and both are derived from the behaviorist paradigm (e.g., Skinner, 1968).

1.3 Importance of the Current Study

The current study is relevant to SLA researchers, language teachers, and instructional designers for several reasons. Beginning with SLA researchers, the current study provides precise language for sorting out the varied definitions of *immediate* and *delayed*, as they are used in both SLA research and research in related fields. The terms *item by item* and *end of test* are defined above for use within this dissertation and beyond. In addition, SLA researchers can gain further information by using conditional probabilities to analyze data, breaking it down into categories that reveal whether correct responses have been reinforced and whether errors have been corrected. This is commonly done in educational psychology research, but has not yet gained a foothold in SLA research. Finally, researchers will be interested in the evidence provided in the following study in support of Doughty's (2001) theory of a cognitive processing

window, within which feedback is most usefully provided, and in support of an SLA attention-based theory (e.g., Gass, 1997; Pica, 1994; Schmidt, 1995).

Teachers and instructional designers will be interested in the results of the current study for other reasons. First, the results provide a starting point for determining when to more effectively and efficiently provide which type of feedback to learners. In addition, the results may help teachers make an informed decision about which commercial CALL products will be most useful to their students.

1.4 Overview

The rest of this dissertation is organized as follows. Chapter 2 explains the theoretical background for the current study and gives an overview of relevant literature in CALL, SLA, language assessment, other language-related fields, and educational psychology. The research questions and predictions for the current study are at the end of Chapter 2. Chapter 3 provides details on the method that I used to investigate the research questions, including the participants, materials, procedure, and analysis. Chapter 4 presents the results, followed by a discussion in Chapter 5. The conclusion is in Chapter 6, followed by the appendices and references.

CHAPTER 2: THEORETICAL BACKGROUND AND LITERATURE REVIEW

Feedback timing is an underresearched area in SLA, perhaps because few theories explicitly or implicitly bear on the question. Therefore, I begin the theoretical discussion by considering which theories address this topic generally, without a concern for the applicability of the theories to the current study. I then look more closely at the interaction approach, which gives a psycholinguistic account for the superiority of immediate feedback. It is this psycholinguistic account, rather than the interaction approach itself, that I will apply to the research study described below. After the SLA theoretical background, I review the language-related empirical research on feedback timing. Next, I examine selected theories from educational psychology that make predictions about the relative effectiveness of immediate and delayed feedback, and I then give an overview of the related research in that field.

After addressing the topic of feedback timing, I move to a discussion of the effectiveness of feedback with and without metalinguistic information. I present the theoretical position of the interaction approach, then language-related research on this topic. Although educational psychology theories do not directly address the topic of metalinguistic feedback, I review theories and research related to informational feedback. Finally, I present the research questions and predictions for the current study.

2.1 SLA Theories and Feedback Timing

I temporarily put aside the problem of whether various SLA theories apply to a situation in which a human answers multiple-choice questions and the computer provides feedback. For the moment, I focus on what the theories can contribute to a general discussion of feedback timing. In fact, most mainstream SLA theories have little to say about when feedback should be provided. Feedback (i.e., negative evidence) plays little, if any, role in theories based on

Universal Grammar (e.g., Cook, 1989; Schwartz & Gubala-Ryzak, 1992; but c.f. White, 1991). In theories in which feedback may play a limited role, such as processability theory (e.g., Pienemann, 1998), the ideal timing of feedback in relation to the error is not specified by the theory, either explicitly or implicitly. Similarly, in purely usage-based approaches (e.g., N. Ellis, 2006, 2008), feedback may play a limited role (without an ideal feedback timing specified by the theory), although Ellis incorporates ideas from the interaction approach to acknowledge feedback and focus-on-form as a means of drawing learners' attention to features of language. I discuss the interaction approach itself below. In skill acquisition theories (e.g., Dekeyser, 1997, 2007), the requirement for feedback to be *timely* is implicit, but no mention is made of why timeliness is important or how to define it (but see Hartshorn et al., 2010, for 24-hour-delayed feedback as one interpretation of *timely*).

In sociocultural theory, feedback is adjusted online (i.e., during interaction) to the needs of the individual learner (Aljaafreh & Lantolf, 1994, p. 466), but this may not imply that feedback immediately follows an error. Rather, researchers have variously operationalized feedback as provided immediately after an error is made, as in the Computerized Dynamic Assessment of Language Proficiency (http://calper.la.psu.edu), delayed until the completion of an oral narrative (e.g., Lantolf, 2008; Poehner, 2008), and delayed longer, as in individual tutoring on a piece of writing that has been completed at an unspecified earlier time (e.g., Aljaafreh & Lantolf, 1994).[1] Thus, SLA researchers have not interpreted sociocultural theory as specifying a particular feedback timing as better than another.

The interaction approach is the SLA theory that most clearly and explicitly specifies the

[1] Arguably, some of the feedback provided in the study by Aljaafreh and Lantolf (1994) came immediately following failed attempts to correct errors. However, from another perspective, the initial errors were made during the writing, not during the later process of reading aloud.

ideal feedback timing. In this approach, negotiation for meaning during a conversation between an L1 and L2 speaker causes interactional adjustments to the speech of the L1 speaker, facilitating language learning on the part of the L2 speaker (Long, 1996). This approach has been formulated in terms of language use in conversation and posits a role for immediate feedback on errors (e.g., Gass, 2010b; Long, 1996). Doughty (2001), for example, wrote the following.

> If the verbatim format of recent speech remains activated in memory and available for use in subsequent utterance formulation, this can be taken to be an important cognitive underpinning for facilitating the opportunity to make cognitive comparisons. With regard to the timing of the information to be compared, the most efficient means to promoting cognitive comparison would seem to be provision of *immediately contingent recasts*. (p. 253; emphasis added)

I interpret *immediately contingent recasts* in a conversational setting as analogous in timing to item-by-item feedback on a multiple-choice test.

The reasoning behind the interaction approach, as described above, is similar to that of the direct contrast hypothesis, which states that in child first language acquisition, negative evidence, typically provided as recasts by adults, allows a child to retreat from overgeneralizations in his or her developing grammar (e.g., Saxton, Backley, & Gallaway, 2005; Saxton, 1997). According to this hypothesis, the immediate contingency of the recast on the error is what leads to the effectiveness of recasts. This hypothesis might also be applicable to L2 learning, with the prediction that recasts will be more effective for L2 speakers than models of accurate speech provided at a time somewhat removed from an error. However, after adapting the hypothesis to L2 learning, it is equivalent to the interaction approach in terms of its predictions for the most effective feedback timing. In addition, no clear psycholinguistic basis

has been proposed for the direct contrast hypothesis, which makes it difficult to argue for its applicability in the current study. Therefore, I do not consider it further here.

Now that I have established that the interaction approach is the SLA theory that most directly deals with the topic of feedback timing, I turn to the question of its applicability to the current study. This approach has been extended to computer-mediated interaction between learners and expert language users, such as video and audio chats (e.g., Yanguas, 2010) and text chats (e.g., Lee, 2008). Moving even further from the approach's origins in face-to-face conversational interaction, Heift (2004) applied the approach to the interaction between a learner, who inputs a sentence into a computer, and a natural-language processing program, which outputs metalinguistic feedback. Indeed, Chapelle (2003) has proposed that the interaction approach could be extended to the input-enhancing interaction that occurs when learners interact with computers. She gave the example of a learner clicking a hyperlinked word to get a definition when reading a passage. From this perspective, the interaction approach may also be extended to the interaction between a human and a computer in the context of feedback on multiple-choice questions, as explained below.

Consider this scenario: A learner reads a sentence with a blank on a computer screen. From a list of answer choices, he or she drags and drops a phrase into the blank. After clicking a "submit" button, the learner is presented with feedback that indicates (a) that the selected response was incorrect, (b) the correct answer, embedded in the original sentence, and (c) a rule explaining why it is correct. Part (a) of the feedback serves as an explicit indication that the learner has made an error. Part (b) provides information equivalent to that in a recast in spoken interaction, and Part (c) is metalinguistic information. Certainly this type of feedback qualifies as enhanced input. Therefore, one might argue that the interaction approach can be applied to this

situation.

Admittedly, the argument above goes quite far afield from the origin of the interaction approach. However, one aspect of the approach in particular may be more readily extendable. I next examine the basis within the interaction approach for the claim that immediate feedback is superior to delayed feedback. This claim is based on the idea of attention to contrast. That is, a language learner who produces a nontargetlike utterance will contrast his or her utterance with a targetlike recast that another speaker produces in response (e.g., Goo & Mackey, 2013; Long, 2007). According to Gass (2010),

> Attention alone is not sufficient. A contrast must be attended to, or in SLA parlance, a gap must be noticed. And conversation provides a forum for the contrast to be detected, especially when the erroneous form and a correct one are in immediate juxtaposition. (p. 230)

I argue that multiple-choice questions with correct answers given as feedback also provide a forum for the contrast to be detected, especially when the feedback is provided on an item-by-item basis. Similarly, Long (1996) stated that "[n]egative feedback of this type (i.e., in the form of implicit correction immediately following an ungrammatical learner utterance) is potentially of special utility because it occurs at a moment in a conversation when the NNS is likely to be attending to see if a message got across, and to assess its effect on the interlocutor" (p. 429). Although Long referred to a spoken conversation, it follows that when negative feedback comes immediately after a learner responds to a multiple-choice question, the learner is likely to be attending to the feedback to find out whether he or she answered the question correctly. Conversely, when the negative feedback comes at a later time, the learner is less likely to be attending.

Of course, conversation typically occurs in a face-to-face setting between two (or more) people, and a multiple-choice computer-based test is taken by one person looking at a computer screen. In addition, feedback provided in the two contexts may be perceived differently by learners for affective reasons. For example, learners may perceive the feedback provided by a computer as less face-threatening than that provided by another person. These differences are potentially important in determining how attention is focused. However, little research exists to clarify how attention to contrast in a written, computer-based format may differ from attention to contrast in spoken interaction. Therefore, I proceed here under the tentative assumption that item-by-item negative feedback on a multiple-choice test on a computer screen will be more effective in drawing a learner's attention to contrast than feedback that comes later.

Another potentially important difference between these contexts is that learners typically focus on meaning during conversation, while they may be more focused on form when they are taking a multiple-choice test. This may prime learners to be more attentive to contrasts in a computer-based test than during conversation.

The attention to contrast described by both Long (1996) and Gass (2010) has a psycholinguistic basis in what Doughty (2001) termed the *cognitive processing window*. According to Doughty (2001), negative feedback must be immediate to be effective because learners' ability to perform cognitive comparisons is limited by working memory (p. 225). Negative feedback that focuses on form can most usefully be provided within this cognitive window, which may last up to 40 seconds (p. 227). On the other hand, Doughty and Long (2003, p. 65) more recently claimed that this window is not well understood and that it may not even depend on working memory, implying that the previous estimate of the duration of the window may not be accurate. However, given no alternative hypothesis as to why the window should be

limited, I proceed here under the assumption that working memory constrains the window.

While Doughty (2001) cites the working memory model of Cowan (1995), the model of Baddeley (2003) is also widely used in psycholinguistic research, and various other models exist. When considering the capacity of working memory, which model is referenced may make little difference. Here, it suffices to say that, according to Baddeley (2003), working memory is "a limited capacity system, which temporarily maintains and stores information, supports human thought processes by providing an interface between perception, long-term memory and action" (p. 829). Similarly, according to Cowan (2005), working memory is "the set of processes that hold a limited amount of information in a readily accessible state for use in an active task" (p. 39). Note that in both definitions, the capacity of working memory is limited.

Information (in this case, an error made by a learner) activated in working memory is likely to remain activated until item-by-item feedback is provided. Information is unlikely to remain activated until end-of-test feedback is provided, not only because of the time delay, but also because of the intervention of other items. Of course, in the end-of-test feedback condition, upon receiving feedback, a learner could reactivate an error in working memory by repeating the processing that occurred when he or she initially saw the item, but this is not guaranteed to occur.

The preceding discussion of feedback has focused on errors. However, receiving feedback on an error is only one of two possible outcomes for a learner answering a multiple-choice question. The other possibility is that the learner answers the question correctly and receives feedback that reinforces the correctness of the response. In this case, when the learner cognitively compares his or her response to the provided correct response, no difference is found. Although the literature on the cognitive processing window does not directly address this case, the theory can be logically extended to predict that the cognitive comparison in this case should

also be more effective when the feedback comes within the cognitive processing window, as limited by working memory. *Effective* should be understood here to mean that the learner continues to answer the item correctly in the future, rather than (erroneously) changing his or her answer, in the case of item learning, and to mean that the learner correctly answers novel items that follow the same rule, in the case of system learning.

2.2 Language Research on Feedback Timing

In the following sections, I summarize the previous research on feedback timing in SLA, CALL, language assessment, and other language-related areas. A summary is shown in Table 1.*

2.2.1 SLA research

Within SLA, most researchers have found no difference between item-by-item and end-of-test feedback. To study the effects of oral feedback timing, Sheen (2012) had adult ESL students perform a narration task using the past tense, and she provided explicit, metalinguistic feedback. In one condition, the feedback was provided immediately after a student made an error (analogous to item-by-item feedback), while in the other condition, the feedback was delayed until the end of the task (analogous to end-of-test feedback). No significant differences were found between the feedback groups on either a posttest or a delayed posttest. As Sheen noted, the end-of-test feedback took more time to provide than the item-by-item feedback, making the conditions difficult to directly compare. Quinn (2013) similarly tested the effects of providing item-by-item and end-of-test oral feedback on oral tasks, in this case, related to English passive constructions. He found no significant difference between the feedback conditions, possibly because the learners in the delayed group were asked to repeat the task in which the error occurred before feedback was provided, making the feedback essentially item-by-item.

Table 1: Summary of Language Learning Studies on Feedback Timing Studies

	More effective feedback timing
Brosvic, Epstein, Dihoff, and Cook (2006a); Brosvic, Epstein, Dihoff, and Cook, (2006b); Opitz, Ferdinand, and Mecklinger (2011); Schroth and Lund (1993)	IBI more effective than delay shorter than EOT.
Nagata (1996)	IBI more effective than EOT
Schroth and Lund (1993)	Delay shorter than EOT more effective than IBI.
Aubrey and Shintani (2014)	Delay shorter than EOT more effective than delay longer than EOT.
Goda (2004); Henshaw (2011); Quinn (2013); Sheen (2012)	No difference between IBI and EOT.
Lavolette, Polio, and Kahng (2013)	No difference between EOT and longer delay.
Lai, Fei, and Roots (2008); Sakai (2004)	IBI recasts noticed more often than EOT recasts/models.

Note. IBI = item by item; EOT = end of test.

Without looking at the effectiveness of feedback on learning, SLA researchers have also examined the effects of the immediate contingency of recasts on the noticing of a gap between the learner's production and the recast. Lai, Fei, and Roots (2008) looked at the effects of the timing on ESL learners' noticing of recasts in typed CMC interactions in a laboratory setting. Noticing was measured using the learners' reports in stimulated recalls or think-aloud protocols. The researchers found that recasts were noticed more often when they immediately followed an error or were separated from the error by only material not related to the content as compared to recasts that were separated from the relevant error by additional content. Sakai (2004) also looked at the effects of contingency on noticing, but in spoken interaction in a laboratory setting. He compared item-by-item recasts and delayed "models," which were similar to recasts, but provided a few minutes later. For the EFL learners in the study, the item-by-item recasts were more effective for noticing, as measured using a stimulated recall.

Loewen (2004) examined the timing of recasts in relationship to uptake, or the response of a learner to a recast. One reason to examine uptake is that it may be an indication of noticing the gap between an erroneous and correct form. Loewen did not define the term *immediate* in the study, but I surmise from the examples given that recasts were immediate if they were produced in the first turn following an erroneous utterance by a learner, analogous to item-by-item feedback. Other recasts were classified as delayed, or in Loewen's terminology, *deferred.* The results showed that for the adult ESL learners in the study, item-by-item recasts were five times as likely to be followed by uptake than recasts produced during later turns. However, as Loewen pointed out, the learners were often given no opportunity for uptake when the recasts were delayed.

2.2.2 CALL research

To my knowledge, only six studies have examined the effect of feedback timing on language learning in a CALL setting (Aubrey & Shintani, 2014; Dabrowski, LeLoup, & MacDonald, 2013; Goda, 2004; Henshaw, 2011; Lavolette et al., 2013; Nagata, 1996). Two of them found significant differences based on feedback timing. The first study that revealed a difference was that of Nagata (1996), who did not design the study with the intention of examining a difference based on feedback timing. Rather, she was interested in the difference between CALL and non-CALL teaching practices. Therefore, the feedback timing factor is confounded with how feedback was provided, preventing any firm conclusions. In Nagata's study, half of the undergraduate participants received item-by-item feedback from a computer program on their usage of Japanese particles, while the other half received the similar (although less individualized) feedback after completing the entire exercise on paper. Both groups participated in the study during their normal class period. The results of immediate and delayed posttests showed that the computerized item-by-item feedback was significantly more effective than the paper-based end-of-test feedback. Second, Aubrey and Shintani (2014) found that feedback provided at a longer delay than item-by-item feedback (but shorter than end-of-test) was more effective than feedback provided at a delay longer than end of test. In their study, English learners in Japan wrote essays using Google Docs, an application that allows multiple people to synchronously edit a document. The researchers provided feedback by inserting comments into the margin of the document, providing only the correct form, and only targeting hypothetical conditionals. One group (synchronous feedback) received feedback while they were writing, generally after they had finished writing the targeted sentence and before they had finished writing the following sentence. A second group (asynchronous feedback) received

feedback a few minutes after they finished writing the entire essay. A third group (control group) received no feedback. Both of the treatment groups performed significantly better than the control group on the immediate posttest, but only the synchronous feedback group outperformed the control group on the delayed posttest.

Four CALL studies showed no difference between various feedback timings. Goda (2004) found no effect of differing feedback timings (item-by-item and end-of-test) on EFL students' scores on TOEFL structure questions. Two test versions were used, with the questions on the treatment test being different from those on the two posttests. However, the questions on the treatment were randomly chosen, so the structures on them may not have been relevant to the structures on the posttest. Dabrowski, LeLoup, and MacDonald (2013) looked at the effects of instructor feedback, provided to one group after a delay, and computer feedback, provided immediately to the other group. Note the confound between the variable of immediate/delayed feedback and that of computer/instructor feedback, preventing conclusions about the relative efficacy of the immediate or delayed feedback in this study. The computer feedback was provided using My Spanish Lab, but it is not clear whether the feedback was item-by-item or end-of-test (which prevents the inclusion of this study in Table 1). No differences were found in the chapter test scores for the two groups. In L2 writing, Lavolette, Polio, and Kahng (2013) examined the effects of feedback timing using Criterion, a program provided by ETS that uses natural language processing to give feedback on ESL students' TOEFL-style essays. One group received feedback immediately upon completing an essay (analogous to end-of-test feedback), while the other received the feedback a week later. The timing of the feedback did not affect the students' responses to the feedback.

The CALL study that is most similar to the current study is that of Henshaw (2011).

Using processing instruction, Henshaw examined the effects of feedback timing on English native speakers' learning of the Spanish subjunctive. The learners in her study were first screened for previous knowledge of the subjunctive using a pretest, then received explicit instruction on this grammar structure. Next, they answered multiple-choice questions testing their recognition and interpretation of the subjunctive, with one group receiving item-by-item feedback, a second group receiving end-of-test feedback, a third group receiving feedback 24 hours after taking the test, and a fourth group receiving no feedback. The feedback was an indication of whether they had answered correctly or incorrectly plus a metalinguistic explanation. A week later, the students took a posttest that included the items that the students had previously seen and new items. No significant differences on old or new items were found among the groups who received feedback, although all feedback groups outperformed the no-feedback group.

2.2.3 Language assessment research

Much of the literature on assessment involving feedback is far removed from the context of the current study. In fact, little empirical research has addressed the question of whether assessments that provide feedback can affect student learning. Of the various types of assessments, three major types provide learners feedback on their responses: formative, dynamic, and diagnostic assessment. To my knowledge, only investigations of diagnostic assessment have considered the timing of the feedback, all of which were reported in a monograph by Alderson (2005).

Alderson (2005) listed some suggested characteristics of diagnostic tests, including that "[d]iagnostic tests provide immediate results, or results as little delayed as possible after test-taking." (p. 11). However, he provides little support for this suggestion. He also reported on two

unpublished studies (Floropoulou, 2002; Yang, 2003) that asked users about their preferences for receiving item-by-item feedback on an early version of DIALANG (http://www.lancaster.ac.uk/researchenterprise/dialang/about), which is a diagnostic test of language proficiency for European languages. Alderson does not mention which languages were tested in either study. All users got end-of-test feedback in the form of a review, and the users had the option of leaving on a default option of item-by-item feedback or turning it off. In both studies, the users' behavior varied as to whether they chose item-by-item feedback. Floropoulou (2002) asked six users about their reasons for their feedback timing preferences. She quoted two users who preferred no item-by-item feedback . One preferred no item-by-item feedback because getting the feedback during the test might influence his or her thinking about the test. The other believed that it allowed him or her to avoid the encouraging effect of getting answers right and the discouraging effect of getting them wrong. One user who preferred item-by-item feedback was also quoted, mentioning only that it is good to know immediately whether your answers are right or wrong.

The second relevant study that Alderson (2005) reported on was an unpublished MA thesis by Yang (2003), who studied 13 users of DIALANG. The users' behavior was somewhat different from that in Floropoulou's (2002) study. While four of Yang's users chose to receive the item-by-item feedback throughout the test and six chose not to get it on any items, two users turned off item-by-item feedback after feeling discouraged due to answering the first four items wrong, and one student selectively turned on item-by-item feedback for items that she was not confident of having answered correctly. Those who always got item-by-item feedback chose it because they wanted to know how they were doing. Those who never got item-by-item feedback chose against it because they wanted to finish the test quickly and because knowing that they got

questions wrong would be demotivating. In addition, “one student said that, since she could not change the answer if it was wrong, she saw no point in using immediate feedback” (Alderson, 2005, p. 215). All users indicated that they found the end-of-test review useful.

Overall, the assessment literature includes little mention of the issue of feedback timing, and even less of the effectiveness of one feedback timing compared to another. One reason for this is that the purpose of assessments is most often to assess learning, rather than to promote it. Another reason may be the potential for item-by-item feedback on items at the beginning of a test to affect students’ answers later on the test. However, this concern would not prevent an investigation into the differences between end-of-test feedback and feedback provided at a longer delay.

2.2.4 Other language-related research

Because little SLA, CALL, or language assessment research has looked at the relative effectiveness of varied feedback timing, as reviewed above, I also review here studies of vocabulary and artificial grammar learning that were undertaken from non-SLA perspectives. The studies of language learning that have examined the variable of timing are summarized in Table 1.

Several groups of researchers working from non-SLA perspectives have studied language learning. Some findings showed an advantage for item-by-item feedback over feedback with a delay shorter than end-of-test. For example, in neuroscience, Opitz, Ferdinand, and Mecklinger (2011) found that participants who received item-by-item feedback while learning an artificial grammar responded correctly to significantly more items than participants who received item-by-item feedback delayed by 1 second. This is interesting in light of the fact that a 1-second delay is likely to still be within the cognitive processing window. In educational psychology, two studies

that looked at students learning form-meaning mappings found that item-by-item feedback was more effective than end-of-test feedback or longer feedback delays. In a laboratory, Brosvic, Epstein, Dihoff, and Cook (2006a) tested undergraduate students on the definitions of Esperanto words. The treatments varied on the timing of feedback (none, item-by-item, end-of-test, or delayed by 24 hours) and whether the student could select another response if the first selection was incorrect. The tests and feedback were provided using special paper forms, which had a coating over the answer choices that was removed by the participants to reveal the feedback for the chosen response, or feedback was provided by an assistant who held up an index card. The students who received item-by-item feedback outperformed the students in the other groups on all posttest measures, with the students who could select multiple responses outperforming those who could only select one. A similar study that manipulated the length of delay and the number of questions answered until feedback was provided had similar results: namely, that receiving item-by-item feedback was most effective (Brosvic et al., 2006b).

A study by Schroth and Lund (1993) showed results both supporting item-by-item feedback as more effective than a delay shorter than end-of-test and the reverse; that is, other results supported a delay shorter than end-of-test as more effective than item-by-item feedback. In a laboratory, undergraduate students learned an artificial grammar that consisted of patterns of letters that followed simple rules. The materials were presented on paper cards, and the experimenter provided the feedback verbally. The results showed that the participants who received item-by-item feedback learned the patterns more quickly than the participants who received feedback delayed by 10, 20, or 30 seconds. However, the participants who received the 30-second delayed feedback were more accurate at transferring what they had learned to a new task on both an immediate and a delayed posttest. Note, however, that the results are

questionable, given that the participants in item-by-item feedback group reached criterion on the training task more quickly than the participants in the delayed groups, resulting in fewer practice trials.

2.3 Educational Psychology Theories of Feedback Timing

Despite the dearth of SLA theories that address feedback timing, many educational psychology theories have been proposed to explain why immediate or delayed feedback is superior for learning. Perhaps most famously, behaviorist psychology contends that feedback (or *reinforcement*, in behaviorist terms) must be immediately contingent upon a response to have a learning (or *conditioning*) effect. In fact, in a discussion of the effects of a teacher's feedback on a student's learning of mathematics, Skinner (1968) wrote, "It can easily be demonstrated that, unless explicit mediating behavior has been set up, the lapse of only a few seconds between response and reinforcement destroys most of the effect" (p. 16). The early work of behavioral psychologists like Skinner continues today in the form of behavior analysis. Here, I specifically consider relational frame theory (e.g., Hayes, Barnes-Holmes, & Roche, 2001). This theory extends behavioral principles to verbal behavior, based on the human ability to create links or relations between stimuli. However, a major change from Skinner's theory is that feedback need not be immediate to be effective (Barnes, 1996). Thus, from a behavior analytic perspective, it is not clear whether item-by-item feedback is predicted to be more effective than end-of-test feedback for either reinforcement of correct responses or error correction, and I do not investigate this type of theory any further.

Next, I concentrate on three theories that predict that delayed feedback is superior: the interference-perseveration hypothesis (Kulhavy & Anderson, 1972), the dual-trace hypothesis (Kulik & Kulik, 1988), and an attention-based account (Phye & Andre, 1989).

To understand the predictions of the three theories, it is helpful to keep in mind the context for which they were developed. This context prototypically has two phases: learning and testing. In the learning phase, the participant answers questions and gets feedback of some sort. In the testing phase, the participant answers the same questions that were presented in the learning phase but does not get feedback. In addition, the three hypotheses are only designed to deal with item learning, not with system learning. That is, the hypotheses predict learning when the items are the same in the learning and testing phases. They do not make predictions about learning the rules associated with the treatment items and being able to extend those rules to new items on a posttest. Finally, note that the terms *immediate* and *delayed* have been interpreted in various ways in the literature, so it is not possible to be more precise here.

The interference-perseveration hypothesis (Clariana, Wagner, & Roher Murphy, 2000; Kulhavy & Anderson, 1972; Smith & Kimball, 2010) predicts differing results for immediate and delayed feedback based on whether a test-taker initially responds correctly or incorrectly. The hypothesis indicates that delayed feedback is superior to immediate feedback for the correction of errors because the time that passes between the incorrect response and the feedback during the learning phase allows the memory trace of the incorrect response to fade. Then, the memory trace of the correct response (provided in the feedback) is stronger than that of the incorrect response. This results in a greater number of correct responses during the testing phase. In the case of a correct response during the learning phase, the interference-perseveration hypothesis predicts that immediate and delayed feedback will produce similar results for reinforcement of the correct response, as measured during the testing phase. That is, because the participant responded correctly during the learning phase, no incorrect response exists or needs to be forgotten; therefore, immediate and delayed feedback are predicted to have similar effects.

Interestingly, this hypothesis and the cognitive processing window lead to exactly opposite predictions for the effectiveness of item-by-item feedback based on similar ideas about the limited capacity of memory.

The dual-trace hypothesis (Clariana et al., 2000; Glover, 1989; Kulik & Kulik, 1988; Rankin & Trepper, 1978) indicates that delayed feedback is more effective because it gives the participant two encoding opportunities. In other words, each time a participant encounters an item provides one encoding opportunity. If the feedback is immediate, the question and the feedback are fused into one encoding opportunity, but if the feedback is delayed, the feedback acts as a second encoding opportunity. Thus, the dual-trace hypothesis predicts that delayed feedback is superior for both the correction of errors and the reinforcement of correct responses because of the additional encoding opportunity that delayed feedback provides.

A final account for why delayed feedback may be superior to immediate feedback for item learning is based on attention. When students are presented with feedback in a normal classroom situation, they choose whether and how long to attend to feedback on a given item. According to the attention-based account (Kulhavy & Anderson, 1972; Phye & Andre, 1989), learners pay more attention to and therefore spend more time studying 24-hour-delayed feedback compared to end-of-test feedback, making the 24-hour-delayed feedback more effective. However, it is not clear what this theory suggests for item-by-item feedback compared to end-of-test feedback, so I will not consider it further.

2.4 Educational Psychology Research on Feedback Timing

In educational psychology and related fields, the study of feedback timing has a history of nearly a century. A summary of the results of relevant meta-analyses is shown in Table 2, and a summary of the results of individual studies is shown in Table 3. Patterns are difficult to distill

from the varied findings, but evidence is available to support both item-by-item and end-of-test as the more effective feedback timing for item learning.

2.4.1 Review articles

Researchers have performed at least four meta-analyses and one other review article that examined the variable of timing. However, the results of these studies (with the exception of that of Bangert-Drowns, Kulik, Kulik, & Morgan, 1991) should be interpreted with caution because the meta-analyzers used the definitions of *immediate* and *delayed* of the researchers of each individual study, which is not consistent from one study to another. This makes the results of the meta-analyses problematic at best. With that caveat in mind, I briefly review the results of the four meta-analyses and one other review article.

Table 2: Summary of Educational Psychology Review Articles on Feedback Timing

Study	More effective feedback timing
Kulik and Kulik (1988)	Most classroom studies found IBI or EOT feedback more effective than feedback that was provided later. Most laboratory studies found later feedback more effective than IBI or EOT feedback.
Kulik and Kulik (1988)	16 studies: more effective learning with IBI feedback. 11 studies: more effective learning with EOT or IBI feedback delayed by seconds.
Bangert-Drowns, Kulik, Kulik, and Morgan (1991)	EOT had larger effect size than longer delay; longer delay had larger effect size than IBI.
Azevedo and Bernard (1995)	Positive effect for immediate over delayed in computer-aided instruction.
Hattie and Timperley (2007)	5 meta-analyses: delayed more effective; 8 meta-analyses: immediate more effective.
Jaehnig and Miller (2007)	Both IBI and feedback provided at various delays are effective in programmed instruction.

Note. IBI = item by item; EOT = end of test.

Table 3: Summary of Individual Educational Psychology and Other Studies on Feedback Timing

Study	More effective feedback timing
English and Kinzer (1966); Kulhavy and Anderson (1972); Metcalfe et al. (2009), Experiment 1; Surber and Anderson (1975); Webb, Stock, and McCarthy (1994), Experiment 1	Delay longer than EOT more effective than EOT.
Butler et al. (2007); King, Young, and Behnke (2000)	Delay longer than EOT more effective than IBI.
Sturges (1978)	Delay longer than EOT more effective than delay shorter than EOT; EOT more effective than delay shorter than EOT
Guzmán-Muñoz and Johnson (2008); Rankin & Trepper (1978); Schooler and Anderson (1990), Experiments 2 and 3	EOT more effective than IBI.
Rankin & Trepper (1978); Schroth (1992); Schroth (1995); Smith and Kimball (2010)	Delay shorter than EOT more effective than IBI.
Brosvic and Epstein (2007); Dihoff, Brosvic, Epstein, and Cook (2004); Lin, Lai, and Chuang (2013)	IBI more effective than EOT.
Brosvic and Epstein (2007); Dihoff, Brosvic, Epstein, and Cook (2004); King, Young, and Behnke (2000)	IBI more effective than delay longer than EOT.
Schroth (1992); Schroth (1995)	IBI more effective than delay shorter than EOT
Clariana et al. (2000); El Saadawi et al., (2008); Gaynor (1981); Lewis and Anderson (1985), Experiments 2 and 3; Schooler and Anderson (1990), Experiment 1; Smith and Kimball (2010); Surber and Anderson (1975); Van der Kleij, Eggen, Timmers, and Veldkamp (2012)	No difference between IBI and EOT/various delays.
Metcalfe et al. (2009), Experiment 2; Webb et al. (1994), Experiment 2	No difference between EOT and longer delay.

Note. IBI = item by item; EOT = end of test.

In a classic meta-analysis of the effects of delayed and immediate feedback, Kulik and Kulik (1988) reviewed both classroom and laboratory studies on the learning of test content in a variety of subjects (e.g., chemistry, psychology, and math). Of the 11 classroom studies, 9 found that immediate feedback was significantly more effective than feedback that was provided later, either at the end of the test or a day to a week later, based on the amount of the materials that students retained during the original period of learning. In 13 of the 14 laboratory studies, the opposite result was found: The participants who received delayed feedback performed better than the participants who received immediate feedback. Note, however, that both Butler, Karpicke, and Roediger (2007) and Metcalfe, Kornell, and Finn (2009) claimed that the classroom versus laboratory distinction made by Kulik and Kulik (1988) was not the reason for the different findings, claiming instead that learner attention to the feedback was the key difference.

Several other meta-analyses have also examined feedback timing. Looking exclusively at computer-aided instruction, Azevedo and Bernard (1995) meta-analyzed 22 studies and found a positive effect for immediate over delayed feedback. In a meta-analysis of meta-analyses, Hattie and Timperley (2007) looked at the results of 74 meta-analyses of the effects of feedback on learning and found seemingly contradictory results regarding immediate versus delayed feedback. Five meta-analyses found delayed feedback more effective, with an effect size of 0.34, while eight meta-analyses found immediate feedback more effective, with an effect size of 0.24. However, as the authors argued, there is a key difference between conditions under which immediate and delayed feedback were beneficial: immediate feedback was beneficial for easier items, whereas delayed feedback was beneficial for more difficult items. The reason for this may be that more difficult items take longer to process, making the extra time before feedback is provided useful for learning, while this extra time is unhelpful for easier items.

Bangert-Drowns et al. (1991) conducted a meta-analysis in which they specified that they analyzed end-of-test and item-by-item feedback. Based on 40 studies, they found that end-of-test feedback produced larger effect sizes than longer delays and that longer delays had larger effect sizes than item-by-item feedback.

Finally, Jaehnig and Miller (2007) reviewed feedback types in programmed instruction. According to the authors, programmed instruction is a type of instruction grounded in behaviorist psychology in which a stimulus is presented to a learner, who has an opportunity to respond. The response is followed by feedback on the correctness of the response. Note that this definition applies to many studies in CALL and SLA as well as educational psychology, and indeed, studies such as those of Nagata (1993) and Rosa and Leow (2004) are included in the review. Although feedback in programmed instruction is defined as immediate (i.e., item-by-item), studies within the review examined various delays. While the authors did not perform a meta-analysis, they concluded that both item-by-item feedback and feedback provided at various delays are effective, with the caveat that the effects of delaying elaborated feedback (i.e., feedback that includes information beyond knowledge of the correct response) have not been fully explored.

2.4.2 Individual studies

The results of studies of feedback timing are varied, and the lack of systematicity in defining *immediate* and *delayed* makes the results challenging to synthesize. Further complicating the issue is the fact that an individual study may contain multiple experiments with differing findings, or even a single experiment whose results suggest an advantage for different feedback timings, based on different independent and dependent variables.

The studies summarized below have two factors in common. First, in nearly all of the

studies, the participants were learning using their L1s. Second, the posttests generally included the same items as the treatments, so any learning demonstrated in the studies is item learning. Exceptions to these two generalizations are noted. A summary of all studies is shown in Table 3.

2.4.2.1 *Delay longer than end-of-test more effective than end-of-test*

In an older classroom study, English and Kinzer (1966) tested undergraduate students on the content of articles using a multiple-choice test completed on paper. There were four treatment groups: The first received end-of-test feedback, and the others received feedback delayed by 1 hour, 2 days, or 1 week. The results showed that the 1-hour and 2-day delays were superior to the other conditions.

Kulhavy and Anderson (1972) tested high school students using multiple-choice questions completed on paper after they had studied material about psychology that provided end-of-test or 1-day delayed feedback. The authors found that the students who received the 1-day delayed feedback did significantly better on a delayed posttest (a week after the initial learning) than the students who received the end-of-test feedback. These results led to the original statement of the interference-perseveration hypothesis.

Metcalfe et al. (2009) studied the learning of L1 English vocabulary using one experiment with child participants and one with adult participants. In the first experiment, sixth-grade children played a computer game that taught them vocabulary. In the testing phase, a definition appeared, and the learner typed the corresponding word. The children received end-of-test feedback on some of the questions and 1- to 4-day delayed feedback on other questions. This experiment is noteworthy because the lag between the feedback and the test was controlled for in half of the participants, while the lag between the questions and the feedback was controlled in the other participants. For all participants, the 1- to 4-day delayed feedback was more effective

than the end-of-test feedback. The second experiment was similar to the first, but the results showed no difference between the two feedback timings, so its results are described below.

Surber and Anderson (1975) tested high school students on their comprehension of an article using multiple-choice questions on paper, providing them end-of-test or 1-day delayed feedback. The 1-day delayed feedback was more effective for error correction (although there was no significant difference in the feedback timings for reinforcement of correct responses).

Webb, Stock, and McCarthy (1994) tested undergraduate students on general knowledge, such as history and geography, using multiple-choice questions in two experiments. Both were conducted in a laboratory setting on computers. Half of the participants received end-of-test feedback, and half of the participants received feedback a day later. In Experiment 1, the 1-day delayed feedback was more effective than the end-of-test feedback. However, note that the lag between the feedback and posttest was shorter for the 1-day delayed feedback than for the end-of-test feedback. In addition, the participants who received 1-day delayed feedback had a significantly higher proportion of errors that they corrected from the training to the posttest than did those who received end-of-test feedback. Experiment 2 was similar to Experiment 1, but with slightly different results, so it is described below.

2.4.2.2 *Delay longer than end-of-test more effective than item-by-item*

Butler et al. (2007) tested the effects of item-by-item versus 10-minute and 1-day delayed feedback on the performance of undergraduate students on a multiple-choice reading comprehension test on computers in a laboratory. The reading passages were taken from study guides for standardized tests. Based on a constructed-response posttest, the researchers found that the students answered more questions correctly when they had the 1-day delayed feedback.

King, Young, and Behnke (2000) looked at the effects of feedback timing on speech

performance by undergraduates in a laboratory setting. Feedback was given to the participants during the speech performance (analogous to item-by-item feedback) or at a 1-day delay. The researchers found that the item-by-item feedback was significantly more effective than the 1-day delayed feedback in getting the participants to make more eye contact, which the researchers characterized as a task that required little processing. The 1-day delayed feedback was significantly more effective than the item-by-item feedback in getting the participants to lengthen their planned introductions, which the researchers characterized as a task that required more processing.

2.4.2.3 *Delay longer than end-of-test more effective than delay shorter than end-of-test; end-of-test more effective than delay shorter than end-of-test*

Sturges (1978) tested undergraduate students on the content of a psychology lecture using a computer. The test included both multiple-choice and short-answer items. Students received 2-second delayed, end-of-test, or 1-day delayed feedback. A posttest given 1 to 3 weeks later showed that the end-of-test and 1-day delayed feedback were significantly more effective than the 2-second delay.

2.4.2.4 *End-of-test more effective than item-by-item*

In a study by Guzmán-Muñoz and Johnson (2008), L1 and L2 Dutch undergraduate students participated in an computerized laboratory experiment in which they dragged Dutch city names to their correct locations on a map. One group viewed a completed map while they performed the task, another group received feedback after placing each city (item-by-item), and a third group received feedback after placing all cities on the map (end-of-test). The learners in the end-of-test group showed the greatest gains on both immediate and 1-week delayed posttests. Note that this task requires visual and spatial skills, making it different from the more traditional

test-like tasks in most of the other studies reviewed here. This may explain why end-of-test feedback was more effective.

Rankin and Trepper (1978) asked undergraduate and graduate students to complete a 10-item multiple-choice test on human sexuality on a computer, presumably in a laboratory. The participants were divided into groups that received item-by-item, 15-second delayed, and end-of-test feedback. A retention test was given 24 hours after the treatment. The 15-second delayed and end-of-test feedback were significantly more effective for retention of knowledge than was the item-by-item feedback. However, note that the participants did not take a pretest and that the total number of questions (10) was very small.

Schooler and Anderson (1990) conducted three experiments in which they taught novices how to use the LISP programming language. The results of the first experiment differed from those of the Experiments 2 and 3, so Experiment 1 is described in a different section below. In Experiments 2 and 3, a LISP tutor program was used to instruct the participants and provide feedback either as they were typing their solutions to problems (item-by-item) or after they submitted a complete solution (end-of-test). Note that the program prevented the participants from continuing to write an incorrect solution in the item-by-item condition, while the program allowed them to continue writing following an error in the end-of-test condition. As a posttest, the participants completed new problems a day later, but without feedback from the tutor. In Experiments 2 and 3, the end-of-test feedback was more effective than the item-by-item feedback in terms of the number of errors that the participants made on the posttest and in the total time they required to complete the posttest.

2.4.2.5 *Delay shorter than end-of-test more effective than item-by-item*

As described above, Rankin and Trepper (1978) found that 15-second-delayed and end-

of-test feedback were significantly more effective for retention of knowledge than was item-by-item feedback.

Schroth conducted two similar studies (1992, 1995) in a laboratory setting in which the experimenter presented undergraduate students with cards and asked them to decide if they fit a concept or not. Because the students were not told the concept in advance, they had to determine the concept based on the feedback. In the first experiment, the feedback was provided item-by-item or at a 10-, 20-, or 30-second delay. Cards were presented until the student correctly responded to 9 out of 10 cards in a row or until 100 cards had been presented. A week later, a transfer test was administered that used a new concept and with item-by-item feedback provided to all participants. The number of trials to criterion during the treatment was significantly lower for the item-by-item group than for any of the delay groups. Note that this means that the item-by-item group had less practice that the delay groups. In the retention test, the 30-second delay group reached criterion faster than any of the other groups.

In Schroth's 1995 study, two further experiments were conducted under conditions similar to those of the 1992 study. In the first experiment of the newer study, instead of the 20-second-delayed feedback, feedback was provided at a delay that randomly varied between 10 and 30 seconds. An immediate transfer test was added after the training, and in both the immediate and delayed transfer tests, the feedback was provided using the same timing as that during the treatment. The results showed that the item-by-item group needed significantly fewer trials to criterion on the treatment than the other groups. On both the immediate and delayed transfer tests, the varied-delay group was the fastest to criterion. The second experiment was similar, but item-by-item feedback was provided to all participants during the two transfer tests. The results of the trials to criterion on the treatment task were the same as those for the first experiment. The

results of the immediate transfer test showed that the 30-second-delayed and variably delayed feedback groups achieved criterion significantly faster than the other two groups. On the delayed transfer test, the variably delayed feedback group was significantly faster than the other groups.

Smith and Kimball (2010) investigated the effects of varied feedback timing on participants' learning of trivia facts using a short-answer test in two experiments conducted on computers in a lab. In both experiments, undergraduate students received item-by-item or 8-minute delayed feedback. That is, feedback in the delayed condition was mixed in with the questions. A posttest was administered a week after the treatment. In the first experiment, the 8-minute delayed feedback was more effective than item-by-item feedback for reinforcement of correct responses, but there was no difference between the conditions for error correction.

2.4.2.6 *Item-by-item more effective than end-of test*

Brosvic, Epstein, and colleagues conducted two studies that found that item-by-item feedback is superior to end-of-test feedback on multiple-choice tests presented on paper. Dihoff, Brosvic, Epstein, and Cook (2004) conducted a classroom experiment in which undergraduate students were provided with item-by-item, end-of-test, or 1-day delayed feedback. The group that received item-by-item feedback did significantly better on the posttest (2 weeks later) than the other groups. Brosvic and Epstein (2007) found that undergraduate students in introductory psychology courses performed better on several posttest measures of learning, including delayed posttests 3, 6, 9, and 12 months after treatment, when they had been trained using item-by-item feedback as opposed to end-of-test feedback, 1-day delayed feedback, or no feedback.

In a computer science class, Lin, Lai, and Chuang (2013) tested the effects of various feedback timings on learning to create database concept diagrams. They developed a system that provided diagnostic feedback at each step of the process of solving a problem (item-by-item) and

compared it to similar systems that provided diagnostic feedback at the end of the process (end-of-test-a) or only information on whether the solution was correct or incorrect at the end of the process (end-of-test-b). The students who received item-by-item feedback scored significantly higher on an immediate posttest than did the students in the other two groups. No delayed posttest was administered.

2.4.2.7 *Item-by-item more effective than delay longer than end-of-test*

Each study in this category is also included in a category above. Dihoff et al. (2004) conducted a classroom experiment using undergraduate student participants. The researchers found that a group that received item-by-item feedback did significantly better on the 2-week delayed posttest than the other groups, including 1-day delayed feedback.

Similarly, Brosvic and Epstein (2007) studied undergraduate students in introductory psychology courses. The students who had been trained using item-by-item feedback performed better than the students who had been trained using 1-day delayed feedback on delayed posttests that were taken 3, 6, 9, and 12 months after treatment.

King, Young, and Behnke (2000) looked at the effects of feedback timing on speech performance by undergraduate in a laboratory setting. They found that 1-day delayed feedback was significantly more effective than item-by-item feedback in getting the participants to lengthen their planned introductions.

2.4.2.8 *Item-by-item more effective than delay shorter than end-of-test*

Both of Schroth's studies (1992, 1995) summarized above found that item-by-item feedback was more effective than various delays shorter than end-of-test for one purpose: achieving criterion on a task in which the participants had to learn a sorting rule based on feedback from the researcher.

2.4.2.9 *No difference between item-by-item and end-of-test/various delays*

Clariana et al. (2000) tested a connectionist model of feedback timing using item-by-item and end-of-test feedback. Their participants were high school students who read passages of varied content and answered comprehension questions on computers. The researchers found no significant difference between item-by-item and end-of-test feedback for item learning. However, they found the interesting trend that item-by-item feedback was more effective with difficult items and end-of-test feedback was more effective with easy items, which is similar to Kulhavy and Anderson's (1972) claim that delayed feedback (operationalized as 1-day delayed) is more effective than immediate feedback (operationalized as end-of-test) for difficult tests.

El Saadawi et al. (2008) looked at the pathology reports written by medical residents who had been trained using a computer program in two conditions: the residents received feedback as they wrote (item-by-item) or after they had written the entire report (end-of-test). All of the residents showed significant improvement from pretest to posttest, but there were no significant differences between the two conditions.

In a study by Gaynor (1981), undergraduate students answered free-response questions about business statistics. A pretest was given on paper during class, then a four-lesson sequence was completed on a computer outside of class. Three types of feedback were provided during the lessons: item-by-item, 30-second delayed, and end-of-test. A paper-based posttest was given during class, 24 hours after the end of the treatment period, and a paper-based delayed posttest was given in class a week later. No significant differences were found between any of the groups.

Lewis and Anderson (1985) conducted three experiments, but only Experiments 2 and 3 addressed feedback timing. In these experiments, participants played a computer game in which they typed instructions to move through rooms in a maze, with the goal of exiting it. They could

choose various actions depending on what features they saw in the rooms, but only one action was correct (i.e., leading to progress out of the maze) in a given room. In Experiment 2, the participants received feedback immediately upon completing an action (item-by-item) or after completing the next action (i.e., after proceeding one room down the wrong path). In an immediate posttest, no significant difference was found between the two feedback timings. No delayed posttest was administered. Experiment 3 was similar to Experiment 2, except that the amount of practice with the computer game was doubled and spread out over two days. In the immediate posttest, the participants in the item-by-item feedback group significantly outperformed those in the 1-step delayed feedback group. Again, no delayed posttest was performed.

Schooler and Anderson (1990) conducted three experiments in which they taught novices how to use the LISP programming language. The results of Experiments 2 and 3 are summarized above, while the results of Experiment 1 fit the current category. As in Experiments 2 and 3, a LISP tutor program was used to instruct the participants and provide feedback either as they were typing their solutions to problems (item-by-item) or after they submitted a complete solution (end-of-test). Experiment 1 differed from the other two in that once a participant in the end-of-test condition submitted a solution and received feedback once, additional feedback on the same problem became item-by-item. No significant difference was found between the end-of-test and item-by-item feedback conditions in terms of the number of errors that the participants made on the posttest and in the total time they required to complete the posttest.

Smith and Kimball (2010) investigated the effects of feedback timing on participants' learning of trivia facts using a cued-response (short-answer) test in two experiments conducted on computers in a lab. The first experiment is summarized above. In the second experiment, as in

the first, undergraduate students received item-by-item feedback or feedback delayed by 8 minutes. In the second experiment, the delay between the feedback and test was controlled. A posttest administered a week after the treatment showed no difference between the two feedback conditions.

As summarized above, Surber and Anderson (1975) tested high school students on their comprehension of an article using multiple-choice questions on paper, providing them end-of-test or 1-day delayed feedback. No difference was found in the feedback timings for reinforcement of correct responses (although the 1-day delayed feedback was more effective for error correction).

Van der Kleij, Eggen, Timmers, and Veldkamp (2012) investigated the effects of three computer-based feedback conditions on L1 adult speakers of Dutch who answered multiple-choice questions related to marketing (in Dutch). The conditions were item-by-item knowledge of the correct response plus elaborated feedback (an explanation of how to arrive at the correct response), end-of-test knowledge of the correct response plus elaborated feedback, and end-of-test knowledge of the correct response. No significant differences were found between the groups on an immediate posttest, and no delayed posttest was administered. However, the participants in this study reported that the item-by-item feedback was most beneficial for their learning.

2.4.2.10 *No difference between end-of-test and longer delay*

Metcalfe et al. (2009) performed two experiments, the first of which is described above. The second experiment was similar to the first, but the participants were college students instead of sixth-grade students. The object of study was again L1 English vocabulary. For the group that did not have the time lag between feedback and test controlled, the 1- to 4-day delayed feedback

was more effective than the end-of-test feedback. For the group that had the lag controlled, there was no significant difference for the two feedback timings. This implies that for the college students, the lag was a more important factor in the learning of vocabulary items than the timing of the feedback.

The first experiment of Webb et al. (1994) is described above. In the second experiment, undergraduate students answered multiple-choice questions on general knowledge and received end-of-test or 1-day delayed feedback, as in Experiment 1. An addition for Experiment 2 was that the participants rated their response confidence in addition to answering the questions, but those results are irrelevant to the current discussion. Although the posttest results for the end-of-test and 1-day delayed feedback were not significantly different, the participants who received 1-day delayed feedback corrected a significantly higher proportion of errors from the training to the posttest than did those who received end-of-test feedback.

2.4.2.11 *Other*

Peeck and Tillema's (1978) study does not fit into the scheme of Table 3 because the researchers did not include a group that received either item-by-item or end-of-test feedback. The researchers studied the effects of varied feedback timings on Dutch fifth grade students' responses to multiple-choice questions about an article in their L1. The study took place in the students' classrooms and was completed on paper. One group received feedback 30 minutes after answering the questions, while another group received feedback 1 day later. The participants took an immediate posttest on the second day of the study and a delayed posttest on the seventh day. No differences were found between the groups on the immediate posttest, but the group that received 1-day delayed feedback performed significantly better than the other group on the delayed posttest. In addition, the students who received 1-day delayed feedback corrected a

significantly higher proportion of errors from the treatment to the posttest than the students who received 30-minute delayed feedback.

2.5 SLA Theories and Metalinguistic Feedback

Although the cognitive processing window theory addresses the timing of negative feedback, it does not indicate whether including metalinguistic information in feedback will be more effective than not including it within the window. To begin to address this question, I consider the idea of attention to form.

According to Schmidt (1995), metalinguistic feedback may be more useful to learners than recasts alone, and attention or *noticing* is the key to this prediction. As suggested by Pica (1994) and Gass (1997), providing metalinguistic feedback draws learners' attention to the correct forms, making metalinguistic feedback more effective than feedback that does not include metalinguistic information. However, the question remains as to whether this prediction can be applied to a situation that does not involve interaction per se, but a learner answering multiple-choice questions and receiving feedback. According to Gass (1997),

> "If what is crucial about interaction is the fact that input becomes salient in some way (i.e., enhanced), then it matters little how salience comes about—whether through a teacher's self-modification, one's own request for clarification, or observation of another's request for clarification. The crucial point is that input becomes available for attentional resources and attention is focused on a particular form or meaning." (p. 129)

Thus, it is again the psycholinguistic basis of the interaction approach that I adopt here, with the key to the effectiveness of metalinguistic feedback being attention.

2.6 Language Research on Metalinguistic Feedback

The language-related research on the relative effectiveness of metalinguistic and nonmetalinguistic feedback is summarized in the following section. First, I cover SLA and CALL research, followed by language assessment research.

2.6.1 SLA and CALL research

The effectiveness of various types of feedback has been a major focus within second language acquisition (SLA) research. Starting with the explicitness of instruction, researchers have shown in two meta-analyses that explicit instruction is generally more effective than implicit instruction (Norris & Ortega, 2000; Spada & Tomita, 2010). First, Norris and Ortega (2000) performed a meta-analysis of empirical studies of the effectiveness of L2 instruction. They defined *explicit instruction* as instruction that included rule explanation or that asked learners to arrive at metalinguistic explanations. *Implicit instruction*, on the other hand, was defined as instruction that did not include rule explanations or ask learners to arrive at them. Learning was measured differently among the studies, with 65% using constrained constructed response, 39% using selected response, 29% using metalinguistic judgment, and 16% using free constructed response. The results showed that explicit instruction was more effective than implicit instruction, with a large mean effect size. This effect size was even larger for studies that used constrained constructed response and selected response. More recently, Spada and Tomita (2010) used the same definitions for *implicit* and *explicit instruction* as Norris and Ortega (2000), and the studies that they meta-analyzed also overlapped with those in the Norris and Ortega study. However, they collapsed the learning outcome measures into two categories: free and controlled responses (including constrained constructed response, selected response, and metalinguistic judgment). They further divided studies into those that examined simple and

complex features, which was determined based on the number of transformations needed to apply the relevant rule. Their results for explicit instruction overall were similar to those of Norris and Ortega, with Spada and Tomita finding larger mean effect sizes for explicit compared to implicit instruction. However, the results for free versus controlled outcome measures were somewhat different, with Spada and Tomita finding the largest effect size for free outcome measures after explicit instruction on complex features.

Although the results of Norris and Ortega (2000) and Spada and Tomita (2010) provide some insight into the effects of metalinguistic instruction, they do not directly address the effects of metalinguistic feedback. Three meta-analyses have investigated this factor, with two showing a greater effect of metalinguistic feedback, and one showing a greater effect of nonmetalinguistic feedback. Li (2010) meta-analyzed studies on both written and oral corrective feedback on L2 learning, and two of the variables he examined were metalinguistic feedback and recasts. The results showed larger effect sizes for metalinguistic feedback in both immediate and delayed posttests. Turning to oral corrective feedback exclusively, Lyster and Saito (2010) performed a meta-analysis of classroom studies. They classified feedback as recasts, explicit correction, and prompts. The category of prompts included metalinguistic feedback in addition to clarification requests, repetition of error, and elicitation. Their results showed larger effect sizes for prompts compared to recasts. Unlike the other meta-analyses summarized above, Mackey and Goo (2007) found nonmetalinguistic feedback to be more effective than metalinguistic feedback. They performed a meta-analysis of studies of synchronous interaction on the acquisition of grammar structures, including both face-to-face and computer-mediated interaction. The types of feedback examined were recasts, negotiation, and metalinguistic feedback, and the mean effect size for metalinguistic feedback was smaller than those for recasts and negotiation.

Like the meta-analyses, somewhat mixed results have come of experimental and quasi-experimental SLA research focused on the relative effectiveness of metalinguistic explanations and other types of feedback. First, Bitchener and Knoch (2009) studied the correction of article errors made in writing by ESL learners and found no advantage for metalinguistic feedback. The authors compared three types of direct corrective feedback, all of which included providing the correct form: written and oral metalinguistic explanation; written metalinguistic explanation; and direct corrective feedback only. They found no significant difference in the groups' accuracy scores on immediate and delayed posttests, although all groups improved over time compared to the pretest scores.

Conversely, many SLA researchers have found an advantage for metalinguistic feedback. Carroll and Swain (1993) studied Spanish L1 learners of ESL. The participants needed to produce the dative alternation for a given sentence or state that it did not alternate. The researchers divided the learners into five groups who got the following types of feedback when they answered incorrectly: metalinguistic information, an explicit indication that the answer was wrong, a recast, a question about whether they were sure, and no feedback. The metalinguistic feedback outperformed all other groups. With a focus on English articles, Muranoi (2000) also found an advantage for L2 learners provided with metalinguistic rule feedback after a group interaction in which implicit negative feedback was provided, compared to the group that received the implicit feedback only. In this study, the treatment was provided as part of a task in which students played roles in oral conversation, and learning was measured using a similar task. Ellis, Loewen, and Erlam (2006) found that explicit, metalinguistic feedback on oral errors in using the past tense in English was more effective than recasts. They measured learning using an oral imitation test, an untimed grammaticality judgment test, and a metalinguistic knowledge test,

but did not test production of the structure. Similarly, Sheen (2007) found that explicit, metalinguistic feedback on oral errors in English article use was more effective than recasts. She measured learning using a speeded dictation test, a writing test prompted by a series of pictures and words, and an error correction test. Goo (2011) studied Korean L1 learners of English in a foreign language setting and the effect of metalinguistic feedback and recasts on their oral production. Learning was measured using oral production and grammaticality judgment tests. The results differed depending on the structure investigated, with metalinguistic feedback being more effective for the that-trace filter, while recasts were more effective for past unreal conditionals. Goo stated that the differing results may be due to the relative complexity of the two rules.

In CALL, several studies by Nagata and Heift on the learning of L2 Japanese and German, respectively, demonstrated an advantage for explicit, metalinguistic feedback over other types of feedback, such as increasing the salience of the error using highlighting or repetition when the student reviews his or her response (e.g., Heift, 2004, 2006; Nagata & Swisher, 1995; Nagata, 1997). In contrast, studies on CALL feedback on the acquisition of Spanish as a foreign language by L1 English speakers have generally revealed either no difference between metalinguistic and no metalinguistic feedback or an advantage for no metalinguistic feedback (e.g., Kregar, 2011; Moreno, 2007; Sanz & Morgan-Short, 2004). CALL feedback has also been studied with ESL learner participants, with mixed results for the relative effectiveness of metalinguistic and nonmetalinguistic feedback (e.g., Loewen & Erlam, 2006; Sauro, 2009).

Finally, a study by Murphy (2010) bridged the divide between the SLA and CALL research reported in this section and the educational psychology literature reported in the next section. That is, although Murphy studied learning in a CALL context, he did not use

metalinguistic feedback, but rather, elaborative feedback, which consisted of hints to foster interaction and repeated attempts to choose the correct answer. This form of elaborative feedback is to some extent similar to the scaffolding provided in sociocultural-theory based interventions (e.g., Aljaafreh & Lantolf, 1994; Poehner & Lantolf, 2013). Other types of elaborative feedback are considered in the next section. In Murphy's study, pairs of English learners in Japan answered multiple-choice reading comprehension questions. One group of learners received feedback that included only the correct response, while another group received elaborative feedback followed by the correct response. Murphy found that the elaborative feedback was superior to the correct response feedback based on the students' accuracy results on a follow-up comprehension exercise.

2.6.2 L2 assessment research on metalinguistic feedback

As mentioned above, three major types of assessments provide learners feedback on their responses: formative, diagnostic, and dynamic assessments. Despite the fact that the feedback is provided with the intention of promoting learning, no L2 empirical research on these types of assessment has addressed the question of the effectiveness of metalinguistic feedback. Research on dynamic assessment comes the closest, so I briefly describe this below.

Dynamic assessment is based on the principles of sociocultural theory. It is described as follows:

> Dynamic assessment integrates assessment and instruction into a seamless, unified activity aimed at promoting learner development through appropriate forms of mediation that are sensitive to the individual's (or in some cases a group's) current abilities. In essence, DA is a procedure for simultaneously assessing and promoting development that takes account of the individual's (or

group's) zone of proximal development (ZPD). (Lantolf & Poehner, 2004, p. 50)

Mediation is a form of feedback that may assist the learner in correctly responding to a question. This mediation may take the form of metalinguistic or nonmetalinguistic feedback.

One of the purposes of dynamic assessment is to promote development, and some research in this area has been conducted on the effectiveness of dynamic assessment to promote language development (e.g., Ableeva, 2010; Poehner & Lantolf, 2013; Poehner, 2007, 2008). For example, Poehner and Lantolf (2013) developed computerized multiple-choice dynamic assessments of Chinese and French reading and listening comprehension. The mediation was a series of hints that gradually narrowed down the search space for the correct answer. In addition to normal items, the test included transfer items, which were in the same format as the normal items, but more difficult. The authors claimed that their computerized dynamic assessments showed that the participants learned during the test, but given that there was no pretest for what they already could do on the transfer tasks, this claim can only be supported anecdotally. The relative effectiveness of the types of mediation (i.e., feedback) used was not compared, and this is true of other studies of dynamic assessment as well.

2.7 Educational Psychology Theories of Informational Feedback

Because educational psychology as a field does not have a special focus on language learning, theories within the field have little to say about the effectiveness of metalinguistic feedback. However, among the categories of feedback that are often distinguished, such as knowledge of results, knowledge of correct response, and informational feedback (e.g., Jaehnig & Miller, 2007), metalinguistic feedback fits into the last category. To reiterate the definitions, *knowledge of results* feedback tells the learner only whether he or she answered correctly or incorrectly. *Knowledge of correct response* feedback includes knowledge of results feedback

(implicitly or explicitly), but also indicates the correct response. *Informational feedback* includes knowledge of correct response feedback plus additional information, such as metalinguistic information.

The literature on these types of feedback in educational psychology is largely atheoretical. Although many researchers have compared the feedback types under varied conditions, most have not proposed theoretical accounts as to why, for example, informational feedback may be more effective than knowledge of correct response feedback. One exception to this is the work of Smits, Boon, Sluijsmans, and Van Gog (2008), who examined the influence of working memory capacity on learning to perform a task. They claimed that for complex cognitive tasks, learners with little prior knowledge benefit from informational feedback that provides details at each step because it helps them avoid overloading working memory capacity, presumably because they can refer to the details in writing, rather than needing to hold them in working memory. On the other hand, learners with more prior knowledge can benefit more from global feedback that does not include details, presumably because the knowledge of the details is already integrated into their knowledge schemata, freeing working memory capacity to look at the task more globally. However, this theory does not directly apply to less complex tasks, such as responding to multiple-choice questions, which do not contain explicit steps. In addition, the distinction between detailed and global feedback is not analogous to that between metalinguistic and nonmetalinguistic feedback.

2.8 Educational Psychology Research on Informational Feedback

Several review articles have compared informational feedback and knowledge of correct response feedback, with the two more recent (Bangert-Drowns et al., 1991; Jaehnig & Miller, 2007) finding an advantage for informational feedback, although the oldest (Kulhavy & Stock,

1989) found no consistent pattern. In a meta-analysis, Bangert-Drowns, Kulik, Kulik, and Morgan (1991) found a larger effect size for informational feedback (referred to as *explanation*; $d = 0.53$) than for knowledge of correct response feedback ($d = 0.22$). Most recently, in a review of feedback types in programmed instruction (a behaviorist approach to individual instruction), Jaehnig and Miller (2007) found that knowledge of results feedback was not effective, while knowledge of correct response feedback was less effective than informational feedback (referred to in the article as *elaborative feedback*).

One additional review article compared informational feedback and knowledge of results feedback. Crooks (1988) reviewed studies and meta-analyses on classroom evaluation, without performing a meta-analysis himself. Crooks concluded that knowledge of results should be provided for all questions, with more informative feedback (e.g., metalinguistic information) only necessary when the student has made an error.

As a counterpoint, at least one study showed an advantage for knowledge of correct response feedback over informational feedback. Kulhavy, White, Topp, Chan, and Adams (1985) had undergraduate students read a passage about the U.S. Navy, then answer multiple-choice questions about it. Four different feedback conditions were used, with the feedback varying along a continuum of complexity from knowledge of correct response to knowledge of correct response plus an explanation of why each of the incorrect alternatives was incorrect plus the section of the passage where the answer could be found. The researchers found an advantage for less complex feedback.

Many studies have found no advantage to providing explanation feedback compared to providing correct answer feedback (e.g., Park & Gittelman, 1992; Whyte, Karolick, Nielsen, Elder, & Hawley, 1995). I will elaborate on a few recent examples here. Mandernach (2005)

studied undergraduate students in a psychology class who received one of a number of different kinds of feedback on multiple-choice questions. One group received no feedback, a second group got knowledge of results, a third group got knowledge of correct results, a fourth group got knowledge of results and was presented with a paragraph in which the correct answer could be found, and a fifth group got knowledge of correct response and received explanations of the selected response and the correct response. The feedback that the fifth group received is clearly a type of informational feedback, while the feedback that the fourth group received could also be considered informational. No significant differences were found among the conditions. As another example, Smits, Boon, Sluijsmans, and Van Gog (2008) had Dutch secondary school students work through genetics problems (presumably in their L1) in a web-based environment, and they got either global feedback, which included the correct answer and a problem-solving approach, or informational feedback, which included the correct answer, problem-solving approach, a fully worked solution, and an explanation for why the answer was correct. The learners who had the higher prior knowledge performed better under the global feedback condition, while no difference between the conditions was found for the learners who had lower prior knowledge.

Note that most of the individual studies of informational feedback and those in the reviews examined item learning only. Counter to this trend is the research of Butler, Godbole, and Marsh (2013), who looked at both item learning, using questions that asked about definitions of concepts, and extension of learning to inference questions in reading comprehension. The definition questions appeared on both 5-minute-delayed posttests and 1-week-delayed posttests, and the inference questions only appeared on the final posttests. The types of feedback that were provided were knowledge of correct response or informational feedback, which included

knowledge of correct response and two additional sentences from the passage that elaborated on the correct answer. The researchers found that the informational feedback was more effective for promoting extension of learning than correct answer feedback, but that the two types of feedback were equivalent for promoting item learning.

Overall, the educational psychology literature concurs with the L2 literature: Informational feedback (such as metalinguistic feedback) generally seems to be more effective than knowledge of correct response for promoting both item learning and system learning.

2.9 Research Questions and Hypotheses

The current study focuses on learning the correct usage of English articles when feedback that includes or does not include metalinguistic information is provided on an item-by-item basis or at the end of a test. This results in four conditions: item-by-item feedback including metalinguistic information, item-by-item feedback without metalinguistic information, end-of-test feedback including metalinguistic information, and end-of-test feedback without metalinguistic information.

To summarize the theory and research introduced so far, SLA and educational psychology theories makes conflicting predictions about the ideal timing of feedback, and empirical studies can be cited to support or refute any given prediction. The effectiveness of metalinguistic feedback is much clearer and has been generally supported by both theory and empirical results. However, none of the theories considered above predicts how providing or not providing metalinguistic feedback will interact with item-by-item or end-of-test timing. For example, it is possible that the effects of increasing attention by providing metalinguistic feedback and of increasing attention by providing that feedback on an item-by-item basis (as predicted by the theories supporting the interaction approach) has an additive effect, leading to

the most attention and therefore the most learning, but it is also possible that attention cannot meaningfully be increased in this way. Because of this gap in both theoretical and empirical knowledge and the potential for this knowledge to help teachers better plan feedback to increase student learning, the first research question is as stated below:

1. On a multiple-choice drag-and-drop test, does the timing (item-by-item or end-of-test) and type (with or without metalinguistic information) of feedback affect ESL students' gain scores on 5-minute-delayed and 1-week-delayed posttests
    a. on the same (repeated) questions?
    b. on new questions?

Note that repeated and new questions are included here in order to measure the effects on both item learning and system learning. (Recall that an item has been learned if the learner can correctly respond to it after a previous exposure to the item, whereas a system has been learned if the learner can correctly apply a rule to a previously unseen situation.)

Next, the cognitive processing window theory, the dual-trace hypothesis, and interference-perseveration hypothesis make different predictions based on whether the learner answers a question correctly, so it is also important to break down the results by correct and incorrect responses. This may help to distinguish which theories are best supported by the data. Thus, the second research question is as follows:

2. On a multiple-choice drag-and-drop test, does the conditional probability of correctly answering a question on 5-minute-delayed and 1-week-delayed posttests differ for groups based on the timing (item-by-item or end-of-test) and type (with or without metalinguistic information) of feedback for questions that are initially answered

a. correctly?

b. incorrectly?

A summary of the predictions based on three theoretical perspectives is shown in Table 4. The two item types that are considered are multiple-choice items that the learners have received feedback on during the treatment (repeated items) and multiple-choice items that the learners have not received feedback on (new items).

2.9.1 Predictions for Research Question 1

The cognitive processing window, as hypothesized by Doughty (2001), provides an opportunity for learners to notice a contrast between their selected response and the correct response. This window is limited by the constraints of working memory. Therefore, item-by-item feedback on errors is predicted to be more effective for language learning than end-of-test feedback. When the feedback comes at the end of the test, the learner may no longer have his or her erroneous response in working memory and may be unable to make the comparison. The situation is similar for correct responses, although the prediction in this case is not completely clear because it is not directly addressed by the theory. Therefore, for all item types (repeated and new), one may predict that the participants in the item-by-item group will have higher gain scores on the 5-minute-delayed and 1-week-delayed posttests than the end-of-test and no-feedback groups.

The dual-trace hypothesis and interference-perseveration hypothesis both predict that end-of-test feedback will be superior to item-by-item feedback for repeated items that are initially answered incorrectly (error correction). However, the dual-trace hypothesis and interference-perseveration hypothesis predict slightly differing results for item-by-item and end-of-test feedback when the test taker responds correctly (reinforcement of correct responses). The

interference-perseveration hypothesis predicts no difference in this case, while the dual-trace hypothesis predicts that the end-of-test feedback will be superior.

Because the dual-trace hypothesis and interference-perseveration hypothesis are not designed to deal with system learning, only speculation is possible for their predictions for the new items. The dual-trace hypothesis may predict the best results for the end-of-test feedback because the rules have, in effect, been reviewed twice as many times in this condition as in the item-by-item feedback condition. That is, a learner receiving item-by-item feedback would have attempted to remember a rule once when answering a question, and would have then read the rule, with the two instances blurring into one. In the end-of-test condition, the learner would have attempted to remember the rule once when answering the question and then read the rule later, at the end of the test, resulting in two separate instances of reviewing the rule. The interference-perseveration hypothesis may predict no difference between the item-by-item and end-of-test feedback because no memory trace of the new items exists.

The results of a pilot of the current study (Lavolette, 2013) do not fit the predictions of any of the theories perfectly. The (nonsignificant) trends seen in the pilot followed the predictions of the educational psychology theories for error correction (i.e., end-of-test > item-by-item). However, an advantage was seen for item-by-item feedback in the reinforcement of correct responses, which is as predicted by the cognitive process window theory. Finally, a very slight, nonsignificant advantage was seen for end-of-test feedback for system learning. Because of its small magnitude, this advantage is likely to remain nonsignificant in the current study, if it appears at all. This may best fit the predictions of the interference-perseveration hypothesis.

2.9.2 Predictions for Research Question 2

The only theory that bears on the issue of metalinguistic feedback is the attention-based

theory supporting the interaction approach. From this perspective, metalinguistic feedback should increase learners' attention to the error and the correct response, leading to more error correction than feedback that does not include metalinguistic information. For the same reasons, metalinguistic feedback should increase learners' attention to feedback on correct responses, leading to better reinforcement of correct responses. This increased attention should also lead to better system learning.

Table 4: Theoretical Predictions for Posttest Results Based on Feedback Timing and Type

	Item learning (repeated items)		System learning (new items)
	Reinforcement of correct responses	Error correction	
Cognitive processing window (SLA)	EOT < IBI	EOT < IBI	EOT < IBI
Dual trace	EOT > IBI	EOT > IBI	EOT > IBI (?)
Interference-perseveration	EOT = IBI	EOT > IBI	EOT = IBI (?)
Attention-based (SLA)	No meta < Meta	No meta < Meta	No meta < Meta

Note. IBI = item by item; EOT = end of test; meta = metalinguistic.

2.9.3 Predictions for interaction effects between feedback timing and feedback type

No empirical evidence exists that would allow me to predict an interaction effect between feedback timing and providing or not providing metalinguistic information. In addition, because the educational psychology theories considered here do not directly address metalinguistic feedback, they cannot make predictions about the potential interaction. Speculatively, I predict

that the effects of feedback provided during the cognitive processing window and the attention-enhancing effects of metalinguistic feedback may be cumulative, leading to the greatest effect on learning when metalinguistic feedback is provided on an item-by-item basis (i.e., within the cognitive processing window). However, it is possible that metalinguistic feedback provided outside of the cognitive processing window is just as effective as metalinguistic feedback provided within it.

CHAPTER 3: METHOD

The first chapter above explained the purpose of the current study, which is to provide evidence of how varied feedback timing and types affect the acquisition of English by adult second language learners. As explained in the following chapters, this was accomplished by measuring how providing or not providing metalinguistic information, item-by-item or at the end of a test, affects ESL learners who answer multiple-choice questions that require them to apply rules for using English articles. In Chapter 2, I reviewed the theories and previous literature from various fields that bear on this question. The theories lead to conflicting predictions about the optimal feedback timing, and the research on feedback timing has been inconclusive. On the other hand, theory and research generally agree that metalinguistic feedback is more effective than nonmetalinguistic feedback, although no theory predicts how feedback type interacts with varied feedback timing, nor do any empirical results bear on this issue. I presented the research questions and predictions at the end of Chapter 2.

Below, I provide details on the participants who took part in the study, then describe the materials used. Following that is a description of the procedure, then details of how the data were analyzed.

3.1 Participants

Participants were recruited from classes at the English Language Center at Michigan State University, and 221 learners agreed to participate. Excluding those who did not complete all portions of the study (n = 35) and those who scored at ceiling on the pretest (n = 74), 112 students participated, with about 25 to 30 randomly assigned to each feedback group. All 112 participants completed the exit questionnaire, and demographic information on the students is shown in Table 5. The students were in the top two levels of the Intensive English Program

(Levels 3 and 4) and the single level of the English for Academic Purposes Program (Level 5). The mean iBT TOEFL scores of the students in these levels overall were about 67 for Level 3, 68 for Level 4, and 72 for Level 5 (D. Reed, personal communication, March 28, 2014). The average age of the participants in each feedback group was from 20 to 23 years. Their native languages were primarily Mandarin Chinese and Arabic, which reflects the population of the English Language Center.

Table 5: Participants' Demographic Information

Group	N	Gender	Level	Age	Language	Years studying English	Months in US
Item-by-item metalinguistic	28	18 male 10 female	10 Level 3 11 Level 4 7 Level 5	22 mean 4.3 SD 18 low 38 high	10 Mandarin 2 Cantonese 11 Arabic 1 Japanese 3 Portuguese 1 Taiwanese	5.5 mean 5.1 SD 1 low 10 high	10.8 mean 9.1 SD 1 low 40 high
Item-by-item without metalinguistic	25	12 male 13 female	5 Level 3 11 Level 4 9 Level 5	20 mean 1.5 SD 18 low 25 high	14 Mandarin 1 Cantonese 3 Arabic 1 Korean 3 Japanese 2 Thai 1 Mandarin-Cantonese bilingual	8.0 mean 4.0 SD 1 low 15 high	9.6 mean 8.2 SD 1 low 30 high
End-of-test metalinguistic	29	13 male 16 female	7 Level 3 13 Level 4 9 Level 5	20 mean 2.7 SD 18 low 30 high	13 Mandarin 1 Cantonese 4 Arabic 2 Korean 5 Japanese 1 Taiwanese 2 Thai 1 missing	7.3 mean 3.4 SD 1 low 13 high	8.4 mean 8.9 SD 1 low 39 high
End-of-test without metalinguistic	30	16 male 13 female 1 missing	6 Level 3 16 Level 4 8 Level 5	23 mean 7.2 SD 18 low 54 high	16 Mandarin 6 Arabic 2 Korean 4 Japanese 1 Mandarin-Cantonese bilingual 1 Mandarin-Taiwanese bilingual	7.7 mean 4.1 SD 1 low 20 high	13.2 mean 15.1 SD 1 low 69 high

Also on the exit questionnaire, the participants indicated whether prior to participating in the study, they had been familiar with the six rules for articles that were tested (Appendix B). All but four participants were familiar with at least one rule, and the mean number of rules that participants knew was 3.4 (SD = 1.6). Fourteen participants indicated that they knew all six rules. The average number of participants who knew each rule was 64 (SD = 8.0), with a minimum of 49 and a maximum of 71 participants knowing a given rule.

3.2 Materials

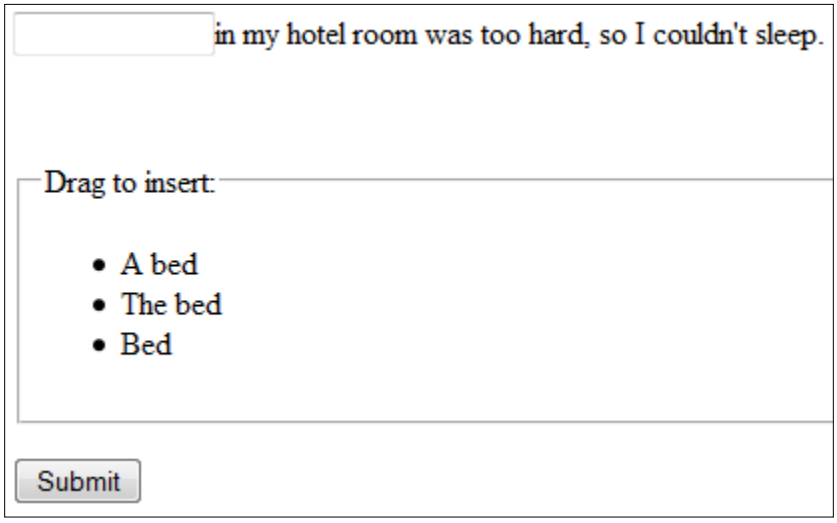
I administered a pretest, two identical treatments, a 5-minute-delayed posttest, a 1-week-delayed posttest, and an exit questionnaire, all via computers. The purpose of the materials was to measure the participants' knowledge of the usage of articles (pretest and posttests), to teach the participants how to use articles (treatments), and find out demographic information about the participants (exit questionnaire). The topic of the items on the tests and treatments was article usage. I chose this topic because of its persistent difficulty for many learners (e.g., Master, 2002). Although using articles in a nonnativelike way may not lead to misunderstandings, in my experience, ESL learners often express an interest in improving their accuracy in the use of articles.

All items on all of the tests and treatments were multiple-choice. This made it possible to give item-by-item and end-of-test feedback that was specific to each participant's responses. I wrote all of the items, which were then tested out on 8 native English speakers to ensure that more than one (for the experimental items) or two (for the filler items) responses were not possible. Items for which 7 out of 8 native speakers did not agree on the intended response were revised and retested.

The participants answered the items by dragging and dropping the selected response into

the blank in the sentence (see Figure 1). Each sentence had one blank, and the three answer choices were “a/an [noun],” “the [noun],” and “[noun].” The drag-and-drop movement was selected because it allowed the participants to see their selected response within the context of the sentence, as opposed to checking a box. In addition, previous research has shown that dragging is more effective for learning than clicking (Heift, 2003). To respond to the questions, the participants dragged and dropped the article and the following noun as a single piece. Because the article and noun were linked in this way, responding to a question always involved the same type of mouse movement, whether the chosen answer contained an overt article or not. If the article and following noun were separate pieces, making a “no article” response would have involved dragging and dropping only one word, while making a response that contained an overt article would have required dragging and dropping two separate words. Therefore, keeping the article and noun linked made the participants’ task more uniform.

Figure 1: Example question.



The target nouns were singular and countable, and each noun appeared an equal number of times in each context (no article, *a/an*, *the*, and either *a/an* or *the*). This ensured that the participants would not receive a score higher than a person guessing randomly by using a

strategy such as always choosing the response with the definite article.

Each item had one blank and three answer choices, which corresponded to the three possible article choices of no article, an indefinite article (*a* or *an* as appropriate), or the definite article (*the*). One item appeared per screen on all tests and treatments, and the participant was required to submit an answer before proceeding to the next item. This prevented missed responses due to clicking the "Submit" button without selecting a response. Participants could not go back to previous items, which prevented the participants from changing previous responses due to what they learned later in the test or treatment. The items on each test and treatment were randomized for each participant. All tests and treatments were presented using a web-based open-source testing platform, Concerto (http://www.psychometrics.cam.ac.uk/). This platform was chosen because of its flexibility to deliver feedback whenever and however the programmer/researcher desired, as well as regulate the timing of breaks. The platform was installed on servers owned by Michigan State University's College of Arts and Letters. It recorded the responses to the questions and the time that each screen was displayed, in addition to recording the responses to the exit questionnaire. No record was made of the clicking or mouse movements that occurred while the participants were choosing their responses. Having this data might provide some insight into what the participants were thinking when they were choosing their responses. However, this information would provide little information to support or refute the hypotheses being tested in the current study, so the collection and analysis of this type of data is left to future studies.

3.2.1 Pretest

The purpose of the pretest was to determine how accurately the participants could use articles following the targeted rules before the treatments began. The pretest contained 48

multiple-choice items, listed in Appendix C. Thirty-six of the items are experimental items that have only one possible answer, and the remaining 12 items are fillers that have two possible answers. The purpose of the filler items was to show the students that the domains of the rules are not mutually exclusive in actual usage. However, these items were not analyzed because two out of three possible answers were correct, giving a high likelihood of answering them correctly by guessing. The pretest took about 15 minutes for the participants to complete.

3.2.2 Treatments 1 and 2

Each treatment contained 32 of the multiple-choice items from the pretest. Using only 32 of the 48 pretest items reserved 16 items from the pretest that did not appear on the treatments and that the participants therefore did not receive feedback on. These *new* items were used to answer Research Question 1b, while the items on the treatment, which the participants received feedback on, were *repeated* items and were used to answer Research Question 1a. Of the 32 treatment items, 24 were experimental items, and 8 were fillers. The number of filler items was chosen in proportion to the number of experimental items included on the treatment. That is, the proportion of experimental to filler items on the pretest (36:12 = 3:1) is the same as the corresponding proportion on the treatments (24:8 = 3:1). An example question is shown in Figure 1.

The treatments provided four types of feedback, depending on the feedback group (i.e., item-by-item with metalinguistic information, item-by-item without metalinguistic information, end-of-test with metalinguistic information, end-of-test without metalinguistic information; see Fig. 2). The feedback for all groups followed the same format, with the feedback provided on one question per screen. The feedback screen first showed the correct answer(s), followed by "You answered **correctly**" for correct responses or "You answered **incorrectly**" for incorrect

responses (Figure 3). I decided on this arrangement of the feedback for the following reason: by learning that he or she answered the question incorrectly before seeing the correct answer, a participant could repeat the processing that occurred when he or she initially saw the item, thus reactivating the error in working memory. This is of particular concern for participants in the end-of-test feedback condition, who would not have seen the question on the immediately preceding screen. Providing the correct answer first without the learner's erroneous response helps to minimize the possibility of reactivating the error. Note that the participants were shown the correct answer, but not their own, potentially erroneous, response. After the indication of the correctness of the learner's response, for the two metalinguistic feedback groups, next on the same screen was a metalinguistic rule, regardless of whether the participant answered correctly or incorrectly. Note that no instruction was provided other than the feedback itself. A week passed between the two identical treatments, which each lasted about 20 minutes.

Figure 2: Division of participants into feedback groups. IBI = item by item; EOT = end of test.

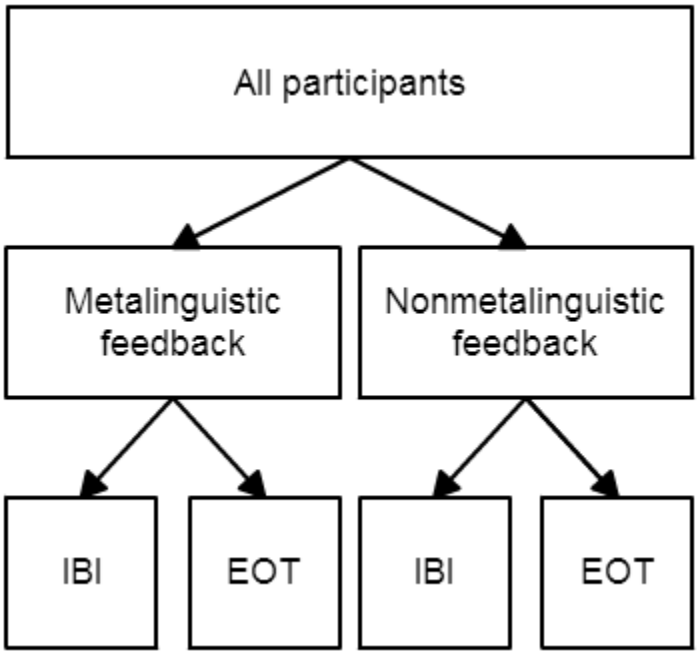
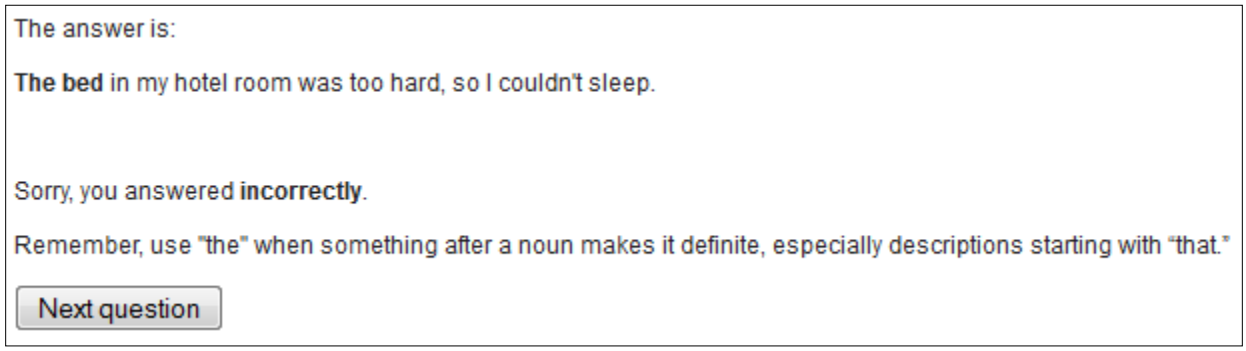


Figure 3: Metalinguistic feedback on an incorrect response.



3.2.3 Five-minute-delayed posttest

The participants completed the 5-minute delayed posttest during the same session as Treatment 2, after a 5-minute break. I included this break to ensure that the 40-second cognitive processing window had fully elapsed for all of the items and feedback on the preceding treatment, as well as to give the participants a chance to rest. The 5-minute-delayed posttest included all items from the pretest so that both item learning and system learning could be measured. No feedback was provided on any of the posttest items to any of the groups because the effects of feedback on the posttest would have been impossible to separate from the effects of the treatment, especially for the participants who received item-by-item feedback. While it is possible that the participants may have remembered the correct answers for the items that appeared on the treatments (once on each treatment, or twice for each item), 16 of the items (12 experimental and 4 filler) appeared only on the pretest and the posttests. Because no feedback was provided on these items, the participants needed to apply the rules that they had learned. The participants completed the posttest in about 15 minutes.

3.2.4 One-week-delayed posttest

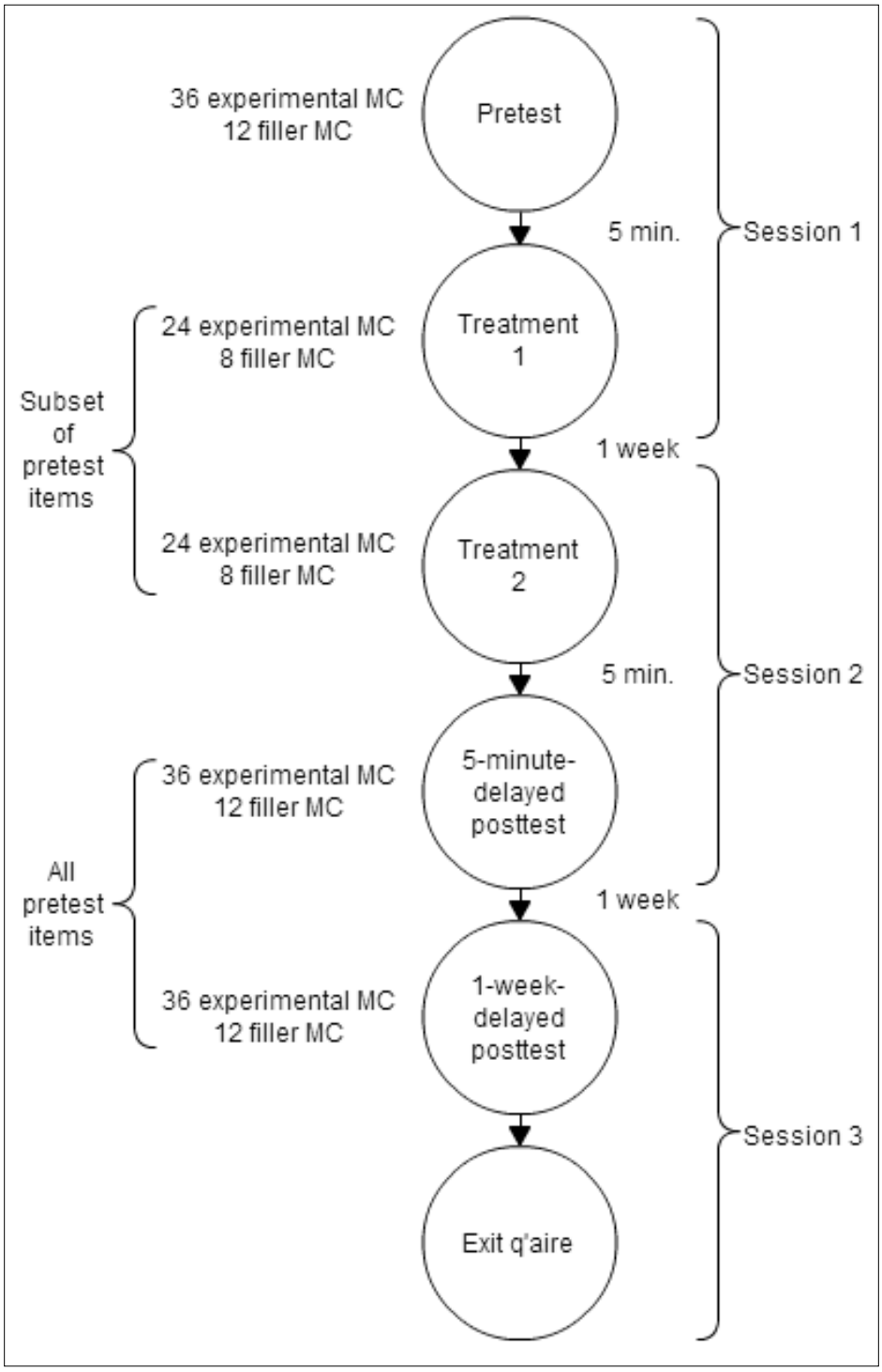
The participants completed the 1-week-delayed posttest 1 week after the 5-minute-delayed posttest, and it contained the same items as the 5-minute-delayed posttest. All of the

participants received end-of-test metalinguistic feedback as a courtesy. This feedback did not affect the posttest results because it came after the final posttest was completed. The participants completed this posttest in about 15 minutes.

3.3 Procedure

Figure 4 shows an overview of the procedure. In advance of Session 1, I randomly assigned the participants to the feedback groups (Figure 2), with approximately the same number of participants in each group per class. During Session 1, I demonstrated how to respond to the items. This was important so that the mechanics of answering the items themselves would not cause the participants to answer them incorrectly. Then, the participants completed the pretest, which took 15 to 20 minutes, depending on the participant. After a 5-minute break, the participants completed Treatment 1. This took 20 to 30 minutes. During this first session, I told the participants that I would come to their class a total of three times. Although I did not tell them that they would encounter the same items on all of the tests and treatments, it is probable that many of them figured this out after the first treatment. Some of them asked me why they saw the same items again, and I told them that repetition was useful for learning. A week later, during Session 2, the participants completed Treatment 2, took a 5-minute break, then completed the 5-minute-delayed posttest, which altogether took 35 to 45 minutes. One week later, they completed the 1-week-delayed posttest and the exit questionnaire, which took 20 to 30 minutes.

Figure 4: Procedure.



I was present in the classes where the study took place only during the experimental sessions themselves, so I do not know the exact content that the participants were exposed to

outside of class. However, the instructors of the classes told me that they did not explicitly study articles in their classes during the class sessions before the study or during the study itself.

3.4 Analysis

As a preliminary step in analyzing the data, I excluded from the analysis participants who scored very high on the pretest. These participants were not of interest because they had very little margin to learn from the feedback. Specifically, I eliminated participants who scored above 85% (an arbitrarily chosen cut-off score) on the new or repeated items: 11 or 12 out of 12 on the new items or 20 or more out of 24 on the repeated items. This eliminated 61 participants. I also calculated the item discrimination of each item on the pretest, considering the repeated items and the new items separately. The item discrimination indicates “the degree to which an item separates the students who performed well from those who performed poorly on the test as a whole” (Brown, 2005, p. 68). That is, an item with low discrimination may be easy for participants who score low on the test overall and difficult for participants who score high on the test overall, which indicates that the item may not be testing the same construct as the other items on the test. There may be something wrong with the statement of the item itself which makes it tricky, or it may be an item that is very easy or very difficult for all participants. Therefore, items with a discrimination of lower than .2 (marked with asterisks in Appendix C) were excluded from the analysis, leaving 10 new items and 17 repeated items. This discrimination level was chosen based on Ebel's (1979, p. 267) recommendations, as cited by Brown (2005, p. 75). After excluding these items, participants who were at the new ceiling of 9 or 10 out of 10 new items or 15 or more out of 17 repeated items were excluded. This eliminated a further 13 participants. After the elimination of these participants, the total number included in the analyses was 112, with demographic information as given above.

All of the analyses described below are mixed-design ANOVAs, which are used when both repeated-measures and between-groups variables are involved in the analysis (Field, 2009, p. 507). This type of analysis was chosen, rather than the other possible analyses (e.g., a repeated-measures ANOVA and a series of t-tests), to minimize the number of analyses needed to answer the research questions, thereby reducing the likelihood of finding a statistically significant result by chance alone. In addition, the ANOVAs provide information on interactions between variables, which cannot be produced by a t-test.

For all of the ANOVAs, I calculated effect sizes as partial eta squared (η^2_{part}). According to Brown (2008), these effect sizes "indicate the percentage of variance in each of the effects (or interaction) and its associated error that is accounted for by that effect (or interaction)" (p. 42). Thus, partial eta squared can be understood as a percentage. To my knowledge, no one has assigned standard interpretations to these effect sizes. However, for eta squared effect sizes, J. Cohen (1988) interpreted .01 as a small effect, .06 as a medium effect, and .14 as a large effect. Using this as guidance for partial eta squared, I interpreted values above .15 as a large effect, between .15 and .05 as a medium effect, between .05 and .01 as a small effect, and below .01 as a very small effect.

I used a significance level of $p \leq .05$ for all analyses. An argument could be made for using a correction to the significance level because of the multiple ANOVAs used. However, as detailed below, I performed the analyses on mutually exclusive portions of the data. For example, I used the gain scores for the repeated questions in one ANOVA, while I used the gain scores for the new items in a separate ANOVA. In that sense, I did not make multiple comparisons of the treatment groups, so I chose not to use a correction for the significance level.

Before performing the ANOVAs, I checked the three assumptions of normality,

homogeneity of variance, and sphericity. For the assumption of normality, I used Kolmogorov-Smirnov tests and examined the Q-Q plots. For each ANOVA, some of the Kolmogorov-Smirnov tests were significant, indicating that the assumption of normality had been violated, and the corresponding Q-Q plots confirmed this in some cases. However, ANOVAs are generally robust to violations of normality when group sizes are equal (Field, 2009, pp. 359–360), so I decided that the assumption of normality was sufficiently met in all cases. The results of checking the other two assumptions vary for each of the ANOVAs, so they are provided below.

3.4.1 Research Question 1

After the preliminary steps of excluding items and participants, I ran the analyses described below to determine whether the participants who received the four treatments differed in how much they learned, addressing Research Question 1. To investigate item learning (Research Question 1a), I analyzed the old items only. I used a 2 x 2 x 2 mixed-design ANOVA with the gain scores (from the pretest to the 5-minute-delayed and 1-week-delayed posttests) as a within-subject factor and feedback timing (item-by-item or end-of-test feedback) and feedback type (with or without metalinguistic information) as between-subjects factors. To investigate system learning (Research Question 1b), I ran a separate 2 x 2 x 2 mixed-design ANOVA using the gain scores for the new items, using the same factors.

Before performing the ANOVAs, I checked the three assumptions of normality, homogeneity of variance, and sphericity. Information about the normality assumption is given above. I used Levene's test to check the homogeneity of variance. No significant results were found, so this assumption was met. Finally, the assumption of sphericity is automatically met because at least three levels of a variable are necessary for sphericity to become a problem (Field, 2009, p. 459), and each variable had only two levels.

3.4.2 Research Question 2

To answer Research Question 2, I further analyzed the posttest results of the repeated and new items based on whether a given participant answered correctly or incorrectly on the corresponding pretest item, following Clariana, Wagner, and Roher Murphy (2000; see also Butler et al., 2007; Peeck & Tillema, 1978; Smith & Kimball, 2010; Surber & Anderson, 1975). The terminology used in this previous literature is *conditional probabilities*, and I adopt this terminology here as well. However, the calculations are performed post-hoc, so *probability* here should not be understood as a prediction. Instead, it is a description of the learners' responses to the posttest questions based on their pretest responses. These conditional probabilities provide information about how the treatment affected reinforcement of correct responses (when a participant answers correctly on both a pretest and posttest) and error correction (when a participant answers incorrectly on a pretest and correctly on a posttest). For both the repeated and new items, the four sets of conditional probabilities that I calculated are R_2/R_1, the probability that an item is answered correctly on the 5-minute-delayed posttest, given that it has been answered correctly on the pretest; R_3/R_1, the probability that an item is answered correctly on the 1-week-delayed posttest, given that it has been answered correctly on the pretest; R_2/W_1, the probability that an item is answered correctly on the 5-minute-delayed posttest, given that it has been answered incorrectly on the pretest; and R_3/W_1, the probability that an item is answered correctly on the 1-week-delayed posttest, given that it has been answered incorrectly on the pretest. For example, I calculated R_2/R_1 by first counting the number of correct responses for a given participant on the pretest (R_1), then counting the number of correct responses for those same questions on the 5-minute-delayed posttest (R_2). I then divided R_2 by R_1 to get R_2/R_1. For the repeated and new items, I performed two 2 x 2 x 2 mixed-design ANOVAs with the

conditional probabilities on the 5-minute-delayed and 1-week-delayed posttests (R_2/R_1 and R_3/R_1; R_2/W_1 and R_3/W_1) as a within-subject factor and the two feedback timing groups and the with/without metalinguistic feedback groups as between-subjects factors.

The analysis of the conditional probabilities is necessary to investigate differences in the predictions of the educational psychology theories. In addition, the results obtained from investigating the conditional probabilities have the potential to provide more detailed insight into what the participants learned correctly or incorrectly through their participation in the study. If, for example, I examine the results from item learning overall, I can determine the gain scores from pretest to posttest. However, gain scores do not tell the whole story. The intended result of the treatment is that participants who answered questions incorrectly on the pretest will answer those questions correctly on the posttest, without incorrectly answering any of the questions that they originally answered correctly. However, imagine this scenario: Out of 20 questions, a participant answers 10 correctly and 10 incorrectly on a pretest. After a treatment, the same participant answers the 20 items again on a posttest. Of the 10 items answered correctly on the pretest, she answers 7 of the items correctly on the posttest, and of the 10 items answered incorrectly on the pretest, she answers 3 correctly on the posttest. Her score is still 10, with a gain score of zero, which may lead to the conclusion that the treatment had no effect. However, a closer investigation of the conditional probabilities would reveal that the treatment had both positive and negative effects. Without separating the results into conditional probabilities, this information is lost. Therefore, while overall gain scores provide useful information, akin to averages, breaking down scores into conditional probabilities may provide further useful information. However, the conditional probabilities will be volatile and therefore unreliable if many of the participants answer relatively few questions either correctly or incorrectly on the

pretest. Ideally, a test for which conditional probabilities are used will provide a large pool of items that are answered both correctly and incorrectly on the pretest for each participant. As the results below will show, relatively few items were answered incorrectly on the pretest, making the results of the analyses of error correction (but not reinforcement of correct responses) unreliable.

Before performing the ANOVAs, I checked the three assumptions of normality, homogeneity of variance, and sphericity. Information about the normality assumption is given above. Next, I used Levene's test to check the homogeneity of variance. A significant result was found for R_2/W_1 for the item-by-item and end-of-test groups for the repeated items, which lends further support to the claim that the analyses involving this conditional probability are not reliable. In all other cases, this assumption was met. Finally, the assumption of sphericity was met because at least three levels of a variable are necessary for sphericity to become a problem (Field, 2009, p. 459), and each variable had only two levels.

3.4.3 Question and feedback display times

Next, I wanted to determine whether the time that the participants spent viewing the questions and feedback on the treatments was significantly different among the with/without metalinguistic feedback groups and two feedback timing groups. Because of the self-paced nature of the tests and treatments, I needed to investigate the possibility that the time spent with the feedback displayed differed among the feedback groups. If so, the time differences, rather than the treatments themselves, could explain any differences in the learning outcomes. Although data that is a set number of standard deviations from the mean is often trimmed from reaction times, in the current study, this would have eliminated display times in which the participant may simply have been viewing the screen for an unusually long time. Although classroom distractions,

rather than long reading times, undoubtedly caused some of these outliers, it is impossible to separate the two types of long display times. That said, these classroom distractions most likely affected all feedback groups in the same way, so including all of these long display times in the analysis should not qualitatively affect the results. Of course, the display times should not be taken as an exact measure of how long the participants were actually reading the screen, but only as an upper bound. With these limitations in mind, I eliminated from the current analysis only the display time for the first question on each test and treatment; which question was eliminated varied by participant because of the randomization of the questions. This first display time was often longer than the other display times for two reasons. First, the participants sometimes waited to be told to proceed before answering the first question. Second, Treatment 1 and the 5-minute-delayed posttest both came after a computer-timed 5-minute break, and the participants were generally not ready to proceed instantaneously after the break. Note, however, that this reasoning does not apply to the time that the feedback was displayed, so none of the feedback display times were excluded from the analysis. I did not remove display times that were exceptionally short for either the questions or the feedback because these very short display times simply indicated that the participant spent very little time with that page displayed.

To analyze the feedback and question display times, I ran three mixed-design ANOVAs on the total display times of each participant for each test or treatment (Table 7). The display times for the 1-week-delayed posttest were not included in any of the analyses because this posttest occurred last, and therefore could not have affected the previous tests. All ANOVAs used feedback timing (item-by-item and end-of-test) and type (with and without metalinguistic information) as between-subjects factors, but the within-subject (repeated measures) factor varied. The first ANOVA was 4 x 2 x 2 and used the display times for the repeated questions on

the pretest, Treatment 1, Treatment 2, and the 5-minute-delayed posttest as the within-subjects variable. The second ANOVA used the display times for the new questions on the pretest and the 5-minute delayed posttest as the within-subjects variable. Note that the new questions did not appear on the treatments, so no question display times were recorded. The third ANOVA was 2 x 2 x 2 and used the feedback display times for Treatment 1 and Treatment 2 as the within-subjects variable. Note that the treatments were the only materials where the feedback appeared and that the participants saw this feedback before they took either posttest.

Before performing the ANOVAs, I checked the three assumptions of normality, homogeneity of variance, and sphericity. Information about the normality assumption is given above. The results of checking the other two assumptions are given for each ANOVA in turn below.

I began with the analysis of the display times for the repeated questions (ANOVA 1 in Table 7). I used Levene's test to check the homogeneity of variance, and the results showed that this assumption was met. Mauchly's test showed that the assumption of sphericity was violated, so I used the Greenhouse-Geisser correction for the significance values for this analysis.

Next, I analyzed the display times for the new questions (ANOVA 2 in Table 7). I used Levene's test to check the homogeneity of variance, and the results showed that this assumption was met. The assumption of sphericity was met because at least three levels of a variable are necessary for sphericity to become a problem (Field, 2009, p. 459), and each variable in this ANOVAs had only two levels.

Finally, I analyzed the display times for the feedback (ANOVA 3 in Table 7). I used Levene's test to check the homogeneity of variance, and the results showed that this assumption was violated. The assumption of sphericity was met because at least three levels of a variable are

necessary for sphericity to become a problem (Field, 2009, p. 459), and each variable in this ANOVA had only two levels. Although not all of the assumptions of a mixed-design ANOVA were met in this case, I decided to run the analyses anyway because no alternative analysis was available. The results, therefore, will be interpreted with caution.

3.5 Summary of Analyses

To sum up this section, the tables below detail the research questions and the analyses used to investigate them. The analyses used on the participants' gain scores and conditional probabilities are shown in Table 6. The analyses used on the reaction times are shown in Table 7.

Table 6: Research Questions and Corresponding Analyses of Gain Scores and Conditional Probabilities

Research question			Within-subject (repeated measures) factor
1. Does timing & type of feedback affect gain scores on 5-minute- & 1-week-delayed posttests	a. on repeated questions?		Gain scores (from pretest to 5-minute- and 1-week-delayed posttests) as within-subject factor
	b. on new questions?		Gain scores (from pretest to 5-minute- and 1-week-delayed posttests) as within-subject factor
2. Does conditional probability of correctly answering question on 5-minute- and 1-week-delayed posttests differ for groups based on timing & type of feedback for questions initially answered	a. correctly	on repeated questions?	Conditional probabilities on 5-minute- & 1-week-delayed posttests (R_2/R_1 & R_3/R_1) as within-subject factor
		on new questions?	Conditional probabilities on 5-minute- & 1-week-delayed posttests (R_2/R_1 & R_3/R_1) as within-subject factor
	b. incorrectly	on repeated questions?	Conditional probabilities on 5-minute- & 1-week-delayed posttests (R_2/R_1 & R_3/R_1) as within-subject factor
		on new questions?	Conditional probabilities on 5-minute- & 1-week-delayed posttests (R_2/R_1 & R_3/R_1) as within-subject factor

Note. All analyses are 2 x 2 x 2 mixed-design ANOVAs with feedback timing (item-by-item or end-of-test) and feedback type (with or without metalinguistic information) as between-subjects factors.

Table 7: Analyses of Feedback and Question Display Times

ANOVA #	Research question	Display times for	Levels of repeated-measures variable	Tests & treatments included
1	1a	Repeated items	4	Pretest, Treatments 1 & 2, 5-minute-delayed posttest
2	1b	New items	2	Pretest, 5-minute-delayed posttest
3	1a & b	Feedback	2	Treatments 1 & 2

Note. All analyses are mixed-design ANOVAs with feedback timing (item-by-item or end-of-test) and feedback type (with or without metalinguistic) as between-subjects factors.

CHAPTER 4: RESULTS

As explained in Chapter 1, the purpose of the current study is to provide evidence of how varied feedback timing and types affect the acquisition of English by adult second language learners. To accomplish this, ESL learner participants answered multiple-choice questions that required them to apply rules for using English articles, and I investigated how providing or not providing metalinguistic feedback, either item-by-item or at the end of the test, affected their responses. In Chapter 2, I reviewed the relevant theories and previous literature. Although theories disagree and the experimental research is inconclusive about the optimal feedback timing, theory and research generally agree that providing metalinguistic feedback is more effective than not providing it. In Chapter 3, I described the participants, who were ESL learners who primarily spoke Chinese and Arabic as their first languages. I also described the materials, which were drag-and-drop multiple-choice questions, feedback, and questionnaires, all delivered via computer. The procedure, shown in Figure 4, took place in the participants' ESL classrooms, and was conducted over three class periods. Finally, I described the analysis of the gain scores (Table 6) and question and feedback display times (Table 7), which was primarily accomplished using mixed-design ANOVAs.

Below, I provide preliminary results of the pretest and each posttest, then address each of the research questions in turn. Following that is a summary of the results.

4.1 Preliminary Results

Descriptive statistics for the pretest and both posttests are shown in Table 8. I calculated the reliability of each test using Cronbach's alpha. Kline (2005) recommended that alpha be at least .7 for low stakes tests (p. 182), and some of statistics for the old and new items do not reach this level. However, this is likely due to the small number of items included (Cortina, 1993).

When all 27 test items are included in the calculation, the minimum alpha is .79, which is well within the acceptable range.

Table 8: Overall Test Results

Test	Items	Number of items	Cronbach's alpha	Mean	SD	Min	Max	Mean item discrimination
Pretest	Old items	17	.57	10.3	2.8	3	14	.37
	New items	10	.09	6.1	1.5	1	8	.32
	All items	27	.79	16.4	4.3	4	22	.35
5-min.-delayed posttest	Old items	17	.81	12.2	3.7	4	17	.49
	New items	10	.40	6.5	1.8	2	9	.40
	All items	27	.86	18.7	5.5	6	26	.46
1-week-delayed posttest	Old items	17	.74	12.0	3.3	4	17	.44
	New items	10	.25	6.7	1.6	3	10	.36
	All items	27	.81	18.6	4.9	7	27	.41

Note. The number of test takers was 112 for all tests.

4.2 Research Question 1

First, I address the question of whether the timing (item-by-item or end-of-test) or type (with or without metalinguistic information) of feedback on a multiple-choice drag-and-drop test affects student scores on 5-minute-delayed and 1-week-delayed posttests on repeated and new questions. That is, I investigated whether feedback timing or type affects item or system learning, in the short or somewhat longer term.

4.2.1 Research Question 1a: Repeated items

To investigate item learning, I first analyzed the repeated items. For all participants combined, the mean gain from the pretest to the 5-minute-delayed posttest was 1.9 points (SD = 2.9), and the mean gain from the pretest to the 1-week delayed posttest was 0.9 points (SD = 2.5). Thus, overall, gain scores were higher on the first posttest, with a loss of about one point from the first to the second posttest. However, the standard deviations were quite high, with a wide range of individual scores. A breakdown by feedback timing and type of feedback is shown in Table 9.

Table 9: Mean (SD) Gain Scores From Pretest to Each Posttest, Repeated Items

	5-minute-delayed posttest	1-week-delayed posttest
Item-by-item	2.2 (2.6)	0.8 (2.4)
End-of-test	1.6 (3.0)	0.9 (2.5)
Metalinguistic	2.2 (2.8)	1.1 (2.3)
No metalinguistic	1.7 (2.9)	0.6 (2.6)
Item-by-item metalinguistic	2.4 (2.8)	1.3 (2.4)
Item-by-item, no metalinguistic	2.1 (2.6)	0.1 (2.3)
End-of-test metalinguistic	2.0 (2.9)	0.9 (2.2)
End-of-test, no metalinguistic	1.3 (3.2)	1.0 (2.8)

I performed a 2 x 2 x 2 mixed-design ANOVA with the gain scores as a within-subject factor and feedback timing and with/without metalinguistic feedback as between-subjects factors. The main effect of time (from the 5-minute-delayed posttest to the 1-week-delayed posttest) was significant, $F(1, 108) = 27.83$, $p < .001$, $\eta^2_{part} = .21$ (a large effect size), with the gain scores on the 1-week-delayed posttest being lower. That is, on average, the participants increased their scores from the pretest to each posttest, but the gains were not fully maintained on the 1-week-delayed posttest. Note that this effect is relatively large. The main effect of feedback timing was not significant, $F(1, 108) = 0.18$, $p = .67$, $\eta^2_{part} = .002$ (a very small effect size), nor was the main effect of feedback type, $F(1, 108) = 1.13$, $p = .29$, $\eta^2_{part} = .01$ (a small effect size). The interaction between feedback timing and feedback type was not significant, $F(1, 108) = 0.24$, $p = .62$, $\eta^2_{part} = .002$ (a very small effect size). The interaction between time and feedback timing was significant, $F(1, 108) = 4.20$, $p = .043$, $\eta^2_{part} = .037$ (a small effect size), with the participants

who got item-by-item feedback doing better than the participants who got end-of-test feedback on the 5-minute-delayed posttest, but the two groups performing similarly on the 1-week-delayed posttest (Fig. 5). The effect size is small. This interaction will be considered in more detail in the discussion. The interaction between time and feedback type was not significant, $F(1, 108)$, = 0.013, $p = .91$, $\eta^2_{part} < .001$ (a very small effect size). The three-way interaction of time, feedback timing, and feedback type was significant, $F(1, 108) = 4.90$, $p = .029$, $\eta^2_{part} = .043$, which is a small effect size. As Figure 6 shows, the groups that received metalinguistic feedback dropped in their gain scores at approximately the same rate from the 5-minute-delayed to the 1-week delayed posttest, while the groups that did not receive metalinguistic feedback showed differing patterns. The group that got item-by-item nonmetalinguistic feedback showed a sharper drop in gain scores from the first to the second posttest, while the group that got end-of-test feedback without metalinguistic information showed very little decrease in gain scores. The interpretation of this interaction will be further considered in the discussion.

Figure 5: Gain scores for two feedback timings on both posttests, repeated items only. IBI = item-by-item feedback; EOT = end-of-test feedback.

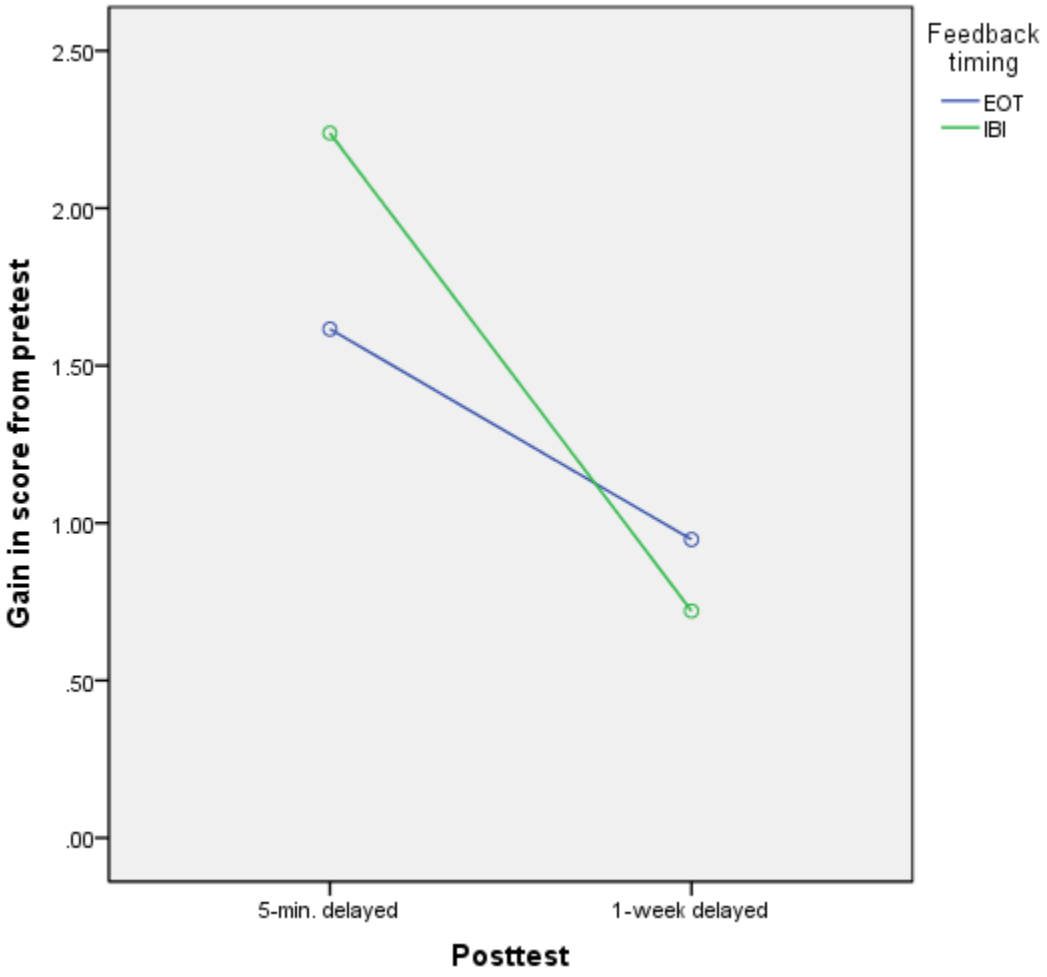
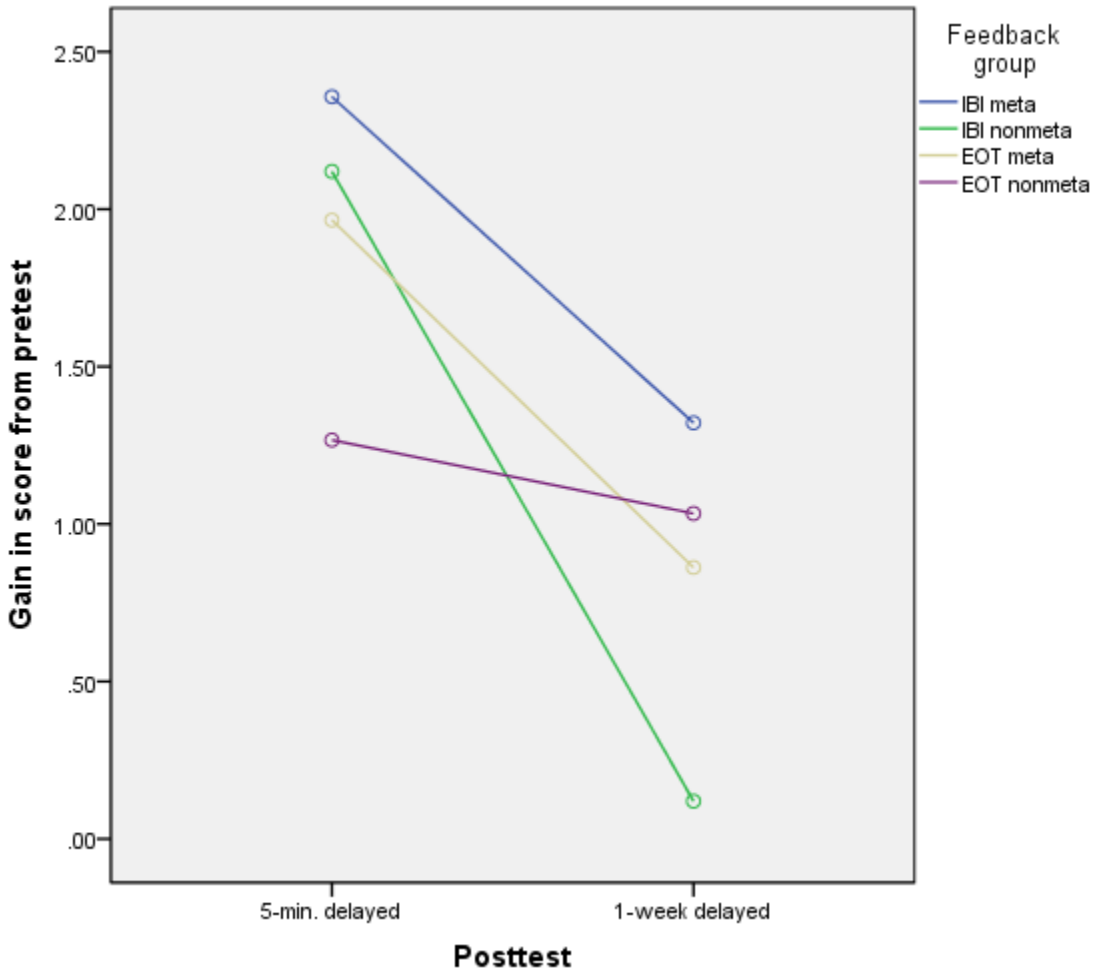


Figure 6: Gain scores of feedback groups on both posttests, repeated items only. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback; nonmeta = no metalinguistic feedback.



To determine whether the display time for the questions on the pretest, two treatments, and first posttest was significantly different among the with/without metalinguistic feedback groups and two feedback timing groups, I performed a 4 x 2 x 2 ANOVA on the total display times per participant. The mean display times are shown in Table 10. The main effect of time was significant, $F(2.4, 254.4) = 192.6$, $p < .001$, $\eta^2_{part} = .64$, which is a large effect size, with the mean total display times decreasing over time. This is not surprising because the participants

became increasingly familiar with the questions over time, which meant that they needed less time to read them and respond to them. No other effects were significant: feedback timing, $F(1, 108) = 0.23$, $p = .64$, $\eta^2_{part} = .002$ (a very small effect size); feedback type, $F(1, 108) = 0.25$, $p = .62$, $\eta^2_{part} = .002$ (a very small effect size); time x feedback timing, $F(2.4, 254.4) = 0.33$, $p = .75$, $\eta^2_{part} = .003$ (a very small effect size); time x feedback type, $F(2.4, 254.4) = 0.99$, $p = .39$, $\eta^2_{part} = .009$ (a very small effect size); feedback timing x feedback type, $F(1, 108) = 2.54$, $p = .072$, $\eta^2_{part} = .023$ (a small effect size); time x feedback timing x feedback type, $F(2.4, 254.4) = 1.77$, $p = .17$, $\eta^2_{part} = .016$ (a small effect size).

Table 10: Mean (SD) Total Time in Seconds Questions Were Displayed, Repeated Items Only

Test or treatment	Pretest	Treatment 1	Treatment 2	5-minute-delayed posttest
IBI	250.3 (72.7)	153.4 (43.4)	171.5 (49.2)	136.4 (39.9)
EOT	249.5 (73.9)	150.7 (38.3)	168.2 (53.9)	126.4 (50.8)
Meta	251.8 (72.1)	158.0 (41.9)	171.3 (55.0)	128.4 (42.2)
Nonmeta	247.9 (74.6)	145.7 (38.6)	168.2 (48.1)	134.0 (50.0)
IBI meta	265.6 (78.9)	162.0 (44.3)	176.4 (55.7)	139.1 (44.2)
IBI no meta	233.1 (62.3)	143.8 (41.1)	166.1 (41.0)	133.4 (35.0)
EOT meta	238.4 (63.3)	154.2 (39.9)	166.4 (54.9)	118.0 (38.1)
EOT no meta	260.2 (82.6)	147.3 (37.1)	170.0 (53.8)	134.6 (60.3)

Note. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback.

To determine whether the display time for the feedback on the two treatments was significantly different among the with/without metalinguistic feedback groups and two feedback timing groups, I performed a 2 x 2 x 2 ANOVA on the total display times per participant. The mean display times are shown in Table 11. Note that the assumption of homogeneity of variance was violated for the analysis of the feedback display time, so the results should be interpreted with caution. However, the results are easily explained by the design of the study, as shown below, which leads me to believe that they may be accurate.

A significant main effect was found for time, $F(1, 108) = 18.37$, $p < .001$, $\eta^2_{part} = .15$ (a large effect size), with the time spent with the feedback displayed decreasing from the first treatment to the second. Like the question display times, the participants were increasingly familiar with the feedback as they progressed through the treatments, so it is unsurprising that they took less time to read the feedback on the second treatment. A significant main effect was also found for feedback timing, $F(1, 108) = 4.14$, $p = .044$, $\eta^2_{part} = .037$ (a small effect size), with the participants who got end-of-test feedback displaying the feedback longer than those who got item-by-item feedback. This may be due to the fact that the participants in the item-by-item feedback group had read the context sentence immediately before getting the feedback, so they did not need to take the time to read the sentence again in detail, while the participants in the end-of-test feedback group needed to take time to refamiliarize themselves with it to understand the feedback. A significant main effect was also found for feedback type, $F(1, 108) = 6.90$, $p = .010$, $\eta^2_{part} = .060$ (a medium effect size), with metalinguistic feedback displayed significantly longer than feedback without metalinguistic information. This effect is expected because there was simply more text on the screen for the metalinguistic feedback condition.

Table 11: Mean (SD) Total Time in Seconds Feedback Was Displayed

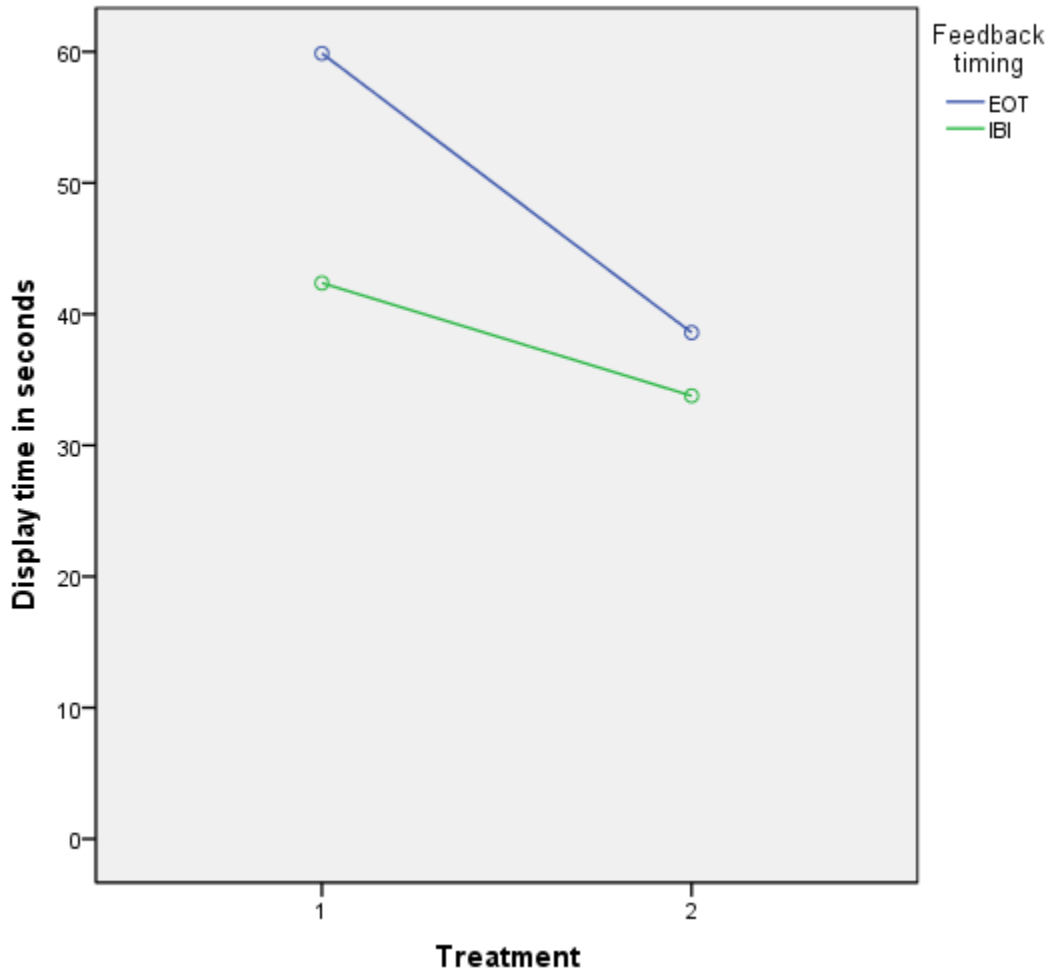
Treatment	1	2
IBI	43.1 (30.0)	34.2 (25.2)
EOT	59.8 (50.1)	38.6 (28.3)
Meta	61.1 (46.2)	40.8 (30.7)
Nonmeta	42.4 (35.5)	32.0 (21.6)
IBI meta	55.8 (33.6)	40.4 (31.8)
IBI no meta	29.0 (12.5)	27.1 (11.8)
EOT meta	66.2 (55.9)	41.1 (30.2)
EOT no meta	53.6 (43.9)	36.1 (26.7)

Note. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback.

A marginally significant interaction was found between time and feedback timing, $F(1, 108) = 3.31$, $p = .072$, $\eta^2_{part} = .030$, which is a small effect size. As Figure 7 shows, the participants who got end-of-test feedback decreased their feedback display times more from Treatment 1 to Treatment 2 than did the participants who got item-by-item feedback. This result may be due to the participants in the end-of-test group spending more time to read the feedback on Treatment 1 because of their need to refamiliarize themselves with the context sentences as explained above, but being familiar enough with the feedback itself on Treatment 2 that they did not read it at all. That is, on Treatment 2, the overall average display time for the feedback was about 36 seconds. Because there were 17 questions, this averages out to just over 2 seconds per feedback screen, which indicates that the participants in general were spending very little time reading the feedback. Thus, although the end-of-test feedback participants needed more time than the item-by-item feedback participants to read the feedback on Treatment 1, both groups

nearly stopped reading the feedback at all on Treatment 2.

Figure 7: Interaction between time and feedback timing for total feedback display time. IBI = item-by-item feedback; EOT = end-of-test feedback.



4.2.2 Research Question 1b: New items

To investigate system learning, I analyzed the new items. For all participants combined, the mean gain from the pretest to the 5-minute-delayed posttest was 0.39 points (SD = 2.8), and the mean gain from the pretest to the 1-week delayed posttest was 1.38 points (SD = 2.0). Thus, overall, gain scores were lower on the first posttest, with a gain of about one point from the first to the second posttest. However, as with the repeated items, the standard deviations were quite

high, indicating a wide range of individual scores. A breakdown by feedback timing and type of feedback is shown in Table 12.

To investigate system learning, I ran a 2 x 2 x 2 mixed-design ANOVA using the new items. The main effect of time (from the 5-minute-delayed posttest to the 1-week-delayed posttest) was significant, $F(1, 108) = 41.22$, $p < .001$, $\eta^2_{part} = .28$ (a large effect size), with the gain scores on the 1-week-delayed posttest being higher. This shows that system learning increased over the time of the study. The main effect of feedback timing was marginally significant, $F(1, 108) = 3.61$, $p = .060$, $\eta^2_{part} = .032$ (a small effect size), with the group that received item-by-item feedback scoring higher. This shows that item-by-item feedback was more effective for system learning than end-of-test feedback. The main effect of feedback type was not significant, $F(1, 108) = 1.04$, $p = .31$, $\eta^2_{part} = .01$ (a small effect size). None of the interaction effects were significant: between feedback timing and feedback type, $F(1, 108) = 0.85$, $p = .36$, $\eta^2_{part} = .008$ (a very small effect size); between time and feedback timing, $F(1, 108) = 0.069$, $p = .79$, $\eta^2_{part} = .001$ (a very small effect size); between time and feedback type, $F(1, 108), = 0.044$, $p = .84$, $\eta^2_{part} < .001$ (a very small effect size); and between time, feedback timing, and feedback type, $F(1, 108) = 2.91$, $p = .097$, $\eta^2_{part} = .025$ (a small effect size).

Table 12: Mean (SD) Gain Scores From Pretest to Each Posttest, New Items

	5-minute-delayed posttest	1-week-delayed posttest
Item-by-item	0.7 (1.5)	1.7 (2.0)
End-of-test	0.1 (2.0)	1.1 (2.0)
Metalinguistic	0.6 (1.8)	1.5 (2.1)
No metalinguistic	0.2 (1.8)	1.2 (1.9)
Item-by-item metalinguistic	0.6 (1.7)	1.9 (2.2)
Item-by-item no metalinguistic	0.8 (1.3)	1.6 (1.8)
End-of-test metalinguistic	0.6 (1.9)	1.2 (2.0)
End-of-test no metalinguistic	−0.3 (2.1)	0.9 (1.9)

To determine whether the display time for the questions on the pretest and first posttest was significantly different among the metalinguistic/nonmetalinguistic feedback groups and two feedback timing groups, I performed a 2 x 2 x 2 ANOVA on the total display times per participant. The mean display times are shown in Table 13. The main effect of time was significant, $F(1, 108) = 179.1$, $p < .001$, $\eta^2_{part} = .62$ (a large effect size), with the display times decreasing over time. As with the repeated question display times, this effect is expected because the participants became more familiar with the questions after reading them the first time, thus needing less time to read them the second time. No significant main effect was found for feedback timing, $F(1, 108) = 0.68$, $p = .41$, $\eta^2_{part} = .006$ (a very small effect size), or for feedback type, $F(1, 108) = 0.21$, $p = .65$, $\eta^2_{part} = .002$ (a very small effect size). No significant interactions were found between time and feedback timing, $F(1, 108) = 0.024$, $p = .88$, $\eta^2_{part} < .001$ (a very small effect size); time and feedback type, $F(1, 108) = 0.37$, $p = .54$, $\eta^2_{part} = .003$ (a very small

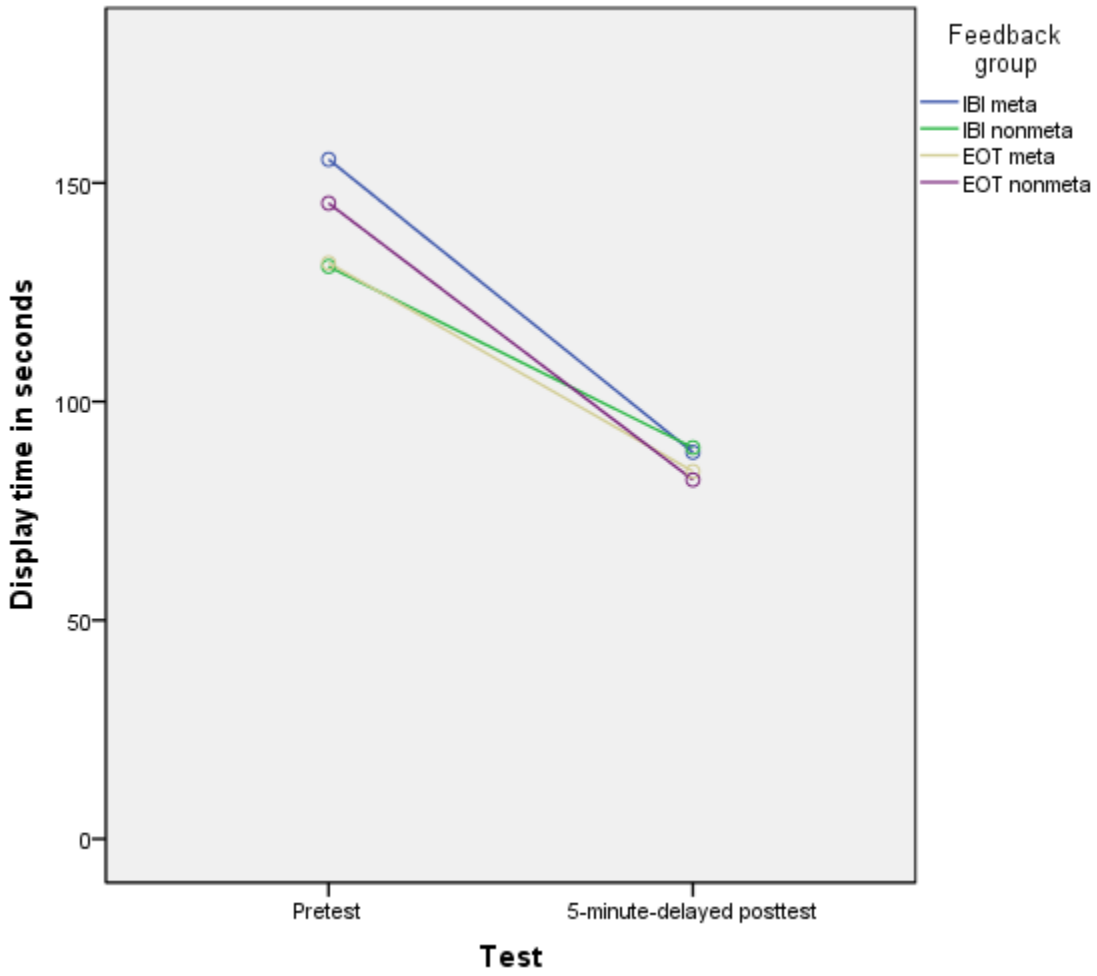
effect size); or feedback timing and feedback type, $F(1, 108) = 1.92$, $p = .17$, $\eta^2_{part} = .017$ (a small effect size). A significant interaction was found between time, feedback timing, and feedback type, $F(1, 108) = 6.29$, $p = .014$, $\eta^2_{part} = .055$ (a medium effect size). This interaction is shown in Figure 8: The four groups show some differences in how long they display the new questions on the pretest, but cluster closer together on the 5-minute-delayed posttest. This interaction cannot be explained by the variables in the study because none of the participants received feedback during the pretest. Therefore, the participants in the groups may have had some differences before the study began that led them to spend different amounts of time reading the questions on the pretest, although these differences were no longer evident by the first posttest. However, it is not clear why these differences affected the display times for the new questions, but not the repeated questions or the feedback.

Table 13: Mean (SD) Total Time in Seconds Questions Were Displayed, New Items Only

Test	Pretest	5-minute-delayed posttest
IBI	143.8 (45.9)	88.9 (29.0)
EOT	138.7 (49.8)	83.1 (31.2)
Meta	143.4 (49.8)	86.2 (32.7)
No meta	138.8 (46.1)	85.5 (27.7)
IBI meta	155.4 (53.0)	88.4 (32.6)
IBI no meta	130.9 (32.9)	89.5 (25.2)
EOT meta	131.7 (44.4)	84.0 (33.3)
EOT no meta	145.4 (54.5)	82.1 (29.6)

Note. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback.

Figure 8: Display time interaction for time x feedback timing x feedback type, new questions only. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback; nonmeta = no metalinguistic feedback.



4.3 Research Question 2

Next, I addressed the question of whether the timing (item-by-item or end-of-test) and type (with or without metalinguistic information) of feedback has a differential effect on questions that are answered correctly and incorrectly.

I calculated four sets of conditional probabilities for the repeated items and for the new items: R_2/R_1, the probability that an item is answered correctly on the 5-minute-delayed posttest, given that it has been answered correctly on the pretest; R_3/R_1, the probability that an item is answered correctly on the 1-week-delayed posttest, given that it has been answered correctly on the pretest; R_2/W_1, the probability that an item is answered correctly on the 5-minute-delayed posttest, given that it has been answered incorrectly on the pretest; and R_3/W_1, the probability that an item is answered correctly on the 1-week-delayed posttest, given that it has been answered incorrectly on the pretest. Then, for each type of question (repeated and new), I performed two 2 x 2 x 2 mixed-design ANOVAs with the conditional probabilities on the 5-minute-delayed and 1-week-delayed posttests (R_2/R_1 and R_3/R_1; R_2/W_1 and R_3/W_1) as a within-subject factor and the two feedback timing groups and the groups who received or did not received metalinguistic information as between-subjects factors.

4.3.1 Repeated items

To investigate item learning, I performed two 2 x 2 x 2 mixed-design ANOVA using the conditional probabilities for the repeated items. The first ANOVA used the probabilities of answering an item correctly on each posttest given that the item had been answered correctly on the pretest. The means for each condition are shown in Table 12. The results showed no significant main effects of time, $F(1, 108) = 0.07$, $p = .78$, $\eta^2_{part} = .001$ (a very small effect size), feedback timing, $F(1, 108) = 0.14$, $p = .71$, $\eta^2_{part} = .001$ (a very small effect size), or feedback

type, $F(1, 108) = 1.66$, $p = .20$, $\eta^2_{part} = .015$ (a small effect size). There were also no significant interaction effects for time x feedback timing, $F(1, 108) = 1.63$, $p = .20$, $\eta^2_{part} = .015$, time x feedback type, $F(1, 108) = 0.95$, $p = .33$, $\eta^2_{part} = .009$ (a very small effect size), feedback timing x type, $F(1, 108) = 0.16$, $p = .69$, $\eta^2_{part} = .002$ (a very small effect size), or time x feedback timing x feedback type, $F(1, 108) = 0.002$, $p = .96$, $\eta^2_{part} < .001$ (a very small effect size).

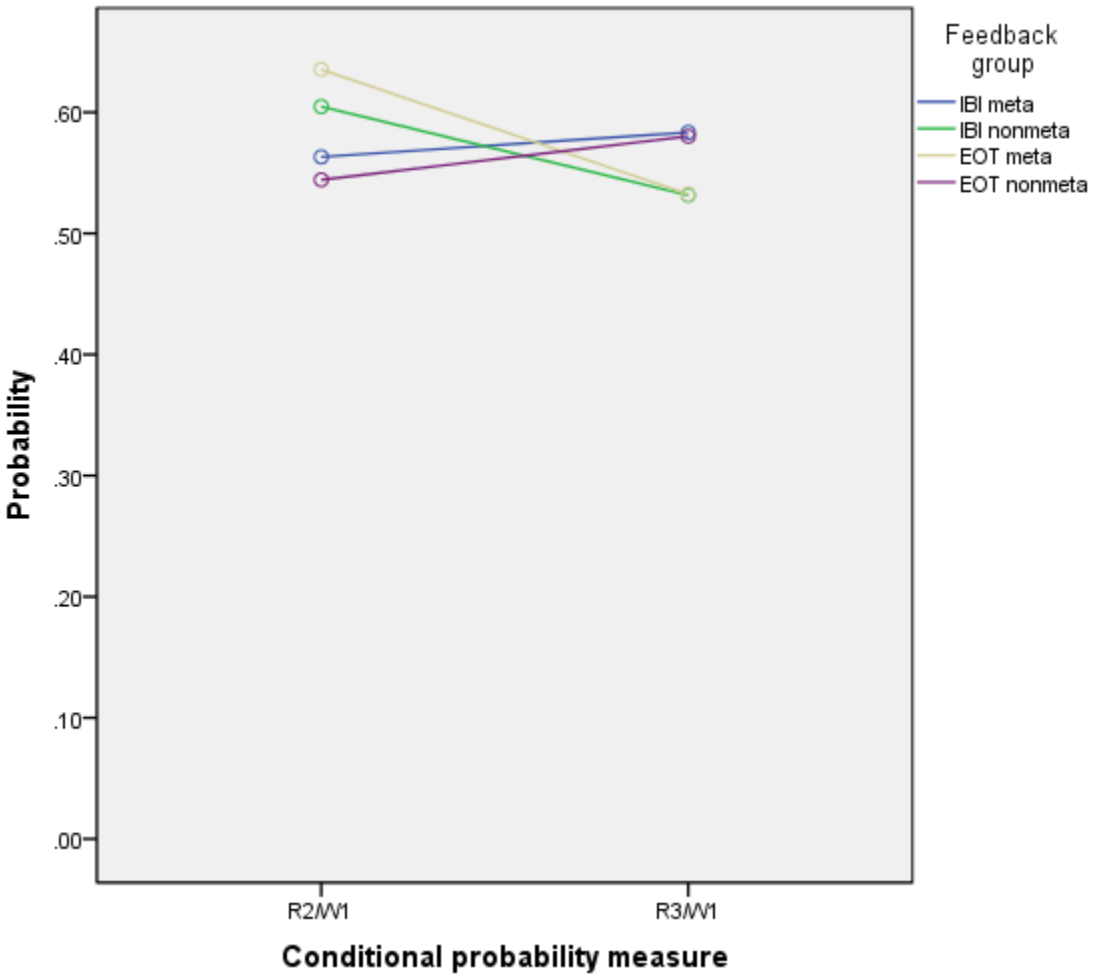
The second ANOVA used the probabilities of answering an item correctly on each posttest given that the item had been answered incorrectly on the pretest. The means for each condition are shown in Table 14. The results showed no significant main effects of time, $F(1, 108) = 2.44$, $p = .12$, $\eta^2_{part} = .022$ (a small effect size), feedback timing, $F(1, 108) = 0.003$, $p = .96$, $\eta^2_{part} < .001$ (a very small effect size), or feedback type, $F(1, 108) = 0.10$, $p = .75$, $\eta^2_{part} = .001$ (a very small effect size). There were also no significant interaction effects for time x feedback timing, $F(1, 108) = 0.034$, $p = .85$, $\eta^2_{part} < .001$ (a very small effect size), time x feedback type, $F(1, 108) = 0.35$, $p = .55$, $\eta^2_{part} = .003$ (a very small effect size), or feedback timing x type, $F(1, 108) = 0.039$, $p = .85$, $\eta^2_{part} < .001$ (a very small effect size). A significant three-way interaction was found for time x feedback timing x feedback type, $F(1, 108) = 9.15$, $p = .003$, $\eta^2_{part} = .078$, which is a medium effect size. Figure 9 shows this interaction: The group that got end-of-test metalinguistic feedback patterned with the group that got item-by-item feedback without metalinguistic information, and the group that got item-by-item metalinguistic feedback patterned with the group that got end-of-test feedback without metalinguistic information. On the 5-minute-delayed posttest, the former two groups were more likely to correctly answer a question that they had answered incorrectly on the pretest, as compared to the latter two groups. On the 1-week delayed posttest, however, the former groups' probabilities of answering correctly dropped, while the latter groups' probabilities increased. Note, however, that

the probability of the item-by-item metalinguistic group did not drop as sharply as that of the end-of-test group without metalinguistic information.

Table 14: Mean Conditional Probabilities (Standard Deviations), Repeated Items

	R_2/R_1	R_3/R_1	R_2/W_1	R_3/W_1
Item-by-item	.81 (.16)	.79 (.17)	.58 (.23)	.56 (.22)
End-of-test	.79 (.18)	.80 (.16)	.59 (.28)	.56 (.24)
Metalinguistic	.79 (.19)	.77 (.18)	.60 (.27)	.56 (.23)
No metalinguistic	.81 (.15)	.82 (.14)	.57 (.24)	.56 (.23)
Item-by-item metalinguistic	.80 (.18)	.76 (.18)	.56 (.25)	.58 (.23)
Item-by-item no metalinguistic	.83 (.11)	.82 (.15)	.60 (.20)	.53 (.20)
End-of-test metalinguistic	.78 (.20)	.78 (.19)	.64 (.29)	.53 (.23)
End-of-test no metalinguistic	.79 (.17)	.82 (.14)	.54 (.27)	.58 (.24)

Figure 9: Probability of selecting the correct response on an item on the 5-minute-delayed posttest (R_2/W_1) or 1-week-delayed posttest (R_3/W_1), given that it was answered incorrectly on the pretest, repeated items only. IBI = item by item feedback; EOT = end of test feedback; meta = metalinguistic feedback; nonmeta = no metalinguistic feedback.



Rather than attempt to interpret this complex interaction at face value, I believe that some caution is needed. For this particular interaction, the conditional probabilities used in the analysis were calculated using only the repeated items that the participants answered incorrectly on the pretest. Some participants (n = 5) included in the analysis got as few as three of the items wrong, and 42 of the participants got 3, 4, or 5 questions wrong on the pretest. This means that the

conditional probabilities were very volatile: If a participant who answered only 3 questions incorrectly on the pretest answered one of those questions correctly on the 5-minute-delayed posttest, then answered two of them correctly on the 1-week-delayed posttest, that participant's conditional probability changed from .33 to 0.66, just by answering one additional question correctly. For this reason, I do not believe that the results of this analysis are reliable or generalizable, especially given the interaction effect, which does not fit well with the results of other analyses. Note that this caution does not apply to questions that were answered correctly on the pretest because only 5 participants answered as few as 5 questions correctly on the pretest, and no participants answered fewer than 5 questions correctly.

To test this, I removed all participants who answered 5 or fewer questions incorrectly on the pretest and ran the ANOVA on the conditional probabilities (R_2/W_1 and R_3/W_1) again. The analysis contained 70 participants, and none of the main or interaction effects were significant. Of course, this analysis differs from the one that includes all participants, not only in the number of participants included, but also that the pretest scores were lower. However, I believe that this analysis adds support to my argument that the analysis above should be interpreted with caution.

4.3.2 New items

To investigate system learning, I performed two 2 x 2 x 2 mixed-design ANOVA using the conditional probabilities for the new items. The first ANOVA used the probabilities of answering an item correctly on each posttest given that the item had been answered correctly on the pretest. The means for each condition are shown in Table 15. The results showed no significant main effects of time, $F(1, 108) < 0.001$, $p = .99$, $\eta^2_{part} < .001$ (a very small effect size), feedback timing, $F(1, 108) = 0.02$, $p = .88$, $\eta^2_{part} < .001$, or feedback type, $F(1, 108) = .002$, $p = .97$, $\eta^2_{part} < .001$ (a very small effect size). There were also no significant interaction effects for

time x feedback timing, $F(1, 108) = 0.57$, $p = .45$, $\eta^2_{part} = .005$, time x feedback type, $F(1, 108) = 0.81$, $p = .37$, $\eta^2_{part} = .007$ (a very small effect size), feedback timing x type, $F(1, 108) = 1.25$, $p = .27$, $\eta^2_{part} < .001$ (a very small effect size), or time x feedback timing x feedback type, $F(1, 108) = 0.78$, $p = .38$, $\eta^2_{part} = .007$ (a very small effect size).

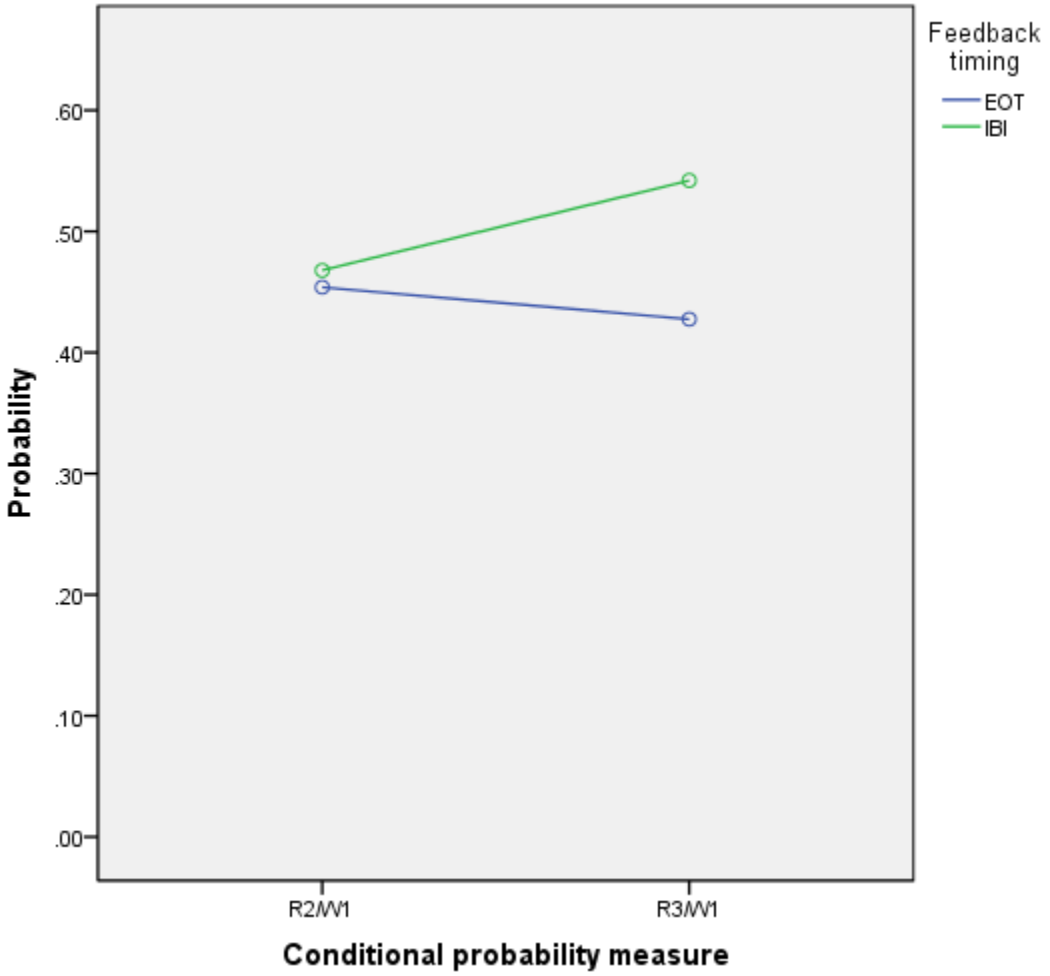
Table 15: Mean Conditional Probabilities (Standard Deviations), New Items

	R_2/R_1	R_3/R_1	R_2/W_1	R_3/W_1
Item-by-item	.77 (.19)	.76 (.19)	.47 (.27)	.54 (.26)
End-of-test	.76 (.22)	.77 (.20)	.45 (.26)	.43 (.23)
Metalinguistic	.75 (.23)	.77 (.23)	.48 (.26)	.53 (.23)
No metalinguistic	.77 (.18)	.76 (.16)	.44 (.27)	.44 (.27)
Item-by-item metalinguistic	.77 (.20)	.79 (.21)	.49 (.27)	.58 (.26)
Item-by-item no metalinguistic	.76 (.18)	.72 (.16)	.44 (.27)	.50 (.27)
End-of-test metalinguistic	.74 (.25)	.75 (.24)	.48 (.25)	.48 (.19)
End-of-test no metalinguistic	.78 (.18)	.79 (.16)	.43 (.27)	.38 (.26)

The second ANOVA used the probabilities of answering an item correctly on each posttest given that the item had been answered incorrectly on the pretest. The means for each condition are shown in Table 14. The results showed no significant main effects of time, $F(1, 108) = 1.15$, $p = .29$, $\eta^2_{part} = .011$ (a small effect size), feedback timing, $F(1, 108) = 2.23$, $p = .14$, $\eta^2_{part} = .02$ (a small effect size), or feedback type, $F(1, 108) = 2.38$, $p = .13$, $\eta^2_{part} = .022$ (a small effect size). There was a significant interaction effect for time x feedback timing, $F(1, 108) = 5.08$, $p = .026$, $\eta^2_{part} = .045$ (a small effect size), with similar probabilities of answering correctly on the 5-minute-delayed posttest after answering incorrectly on the pretest for participants who

received item-by-item and end-of-test feedback, but higher probability of answering correctly on the 1-week-delayed posttest for participants who received item-by-item feedback (Fig. 10). There were no significant interaction effects for time x feedback type, $F(1, 108) = 0.73$, $p = .40$, $\eta^2_{part} = .007$ (a very small effect size), feedback timing x type, $F(1, 108) = 0.008$, $p = .93$, $\eta^2_{part} < .001$ (a very small effect size), or time x feedback timing x feedback type, $F(1, 108) = 0.13$, $p = .72$, $\eta^2_{part} = .001$ (a very small effect size).

Figure 10: Probability of selecting the correct response on an item on the 5-minute-delayed posttest (R_2/R_1) or 1-week-delayed posttest (R_3/R_1), given that it was answered incorrectly on the pretest, new items only. IBI = item by item; EOT = end of test.



As above for the repeated items, I do not believe that this analysis is reliable because of the volatility of the conditional probabilities for participants who answered very few question incorrectly on the pretest. I ran the same analysis after removing all participants with 3 or fewer incorrect responses on the pretest (n = 64), and the results showed no significant effects. This lends some support to the claim that the analysis of the conditional probabilities above is not reliable.

4.4 Summary of Results

Below, the significant results for each research question are summarized in Table 16. The display time results are summarized in Table 17. Following that, I summarize the answer to each research question based on the results.

4.4.1 Research Question 1a: Item learning

This question addresses whether the timing (item-by-item or end-of-test) and type (metalinguistic/no metalinguistic) of feedback affected students' scores on posttests on the same questions (i.e., items that the participants got feedback on during the treatments). Thus, this question is about item learning. Overall, the item-by-item metalinguistic feedback was superior to the other timing and type combinations, although the effect sizes involved were relatively small.

Table 16: Research Questions and Results

Research question		Significant ANOVA results	Direction of effect
1. Does timing & type of feedback affect gain scores on 5-minute- & 1-week-delayed posttests	a. on repeated questions?	Main effect of time, $F(1, 108) = 27.83$, $p < .001$, $\eta^2_{part} = .21$	5-min.-delayed posttest > 1-week-delayed posttest
		Interaction of time x feedback timing, $F(1, 108) = 4.20$, $p = .043$, $\eta^2_{part} = .037$	5-min.-delayed posttest: IBI > EOT 1-week-delayed posttest: IBI ≈ EOT (Fig. 6)
		Interaction of time x feedback timing x feedback type, $F(1, 108) = 4.90$, $p = .029$, $\eta^2_{part} = .043$	Meta groups dropped in gain scores at same rate from 5-min.-delayed to 1-week delayed posttest. IBI group without meta showed sharper drop in gain scores from 5-min.-delayed to 1-week-delayed posttest, while EOT group without meta showed very little decrease in gain scores (Fig. 7)
	b. on new questions?	Main effect of time, $F(1, 108) = 41.22$, $p < .001$, $\eta^2_{part} = .28$	5-min.-delayed posttest < 1-week-delayed posttest
		Nearly sig. main effect of feedback timing, $F(1, 108) = 3.61$, $p = .060$, $\eta^2_{part} = .032$	IBI > EOT

Table 16 (cont'd)

2. Does conditional probability of correctly answering question on 5-minute- and 1-week-delayed posttests differ for groups based on timing & type of feedback for questions initially answered	a. correctly	on repeated questions?	No sig. results	
		on new questions?	No sig. results	
	b. incorrectly	on repeated questions?	Interaction of time x feedback timing x feedback type, $F(1, 108) = 9.15$, $p = .003$, $\eta^2_{part} = .078$	5-min.-delayed posttest: EOT meta & IBI no meta > IBI meta & EOT no meta 1-week delayed posttest: EOT meta & IBI no meta < IBI meta & EOT no meta (Fig. 10)
		on new questions?	Interaction of time x feedback timing, $F(1, 108) = 5.08$, $p = .026$, $\eta^2_{part} = .045$	5-min.-delayed posttest: IBI ≈ EOT 1-week-delayed posttest: IBI > EOT (Fig. 11)

Note. IBI = item by item; EOT = end of test; meta = metalinguistic feedback.

Table 17: Results of ANOVAs of Display Times

Main effect or interaction	Repeated items (questions)	New items (questions)	Feedback
Time	**$F(2.4, 254.4) = 192.6$, $p < .001$, $\eta^2_{part} = .64$**	**$F(1, 108) = 179.1$, $p < .001$, $\eta^2_{part} = .62$**	**$F(1, 108) = 18.37$, $p < .001$, $\eta^2_{part} = .15$**
		Display times became shorter over time	
Feedback timing	$F(1, 108) = 0.23$, $p = .64$, $\eta^2_{part} = .002$	$F(1, 108) = 0.68$, $p = .41$, $\eta^2_{part} = .006$	**$F(1, 108) = 4.14$, $p = .044$, $\eta^2_{part} = .037$** **EOT > IBI**
Feedback type	$F(1, 108) = 0.25$, $p = .62$, $\eta^2_{part} = .002$	$F(1, 108) = 0.21$, $p = .65$, $\eta^2_{part} = .002$	**$F(1, 108) = 6.90$, $p = .010$, $\eta^2_{part} = .060$** **Meta > no meta**
Time x feedback timing	$F(2.4, 254.4) = 0.33$, $p = .75$, $\eta^2_{part} = .003$	$F(1, 108) = 0.024$, $p = .88$, $\eta^2_{part} < .001$	$F(1, 108) = 3.31$, $p = .072$, $\eta^2_{part} = .030$
Time x feedback type	$F(2.4, 254.4) = 0.99$, $p = .39$, $\eta^2_{part} = .009$	$F(1, 108) = 0.37$, $p = .54$, $\eta^2_{part} = .003$	$F(1, 108) = 2.30$, $p = .13$, $\eta^2_{part} = .021$
Feedback timing x feedback type	$F(1, 108) = 2.54$, $p = .072$, $\eta^2_{part} = .023$	$F(1, 108) = 1.92$, $p = .17$, $\eta^2_{part} = .017$	$F(1, 108) = 1.06$, $p = .31$, $\eta^2_{part} = .010$
Time x feedback timing x feedback type	$F(2.4, 254.4) = 1.77$, $p = .17$, $\eta^2_{part} = .016$	**$F(1, 108) = 6.29$, $p = .014$, $\eta^2_{part} = .055$**	$F(1, 108) = 0.18$, $p = .67$, $\eta^2_{part} = .002$

Note. Bolded results are significant at the level of $p < .05$. IBI = item-by-item feedback; EOT = end-of-test feedback; meta = metalinguistic feedback. Significant three-way interaction effect for new items: The four groups show some differences in how long they displayed the new questions on the pretest, but clustered closer together on the 5-minute-delayed posttest.

Two significant or nearly significant results support this conclusion. First, the larger gain scores for the item-by-item group on the 5-minute delayed posttest show that the item-by-item feedback was more effective in the short term for item learning, although in the longer term, both feedback timings were equally effective. At first glance, it may seem that neither timing was terribly effective in the longer term, with both groups gaining less than one point on average from the pretest to the 1-week-delayed posttest. However, the total number of repeated items was 17, and many of the participants (n = 42) answered only 3, 4 or 5 of these items incorrectly on the pretest. A second result supports the conclusion that item-by-item metalinguistic feedback was superior: The interaction between time, feedback timing, and feedback type showed that the group that received item-by-item metalinguistic feedback had higher gain scores on both posttests than all other groups. Still, the item-by-item metalinguistic feedback group's average gain from the pretest to the 1-week-delayed posttest was less than 1.5 points.

4.4.2 Research Question 1b: System learning

This question addresses whether the timing (item-by-item or end-of-test) and type (metalinguistic/no metalinguistic) of feedback affected participants' scores on posttests on new questions (i.e., items that the participants did not encounter or get feedback on during the treatments). Thus, this question is about system learning. The results support the superiority of item-by-item feedback generally, regardless of the type of feedback, with the caveat that the effect approached significance, with a relatively small effect size.

4.4.3 Research Questions 2a and b: Reinforcing correct responses and correcting errors

Research Question 2a addresses whether differing feedback timings and types affect the degree to which correct responses were reinforced from the pretest to each posttest. However, no significant main effects or interactions were found for any of the groups on either repeated or

new items. Therefore, I found no evidence to support a claim of differences due to feedback timing or type on these conditional probabilities.

Research Question 2b addresses whether differing feedback timings and types affect the degree to which a participant corrects initially incorrect responses from the pretest to each posttest. The results indicate a complicated answer.

First, looking at item learning, the interaction of time x feedback timing x feedback type was significant, with a small to medium effect size. The results showed that in the shorter term, the group that got end-of-test metalinguistic feedback and the group that got item-by-item feedback without metalinguistic information corrected more of their initially incorrect responses than did the group that got item-by-item metalinguistic feedback and the group that got end-of-test feedback without metalinguistic information. In the longer term, this trend reverses, such that the groups most likely to correct their initial errors were those that got item-by-item metalinguistic feedback and the group that got end-of-test feedback without metalinguistic information. However, due to the volatility of the conditional probabilities, I do not believe that this interaction should be interpreted.

Next, for system learning, the group that got item-by-item feedback showed similar or better conditional probabilities than the end-of-test group on both posttests, with a small effect size. If one were to interpret this effect at face value, it would indicate that for system learning, item-by-item feedback was superior to end-of-test feedback in helping students to correct errors, especially on the 1-week-delayed posttest. However, as with the interaction effect above, I believe that the volatility of these conditional probabilities limits their usefulness.

CHAPTER 5: DISCUSSION

Below, I summarize the findings of the current study in relation to the purpose, which was to provide evidence of how varied feedback timing and type affect the learning of English article rules by adult learners. Following that is a discussion of possible explanations for the findings related to Research Question 1, then a discussion of the two interaction effects found for item learning. Finally, I discuss some potential reasons why the results of the current study differed from those of previous studies with similar designs.

5.1 Summary of Findings

Some of the results reported above support the cognitive processing window theory (Doughty, 2001) and the SLA attention-based theory (e.g., Gass, 1997; Pica, 1994; Schmidt, 1995). For item learning, the item-by-item metalinguistic feedback group showed greater mean gain scores than the other feedback groups, lending support to both the cognitive processing window theory and the SLA attention-based theory. For system learning, the item-by-item feedback groups showed higher mean gain scores than the end-of-test feedback groups on the 5-minute-delayed posttest, and the item-by-item metalinguistic feedback group had the highest mean gain scores on the 1-week-delayed posttest, again lending support to the cognitive processing window and the SLA attention-based theory. The results for the reinforcement of correct responses provided no evidence to support either theory, either for item or system learning. The results of error correction for item and system learning also provided no evidence due to the unreliability of the conditional probabilities. Because none of the results showed that end-of-test feedback was more effective than item-by-item feedback, no support was found for either of the educational psychology theories examined above, that is, the interference-perseveration hypothesis (Kulhavy & Anderson, 1972) and the dual-trace hypothesis.

5.2 Research Question 1

In addition to the analyses of the gain scores, I further investigated the results for Research Questions 1a and b by analyzing the time that the participants spent with the questions and feedback displayed on their screens. Below, I discuss how these further analyses show that this display time does not explain the item and system learning results, and I provide more plausible explanations for these results.

5.2.1 Research Question 1a: Item learning

As a possible explanation for why item-by-item metalinguistic feedback was generally superior to end-of-test feedback and feedback without metalinguistic information for item learning, I explored the relative time that the participants spent with questions and feedback displayed. No statistically significant differences were found in the question display times for the repeated questions, other than an overall decrease in display time as the participants moved through the tests and treatments, which was also seen for the new item and feedback display times. For the feedback display times, the metalinguistic feedback was displayed longer than the feedback without metalinguistic information, and the end-of-test feedback was displayed longer than the item-by-item feedback. Therefore, the length of time that the feedback was displayed might explain why the metalinguistic feedback was more effective, but not why the item-by-item feedback was more effective.

None of the predictions of the educational psychology theories for item learning, which include the dual-trace hypothesis (Clariana et al., 2000; Glover, 1989; Kulik & Kulik, 1988; Rankin & Trepper, 1978) and interference-perseveration hypothesis (Kulhavy & Anderson, 1972), were borne out by the results. A more convincing theoretical explanation of the differences in the gain scores among the groups is provided by two of the theories underlying the

interaction approach: the limitations of the cognitive processing window (Doughty, 2001) in combination with the attention-based explanation (e.g., Gass, 1997; Pica, 1994; Schmidt, 1995). As predicted by the cognitive processing window theory, the feedback that was provided immediately after a participant answered a question was more effective than feedback that was provided at the end of the test. In addition, as predicted by the attention-based explanation, metalinguistic feedback was more effective than feedback without metalinguistic information. Although it was not possible to make a theoretical prediction for the interaction between feedback timing and type, the results showed that the effects were, to some degree, additive, with item-by-item metalinguistic feedback the most effective condition.

5.2.2 Research Question 1b: System learning

For system learning, the item-by-item feedback groups showed higher mean gain scores than the end-of-test feedback groups on the 5-minute-delayed posttest, and the item-by-item metalinguistic feedback group had the highest mean gain scores overall. The display times for the questions and feedback were analyzed in an effort to explain this result. For the question display times, the mean display times decreased over the course of the study for all participants. The only other significant result was a three-way interaction of time x feedback timing x feedback type. This showed that on the pretest, the item-by-item metalinguistic feedback group spent, on average, longer with the new items displayed than the other groups did. This could explain why the item-by-item metalinguistic group performed better than the other groups. However, because the pretest was identical for all feedback groups, I cannot explain why the item-by-item metalinguistic group spent longer reading the questions on this test. It may have been due to individual differences that randomly were concentrated in this group. Thus, if the time spent reading the questions is indeed the reason for the higher gain scores, this result may

not be replicable. If, on the other hand, the nature and timing of the feedback is what caused the higher gain scores, rather than the time spent reading the questions, then the results will be replicable.

No analyses could be run for the feedback display times on the new items because, by design, no feedback was ever given on these items. However, the feedback display times for the repeated items might explain the gain scores on the new items, given that the feedback was also relevant to the new items. The groups that got end-of-test feedback displayed the feedback significantly longer than the item-by-item groups did, which cannot explain the fact that the item-by-item groups had higher gain scores. The metalinguistic feedback groups also displayed the feedback significantly longer than the feedback groups who did not receive metalinguistic information, and no corresponding relationship was seen in the gain scores. Thus, the feedback display times do not explain the gain scores.

Given the inconsistent relationships seen between the gain scores and the question and feedback display times for the new questions, the display times are not a convincing explanation for the difference between the groups in gain scores. As with item learning, the results for system learning corresponded to the predictions made based on the cognitive processing window theory, with no support for any of the educational psychology theories mentioned. However, unlike the results for item learning, the results for system learning do not provide any support for the SLA attention-based theory because no significant differences were found between the groups that got feedback with and without metalinguistic information.

Another interesting difference between the results for item learning and system learning can be seen in the trends from the 5-minute-delayed posttest to the 1-week-delayed posttest. Although item learning decreased on average from the first to second posttest, system learning

conversely increased. In both cases, the effect sizes were large. A possible explanation for this is that for item learning, the learners did not necessarily need to deeply process the metalinguistic information they received as feedback; they only needed to memorize the correct responses. Thus, a week after receiving feedback for the last time, they had forgotten some of the feedback. While many educational psychology studies used only one posttest, those that used more than one tended to find a similar drop in item learning over time (e.g., Brosvic et al., 2006a, 2006b; Clariana, Ross, & Morrison, 1991). On the other hand, for system learning, it may be that the learners benefitted from the additional time to process the rules and/or exemplars provided in the feedback (e.g., Mackey, Gass, & Mcdonough, 2000, p. 474).

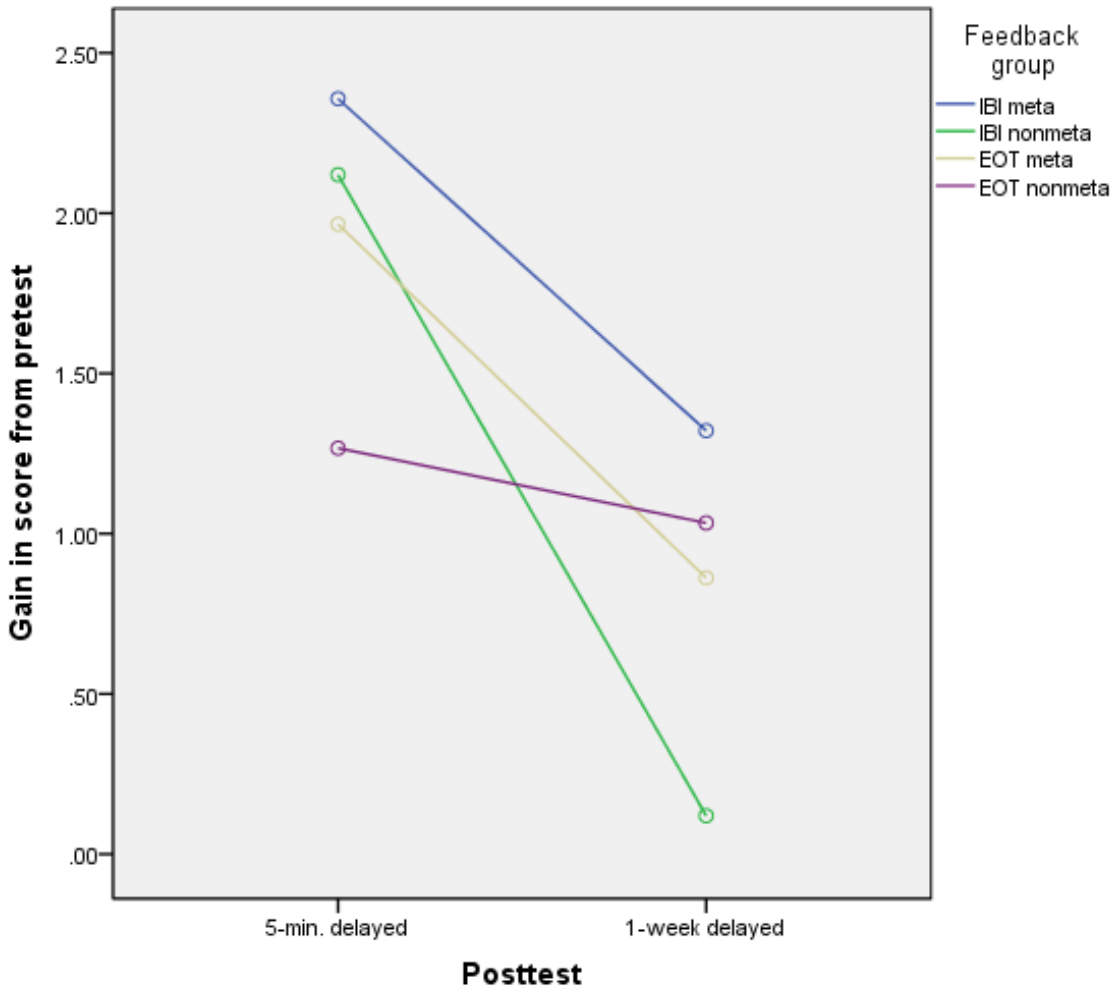
5.3 Interaction Effects

I found two significant interaction effects, both for item learning. Because these effects were not fully predicted by any of the theories detailed above, I consider here what may have caused them.

5.3.1 Time x feedback timing x feedback type interaction for item learning

First, an interaction effect was found for time x feedback timing x feedback type when examining item learning. The graphical representation of this effect is repeated below as Figure 11. The two metalinguistic feedback groups dropped in their gain scores at approximately the same rate from the 5-minute-delayed to the 1-week delayed posttest, while the groups that received feedback without metalinguistic information showed differing patterns. The gain scores of the group that got item-by-item feedback without metalinguistic information dropped sharply from the first to the second posttest, while the gain scores of the group that got end-of-test feedback without metalinguistic information remained nearly the same.

Figure 11: (Duplicate of Figure 6.) Gain scores of feedback groups on both posttests, repeated items only. IBI = item by item feedback; EOT = end of test feedback; meta = metalinguistic feedback; nonmeta = no metalinguistic feedback.



Before proceeding, consider that although I have mentioned that item learning may be the result of memorization, in fact, the repeated items can be answered correctly either by memorizing the correct response or by applying a metalinguistic rule. Therefore, it should not be unexpected that on the repeated items, the groups that got metalinguistic feedback did as well as or better than the groups that did not get metalinguistic feedback, simply because they had more potential resources to draw on. In addition, the downward trend in gain scores for all groups is

likely due to the participants forgetting the memorized items (e.g., Brosvic et al., 2006a, 2006b; Clariana, Ross, & Morrison, 1991), although forgetting may have also affected their memory of the metalinguistic rules. In addition, getting feedback on an item-by-item basis, even without metalinguistic information, may have provided an advantage on the 5-minute-delayed posttest for memorizing the correct answers because the feedback was provided within the cognitive processing window. This explains why the group that got end-of-test feedback without metalinguistic information had the lowest mean gain score on the 5-minute-delayed posttest. What remains to be explained is why the downward trends occurred at different rates.

To explain these effects, a few assumptions are necessary. Because the results in the current study that agree well with the cognitive processing window theory and SLA attention-based theory, I assume that these theories apply to this interaction effect as well. First, let us assume that the groups that got metalinguistic feedback were better able to explicitly learn the metalinguistic rules than the groups that did not get metalinguistic feedback, as predicted by the SLA attention-based theory. Let us also assume that the item-by-item feedback groups had an easier time memorizing the answers because the feedback was provided within the cognitive processing window. Finally, while both memorized responses and metalinguistic rules may be forgotten, metalinguistic information (implicit or explicit) may take time to be fully processed (e.g., Mackey, Gass, & Mcdonough, 2000, p. 474). Therefore, a stronger effect of forgetting may be seen for memorized responses than for metalinguistic rules on the 1-week-delayed posttest.

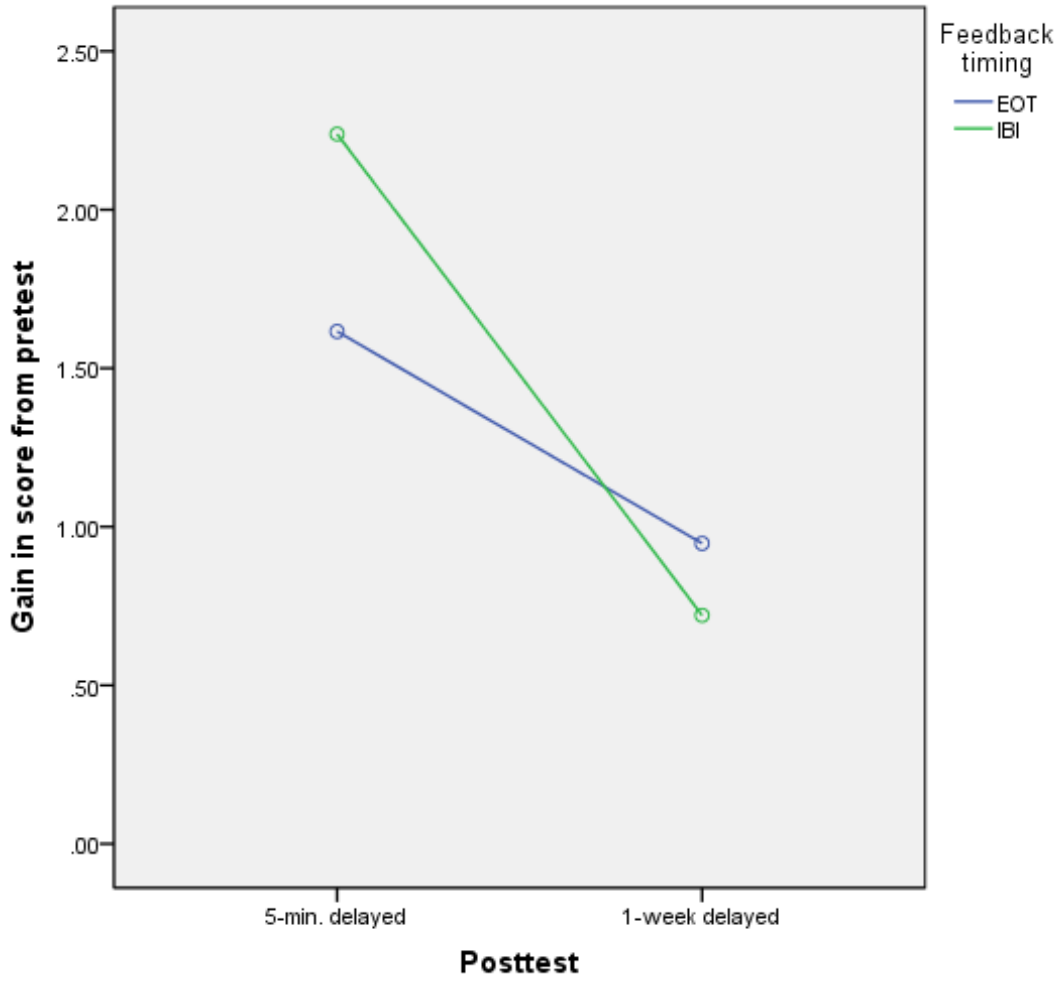
With these assumptions in place, I can explain the differing rates at which the four groups decreased in gain scores from the 5-minute-delayed posttest to the 1-week-delayed posttest. The item-by-item feedback group that did not get metalinguistic information had a relatively easy time memorizing the correct responses to answer the questions on the 5-minute-delayed posttest

because these participants received feedback within the cognitive processing window, but they forgot most of the correct responses by the time of the 1-week delayed posttest. Without much metalinguistic information to fall back on, this group's mean gain scores fell to around zero by the time of the delayed posttest. On the other hand, the end-of-test feedback group that did not get metalinguistic information did not do as well at memorizing the correct responses because the answers were not provided within the cognitive processing window. However, because all of the feedback was provided consecutively, without being interrupted by the need to respond to new questions, this group may have been better able to implicitly or explicitly derive the rules from the feedback they did receive, compared to the item-by-item groups. Therefore, their mean gain score fell less from the first to second posttest.

5.3.2 Time x feedback timing interaction for item learning

Next, an interaction effect was found for time x feedback timing in item learning (Figure 12). On the 5-minute-delayed posttest, the participants who got item-by-item feedback outperformed the participants who got end-of-test feedback, while the mean scores were nearly the same on the 1-week-delayed posttest, with the end-of-test group even doing somewhat better than the item-by-item group. This interaction may best be understood by considering the three-way interaction that was explained above.

Figure 12: (Duplicate of Figure 5.) Gain scores of feedback groups on both posttests, repeated items only. IBI = item by item feedback; EOT = end of test feedback.



Because these two interaction effects come from the same analysis, this two-way interaction can be understood as the result of averaging the effects seen in the three-way interaction above. That is, the item-by-item feedback group here is the average of the item-by-item feedback groups (that did and did not get metalinguistic information) in the three-way interaction, and the end-of-test feedback group here is the average of the end-of-test feedback groups (that did and did not get metalinguistic information) in the three-way interaction. Thus, the explanation above also applies here.

5.4 Why Results Differ Between Current Study and Previous Literature

Given that previous research on feedback timing and metalinguistic feedback found a wide range of results from a wide range of initial conditions, the results of the current study inevitably differ from some of the previous results. Below, I consider why my results differ from those of studies that were similar in design to the current study and from those of studies that made generalizations that appear to apply to the current study.

The design of the current study is most similar to that of Henshaw (2011). However, some of the results differ, with Henshaw finding no significant difference between feedback provided on an item-by-item basis and at the end of a test. A potential explanation for this difference is that the design of Henshaw's (2011) study included instruction that was provided before the treatments and feedback, whereas no instruction was provided in the current study. According to Goo and Mackey (2013), the effects of instruction may mitigate the effects of feedback (p. 152). Thus, the lack of a significant difference between the item-by-item and end-of-test feedback found by Henshaw may have been due to the effectiveness of the instruction, which all participants received.

In addition, the current results are somewhat at odds with the conclusions of Shute (2008), who distilled the disparate results of the educational psychology literature into recommendations for providing formative feedback. She found that immediate feedback (left defined relative to delayed feedback) was generally preferable to delayed feedback, which is clearly consistent with the current results. However, she also suggested that delayed feedback might be better for promoting the transfer of learning, citing Kulhavy et al. (1985) and Schroth (1992). In contrast, the current study showed that item-by-item feedback was superior to end-of-test feedback for both item and system learning. One obvious difference between the contexts is that Shute's

conclusion was drawn based on studies that did not involve language learning, and she may not have intended transfer to include rule learning. In addition, Shute's suggestion about delayed feedback, by her own admission, needed more research.

Two researchers have demonstrated that relatively immediate feedback is more effective for easy items, while relatively delayed feedback is more effective for difficult items (Clariana et al., 2000; Kulhavy & Anderson, 1972). I did not determine difficulty ratings for the items in the current study, and it would be useful to address this in future studies. However, to some extent, the participants themselves determined which questions were easy for them personally by answering them correctly on the pretest. These questions can be considered subjectively easy for each participant. Conversely, the questions that were difficult for individual participants were those that they answered incorrectly on the pretest. These questions can be considered subjectively difficult. Thus, the analyses of the conditional probabilities in which the participant answered correctly (R_2/R_1 and R_3/R_1) or incorrectly (R_2/W_1 and R_3/W_1) on the pretest are relevant here. The former analyses did not produce any significant results, and I argued that the later results were unreliable, both for item and system learning. None of these results support the generalization that immediate feedback is better for easy items and delayed feedback better for difficult items. The reason for this discrepancy between the current study and the previous research may be that previous studies determined item difficulty not by individual participant, but for the population overall, using data from previous research (Clariana et al., 2000) or the researchers' intuition (Kulhavy & Anderson, 1972). In addition, the previous research did not involve second-language learning, which by its nature may be different from other types of learning.

On a related note, Goo (2011) found that metalinguistic feedback was more effective

when the rule involved was a simple one, whereas nonmetalinguistic feedback was more effective when the rule was complex. Goo does not precisely define complexity, so it is difficult to say how the rules in the current study would be classified. Because they are not purely syntactic rules (i.e., semantic features must be taken into account to use the rules appropriately), they might be classified as complex. Therefore, simpler rules than the ones used in the current study might show a stronger effect for metalinguistic feedback and potentially provide more support for the SLA attention-based theory.

CHAPTER 6: CONCLUSION

Below, I summarize the major findings of the current study, then explore the pedagogical, CALL, theoretical, and research implications. Following that, I examine the limitations of the study and some possible future directions.

6.1 Summary of Findings

Overall, the results of the current study show that item-by-item feedback is superior to end-of-test feedback for both item learning and system learning, a result suggested by the findings of Lai, Fei, and Roots (2008), Nagata (1996), and Sakai (2004). The results also provide evidence that providing metalinguistic feedback is superior to not providing metalinguistic feedback for item learning, in agreement with the findings of Li (2010) and Lyster and Saito (2010). Little evidence was found in support of either item-by-item or end-of-test feedback timing or providing or not providing metalinguistic feedback for the reinforcement of correct responses or error correction, either for item or system learning. Thus, the bulk of evidence in this study is in support of the cognitive processing window theory (Doughty, 2001), and some of the evidence also supports the SLA attention-based theory (e.g., Gass, 1997; Pica, 1994; Schmidt, 1995). In addition, the results show that metalinguistic feedback provided on an item-by-item basis may provide more of an advantage than other combinations of feedback timing and type.

Interestingly, the results of the current study support both the efficacy of item-by-item feedback and its efficiency. That is, the participants who got the item-by-item feedback spent less time with the feedback displayed on the screen than did the learners who got the end-of-test feedback. No significant differences were found between reading times for the questions themselves in terms of feedback timing, with the exception of the item-by-item metalinguistic feedback group spending longer reading the new item questions on the pretest, which is

unrelated to the treatment. Although the results showed some evidence for the effectiveness of metalinguistic feedback, the participants spent more time reading this type of feedback than did those who did not get metalinguistic feedback. Thus, metalinguistic feedback is not more efficient than feedback without metalinguistic information.

6.2 Pedagogical and CALL Implications

One could argue that immediate feedback is the current best practice in computer-assisted language learning. However, I have shown in the literature review, *immediate* is interpreted in various ways, and both item-by-item and end-of-test feedback have been regarded as immediate. We cannot blithely assume that all "immediate" feedback timings are the same; in fact, the current study shows that item-by-item and end-of-test feedback may not be equivalent in terms of student learning. From a practical standpoint, a teacher might most effectively provide students with both item-by-item feedback, which may offer learning advantages, and end-of-test feedback, which students could use for further review, as suggested by Cohen (1985). However, if efficiency is a concern, learners may have the least feedback to read while still getting the most effective results if feedback is provided on an item-by-item basis, but with metalinguistic feedback provided only for questions that are answered incorrectly, similarly to Crooks's (1988) suggestion of providing informational feedback beyond knowledge of results only for errors.

The current study provides evidence that item-by-item feedback is superior to end-of-test feedback, which lends some support to the effectiveness claims of the companies that produce CALL applications, such as Rosetta Stone, Tell Me More, Duolingo, Open English, and Pimsleur, which all provide item-by-item feedback on at least some of their exercises. It also lends some support to researchers who have claimed that immediate feedback is superior to delayed feedback (e.g., Amaral & Meurers, 2011; Heift, 2010; Nagata & Swisher, 1995; Nagata, 1999).

However, I caution researchers, materials designers, and companies to use precise terms, such as those in Dempsey and Wager's (1988) taxonomy of feedback timing in computer-based instruction, rather than *immediate* and *delayed*. I expand on this caution in the following section. In addition, while I did find a statistically significant difference between item-by-item and end-of-test feedback in some cases in the current study, the effect sizes were small. Therefore, I question the emphasis that is placed on the immediacy of feedback by both researchers and CALL companies. While item-by-item feedback may indeed be better than end-of-test feedback in some situations, simply providing feedback on this schedule is unlikely to make a large difference in what students learn. Rather, it is one factor among many that may help students to learn more effectively and efficiently. In other situations, item-by-item feedback may not be appropriate or as effective as end-of-test feedback, and future research is needed to determine when item-by-item feedback is best.

6.3 Theoretical and Research Implications

I draw several implications for SLA research and theory from the current study. First, the current results must not be extrapolated to claim that immediate feedback is better than delayed feedback. Second, in certain situations, researchers can use conditional probabilities to glean information from data that may otherwise remain hidden. Third, multiple-choice questions can provide a forum for learners to notice a contrast between their own output and a native-like version. Finally, I consider the suggestion that metalinguistic feedback and item-by-item feedback timing have additive effects. I elaborate on each of these implications below.

While the results of the current study show that item-by-item feedback was on average more effective than end-of-test feedback, this should by no means be extrapolated to suggest that in general, relatively immediate feedback is more effective than relatively delayed feedback. For

example, the current study provides no evidence related to the relative efficacy of end-of-test feedback and feedback provided at a 24-hour delay. Based on the cognitive processing window theory, which was generally supported by the current results, I would expect end-of-test and 24-hour-delayed feedback to both be less effective than item-by-item feedback. However, their relative effectiveness that is an empirical question that is worth addressing in future studies.

Next, I consider the utility of examining conditional probabilities. My original reason for analyzing them was to investigate differences in the predictions of the educational psychology theories, and no evidence was found to support any of those theories that would necessitate examining the conditional probabilities. However, the results obtained from investigating the conditional probabilities also had the potential to provide more detailed insight into what the participants learned correctly or incorrectly through their participation in the study. The conditional probabilities for error correction were volatile due to many of the participants answering relatively few questions incorrectly on the pretest, which meant that the results were not reliable. Ideally, tests for which conditional probabilities are used would be more difficult and longer than the one used here, providing a bigger pool of items that are answered both correctly and incorrectly on the pretest for each participant. Given appropriate data, I believe that conditional probabilities can make a useful contribution to an analysis. In addition, in the current study, I did not examine whether the incorrect responses selected were the same from the pretest to each posttest (in part because there were only three choices), but that has the potential to provide yet more information about the effect of a treatment. Despite the common use of conditional probabilities in educational psychology research (e.g., Butler et al., 2007; Clariana et al., 2000; Peeck & Tillema, 1978; T. A. Smith & Kimball, 2010; Surber & Anderson, 1975), in the SLA literature, results have not traditionally been broken down in this fashion. SLA

researchers are (or should be) interested in both corrective feedback and in feedback that helps reinforce the knowledge of correct answers. Therefore, we should be interested in the type of conditional probabilities that were used in the current study. We should not neglect the potentially revealing information available to us from conditional probabilities, given appropriate data.

Next, I consider the role of multiple-choice questions in providing an opportunity for learners to notice a contrast between their own language and that of a native speaker. Long (1996) stated that implicit negative feedback may be especially useful when it comes immediately after an ungrammatical learner utterance because of learner attention to a response at that point in a conversation. In addition, Lai, Fei, and Roots (2008) and Sakai (2004) reported more noticing when recasts were provided on an item-by-item basis, either in typed CMC or spoken interaction, respectively. Based on this, I argued above that multiple-choice questions with correct answers given as feedback provide a forum for the contrast to be detected, especially when the feedback is provided on an item-by-item basis. Based on the results of the current study that item-by-item feedback was generally more effective than end-of-test feedback, multiple-choice questions may indeed provide this forum for noticing a contrast.

Finally, the results showed that item-by-item metalinguistic feedback was the most effective condition, which suggests that the effects of metalinguistic feedback and item-by-item feedback were complementary to some extent. This indicates that metalinguistic feedback may be more effective during the cognitive processing window than outside it, perhaps because the learner is cognitively comparing his or her current explicit metalinguistic understanding with the metalinguistic information provided in the feedback. However, this reasoning is speculative and should be investigated in future research.

6.4 Limitations

While the findings in the current study are a valuable first step toward determining the most effective and efficient feedback timing, some limitations should be kept in mind. Below, I expand on some of the important limitations.

Several limitations of the study are related to control. First, no true control group was included in the study. That is, no group received no feedback, so the results cannot tell us whether the feedback in the current study was more effective than no feedback. However, a vast literature has shown the advantages of feedback (e.g., the following meta-analyses: Li, 2010; Mackey & Goo, 2007; Russell & Spada, 2006), and denying a group the predicted benefits of feedback raises ethical issues (Gass, 2010a).

Because the current study was conducted in classrooms, some variables were uncontrollable. For instance, normal classroom interruptions occurred that may have distracted the participants from the tests, treatments, and feedback. The effect of these interruptions was not unduly distributed to any given feedback group, however, because the participants were randomly assigned to the groups within each class. Although the learners' ability to use articles was controlled by eliminating those who scored at ceiling on the pretest, I did not control for more general English proficiency. In addition, nearly all of the participants indicated that they had prior knowledge of some of the rules used in the study. It is possible that the results would have differed if the rules were entirely new to them.

Each participant received feedback on a given item only twice (once per treatment), and the last posttest was completed only one week after the last treatment and two weeks after the pretest. Thus, the current study does not contribute to our understanding of what the participants may have retained in the longer term.

The end-of-test feedback was by design given closer to the time of the 5-minute-delayed posttest than most of the item-by-item feedback. This may have given the participants who got end-of-test feedback an advantage on the 5-minute-delayed posttest, although the advantage should have been largely eliminated by the time of the 1-week-delayed posttest. The 5-minute delay was incorporated in an attempt to mitigate this advantage, but 5 minutes may not have been long enough to do so.

In the end-of-test feedback condition, upon receiving feedback, a learner could have reactivated an error in working memory by repeating the processing that occurred when he or she initially saw the item. I do not know how often this occurred in the current study. This reactivation could be the reason for the small effect sizes seen between the item-by-item and end-of-test feedback. It may possible to control this reactivation with a clever design, but a real classroom situation is nearly always going to allow for this reactivation.

The instructions for the pretest (and for the treatments and posttests) stated "Choose the best answer to fit the blank. For some questions, more than one answer is possible, so choose the best one" (Appendix C). This wording was chosen so that the participants would not be confused when they encountered the questions that had two possible answers. However, this choice of words may have been confusing to the participants because it implies that one answer is better than the other. A more helpful choice of wording may have been "For some questions, more than one answer is possible, so choose the one that you like best."

Finally, the small number of questions used in the tests may have reduced the effect sizes in this study. For example, only 10 new items were included in the analyses, and some participants who correctly answered 8 out of 10 of these items on the pretest were included in the analysis. Even if the treatments were as effective as possible, these participants could only have

gained 2 points on the posttests. Thus, a more difficult test may have led to larger effect sizes.

6.5 Future Directions

The optimal timing of feedback in the context of second language acquisition is an underresearched area, and interest in this area is only beginning to pick up speed (Aubrey & Shintani, 2014; Quinn, 2013; Sheen, 2012). Given the potential applications of this research for teachers, instructional designers, and companies that design CALL applications, I expect to see more research on this topic in the near future. In addition, there is a great need to understand how the timing of feedback interacts with the type of feedback and learners' prior knowledge to more effectively design learning. Below, I describe some future directions for this area of inquiry.

The context of this study was multiple-choice questions, and no writing, listening, speaking, or extended reading comprehension data was collected. Future studies are needed to determine whether the current results also apply to less controlled learner production and contexts in which learners cannot readily access their metalinguistic knowledge. Studies of other English structures as well as structures in other languages are also needed to extend the results.

Although I know how long the participants in the current study left the feedback and questions displayed on their screens, I do not know how long they spent reading. Previous studies have shown that users read feedback (Heift, 2001), but may skip or skim over feedback on questions that they have answered correctly (Pujolà, 2001) and feedback over three lines long (van der Linden, 1993). Given that the feedback in the current study was longer than three lines and that the feedback was displayed for relatively short mean times, the participants may indeed have skipped over much of it. Future studies could use eye-tracking technology to better determine how long participants actually spend reading feedback. Further insight into what the participants were thinking while they were choosing their response could also be gained by

recording the screen during the treatments and tests, possibly in combination with think-aloud protocols or stimulated recalls.

By manipulating whether the participants in the current study received metalinguistic feedback and when they received feedback, I tested the predictions of the cognitive processing window (Doughty, 2001) and SLA attention-based theories (e.g., Gass, 1997; Pica, 1994; Schmidt, 1995). However, other theories may also apply to the questions of when feedback should be provided and what the feedback should include. For example, cognitive load theory (e.g., Sweller, 1994) suggests that including more text on the screen may increase the cognitive load of participants with low memory resources, thus reducing their reading comprehension (Ardaç & Unal, 2008). This in turn suggests that providing metalinguistic feedback in addition to correct response feedback may increase L2 learners' cognitive load and reduce their learning. This prediction is not consistent with the results of the current study but merits further investigation. The current study raises the important question of when feedback should be delivered to learners, and the results provide the tentative answer that item-by-item feedback is both more efficient and more effective than end-of-test feedback. Based on these results, I hope that researchers, teachers, and instructional designers will be mindful of how they use the terms *immediate* and *delayed* regarding feedback and consider providing item-by-item feedback where possible.

APPENDICES

Appendix A: Consent Form

Michigan State University
IRB ID# i044441
Consent Form
Study Title: Effects of Feedback Timing on ESL Students' Learning of Articles

Researcher and Title: Elizabeth (Betsy) Lavolette, PhD candidate, Second Language Studies
Department and Institution: Department of Linguistics and German, Slavic, Asian and African Languages, Michigan State University
Address and Contact Information:
451 Glenmoor #2, East Lansing, MI 48823.
Phone: 808-224-0949
Email: betsy@msu.edu

Researcher and Title: Dr. Charlene Polio, Professor, Second Language Studies
Department and Institution: Department of Linguistics and German, Slavic, Asian and African Languages, Michigan State University
Address and Contact Information:
B251 Wells Hall, East Lansing, MI 48823.
Phone: 517-884-1502
Email: polio@msu.edu

1. **EXPLANATION OF THE RESEARCH and WHAT YOU WILL DO:**
   You are being asked to participate in a research study of how students learn from computers in language classrooms.
   In this study, you will take five tests on articles (a/an, the, no article). You will also fill out an opinion questionnaire and a background questionnaire.
   You must be at least 18 years old to participate in this research.

2. **YOUR RIGHTS TO PARTICIPATE, SAY NO, OR WITHDRAW:**
   This research is being done as part of your regular classroom work. You are being asked for your consent to use this classroom work for research purposes. Participation in this study is strictly voluntary; you have the right to say no. You are free to change your mind and withdraw at any time without consequence or penalty. Your grades will not be affected by your participation or not in this research.

3. **COSTS AND COMPENSATION FOR BEING IN THE STUDY:**
   There are no costs or compensation for participating in this study.

4. **CONTACT INFORMATION FOR QUESTIONS AND CONCERNS**
   If you have concerns or questions about this study, such as scientific issues, how to do any part of it, or to report an injury, please contact the researcher (Betsy Lavolette, 451 Glenmoor #2, East Lansing, MI 48823. Phone: 808-224-0949. Email: betsy@msu.edu).

Name _____

____ I agree to participate.

____ I do not agree to participate.

Appendix B: Article Rules

Rule 1: Use **no article** for transportation when it follows *by* or *via*.

Rule 2: Use **no article** for some places when they are used for their main purpose.

Rule 3: Use ***a/an*** when a noun does not refer to a specific thing or person—it's ANY [noun].

Rule 4: Use ***a/an*** when mentioning a thing or person that the reader (or listener) does not know about.

Rule 5: Use ***the*** when something after a noun makes it definite, especially descriptions starting with *that*.

Rule 6: Use ***the*** when the context makes a noun known to the reader (or listener).

Appendix C: Test Items

The experimental items are shown below, sorted according to the rule they exemplify. The filler items, which have two possible answers, are shown after the experimental items. Each item has three answer choices: a/an [noun], the [noun], and [noun] (without an article). The correct answers are shown below in brackets. For each type of experimental item below, all items appeared on the pretest and posttests, and the first four only appeared on the treatments. The items with a star had poor item discrimination and were not included in the analysis. All of the filler items appeared on the pretest and posttests, and the first eight only appeared on the treatments.

Instructions:

Choose the best answer to fit the blank.

For some questions, more than one answer is possible, so choose the best one.

Rule 1

1.1) Instead of driving, I want to travel by [horse].

1.2) The cheapest way is to go by [train].

1.3) A: How are you traveling to Washington? B: Via [airplane].

*1.4) You'll get there in 10 minutes if you go via [bicycle].

1.5) I don't like going to school by [bus] in the winter.

1.6) Let's not go via [car] because that will be slow.

Rule 2

2.1) She didn't go to [bed] until 1:00 a.m.

2.2) A: Did you knock on their door? B: Yes, but no one was at [home].

2.3) A: My son is a high school student and my daughter is a college student. B: Can I meet them? A: No, they're both at [school] now.

*2.4) Customer: If you don't return my money, I will take you to [court]!

2.5) A: I was too sick to go to my three classes last week. B: Do you feel well enough to go to [class] this week?

2.6) Even when I am traveling, I always go to [church] on Sundays.

Rule 3

*3.1) What kinds of animals have you ever ridden? Have you ridden [a horse]?

3.2) A: If you were rich, would you buy expensive cars? B: No, I would buy [an airplane].

3.3) I just moved into my house, and I need [a bed].

*3.4) They were looking for [a home] near the university.

3.5) I'm thinking of taking [a class] at the college this summer.

3.6) A: Have you seen everything in town? B: Not yet. I'd like to see [a church] if there are any here.

Rule 4

4.1) A: There are many trains in San Francisco, aren't there? B: Yes, and I read that today, [a train] caught fire during rush hour.

4.2) A: Did you buy the bicycle that we looked at yesterday? B: Not that one, but I did buy [a bicycle].

4.3) A: Did you take your problem to the court in Michigan? B: No, it was [a court] elsewhere.

4.4) I read that computers were stolen from [a school] somewhere in California.

*4.5) He is tired of riding the bus, so he's looking for [a car].

4.6) A: Are you OK? You look like you were hit by a car! B: No, it was [a bus].

Rule 5

*5.1) I was late because [the train] that I took to work was late.

5.2) A: Did you ride today? B: Yes, I rode [the horse] that is eating grass.

*5.3) I couldn't sleep because [the bed] in my hotel room was too hard.

5.4) My parents still live in [the home] that I grew up in.

5.5) In front of my house, I saw [the car] that my friend had just bought.

5.6) This morning at 8:00, [the class] that I had was not very interesting.

Rule 6

6.1) The captain checked the engines and wings, then we boarded [the airplane].

6.2) Its front tire was flat, so I could not use [the bicycle].

6.3) The judge asked me to read a statement to [the court].

*6.4) Many of the teachers spoke English, but [the school] did not have English classes.

*6.5) When our driver comes back, [the bus] will leave.

6.6) Do you see the large cross at the front of [the church]?

Fillers (Either "an/a" or "the" is possible)

7.1) [A/The horse] with black spots was standing in the field.

7.2) I took [a/the train] to Chicago.

7.3) Before my flight, I watched [a/the airplane] land smoothly.

7.4) He went back to the store to buy [a/the bicycle] with wide tires.

7.5) The room was only big enough to contain [a/the bed].

7.6) We plan to buy [a/the home] near the lake.

7.7) This desk is from [a/the school] in my town.

7.8) He grew up in [a/the town] directly west of Lansing.

7.9) A: Why did you stop? B: [A/The car] ahead of me stopped.

7.10) Just before I got to the bus stop, [a/the bus] drove past.

7.11) [A/The class] that I am taking now is interesting.

7.12) We saw [a/the church] in the middle of town.

Appendix D: Exit Questionnaire

1. Do you think that your ability to use articles (a, an, the) improved from practicing with the computer?

   yes/no

2. What part of the practice was helpful? What part was not helpful?
3. What would make the computer practice of articles (a, an, the) more useful to you?
4. Did you learn the following rules for using articles (a, an, the) before participating in this study?
   a. Use no article for transportation when it follows *by* or *via*.
   b. Use no article for some places when they are used for their main purpose.
   c. Use *a/an* when a noun does not refer to a specific thing or person—it's ANY [noun].
   d. Use *a/an* when mentioning a thing or person that the reader (or listener) does not know about.
   e. Use *the* when something after a noun makes it definite, especially descriptions starting with *that*.
   f. Use *the* when the context makes a noun known to the reader (or listener).
5. Age in years:
6. Gender

   Male Female Other

7. What is your first language?

   Mandarin (Chinese) Cantonese (Chinese) Arabic Korean

   Japanese Spanish Other ____

8. How many years have you studied English?
9. How many months total have you lived in an English-speaking country?
10. Any comments?

REFERENCES

REFERENCES

Ableeva, R. (2010). Dynamic assessment of listening comprehension in second language learning. (Doctoral dissertation). Available from ProQuest Dissertation and Theses Database. (UMI Number: 3436042).

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, England: Continuum.

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal*, *78*(4), 465–483.

Amaral, L. A., & Meurers, D. (2011). On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, *23*(01), 4–24. doi:10.1017/S0958344010000261

Ardaç, D., & Unal, S. (2008). Does the amount of on-screen text influence student learning from a multimedia-based instructional unit? *Instructional Science, 36*, 75–88. doi:10.1007/s11251-007-9035-4

Aubrey, S. C., & Shintani, N. (2014, March). The effects of synchronous and asynchronous written corrective feedback on grammatical accuracy in a computer-mediated environment. Paper presented at the meeting of the American Association of Applied Linguistics, Portland, OR.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research*, *13*(2), 111–127.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews: Neuroscience, 4*, 829–839. doi:10.1038/nrn1201

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213–238. doi:10.3102/00346543061002213

Barnes, D. (1996). Naming as a technical term: Sacrificing behavior analysis at the alter of popularity? *Journal of the Experimental Analysis of Behavior*, *65*(1), 264–267.

Bitchener, J., & Knoch, U. (2009). The relative effectiveness of different types of direct written corrective feedback. *System*, *37*(2), 322–329. doi:10.1016/j.system.2008.12.006

Brosvic, G. M., & Epstein, M. L. (2007). Enhancing learning in the introductory course. *The Psychological Record, 57*(3), 391–408.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006a). Acquisition and retention of Esperanto: The case for error correction and immediate feedback. *The Psychological Record, 56,* 205–218.

Brosvic, G. M., Epstein, M. L., Dihoff, R. E., & Cook, M. J. (2006b). Retention of Esperanto is affected by delay-interval task and item closure: A partial resolution of the delay-retention effect. *The Psychological Record, 56,* 597–615.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology, 1*(1), 412–414.

Brown, J. D. (2005). Item analysis in language testing. In *Testing in language programs* (pp. 66–88). New York, NY: McGraw-Hill.

Brown, J. D. (2008). Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 38–43.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290–298. doi:10.1037/a0031026

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13*(4), 273–281. doi:10.1037/1076-898X.13.4.273

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.

Chapelle, C. A. (2003). The potential of technology for language learning. In *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology* (pp. 35–68). Amsterdam, The Netherlands: John Benjamins.

Chun, D. M., & Brandl, K. (1992). Beyond form-based drill and practice: Meaning-enhancing CALL on the Macintosh. *Foreign Language Annals, 25*(3), 255–267.

Clariana, R. B., Ross, S. M., & Morrison, G. R. (1991). The effects of different feedback strategies using computer-administered multiple-choice questions as instruction. *Educational Technology Research & Development, 39*(2), 5–17.

Clariana, R. B., Wagner, D., & Roher Murphy, L. C. (2000). Applying a connectionist description of feedback timing. *Educational Technology Research & Development, 48*(3), 5–22.

Cohen, J. (1988). The analysis of variance. In *Statistical power analysis for the behavioral sciences* (2nd ed., pp. 273–406). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, V. B. (1985). A reexamination of feedback in computer-based instruction: Implications for instructional design. *Educational Technology*, *25*(1), 33–37.

Cook, V. (1989). Universal Grammar theory and the classroom. *System*, *17*(2), 169–181.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104. doi:10.1037//0021-9010.78.1.98

Cowan, N. (2005). The present theoretical approach. In *Attention and memory: An integrated framework.* (pp. 39–73). Oxford, England: Oxford University Press.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481.

Dabrowski, R., LeLoup, J. W., & MacDonald, L. (2013). Effectiveness of computer-graded vs. instructor-graded homework assignments in an elementary Spanish course: A comparative study at two undergraduate institutions. *The IALLT Journal*, *43*(1), 79–100.

Dekeyser, R. M. (1997). Beyong explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221.

Dekeyser, R. M. (2007). Skill acquisition theory. In B. Vanpatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Erlbaum.

Dempsey, J. V., & Wager, S. U. (1988). A taxonomy for the timing of feedback in computer-based instruction. *Educational Technology*, *28*(10), 20–25.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *The Psychological Record*, *54*, 207–231.

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

Doughty, C. J., & Long, M. H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Language Learning & Technology*, *7*(3), 50–75.

Ebel, R. L. (1979). How to improve test quality through item analysis. In *Essentials of educational measurement* (3rd ed., pp. 258–273). Englewood Cliffs, NJ: Prentice-Hall.

El Saadawi, G. M., Tseytlin, E., Legowski, E., Jukic, D., Castine, M., Fine, J., … Crowley, R. S. (2008). A natural language intelligent tutoring system for training pathologists: Implementation and evaluation. *Advances in health sciences education: Theory and practice*, *13*(5), 709–722. doi:10.1007/s10459-007-9081-3

Ellis, N. (2006). Cognitive perspectives on SLA: The associative-cognitive CREED. *AILA Review*, *19*, 100–121.

Ellis, N. (2008). Usage-based and form-focused SLA: The implicit and explicit learning of constructions. In A. Tyler, Y. Kim, & M. Takada (Eds.), *Language in the context of use: Discourse and cognitive approaches to language* (pp. 99–126). Berlin, Germany: Mouton de Gruyter.

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

English, R. A., & Kinzer, J. R. (1966). The effect of immediate and delayed feedback on retention of subject matter. *Psychology in the Schools*, *3*(2), 143–147.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, England: Sage.

Gagné, R. M. (1985). What is learned—Varieties. In *The conditions of learning and theory of instruction* (3rd. ed., pp. 46–69). New York, NY: Holt, Rinehart and Winston.

García, M. R., & Arias, F. V. (2010). A comparative study in motivation and learning through print-oriented and computer-oriented tests. *Computer Assisted Language Learning*, *13*(4–5), 457–465.

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

Gass, S. M. (2010a). Experimental research. In B. Paltridge & A. Phakiti (Eds.), *Continuum companions to research methods in applied linguistics* (pp. 7–21). London, England: Continuum.

Gass, S. M. (2010b). Interactionist perspectives on second language acquisition. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd. ed., pp. 217–231). Oxford, England: Oxford University Press.

Gaynor, P. (1981). The effect of feedback delay on retention of computer-based mathematical material. *Journal of Computer-Based Instruction*, *8*(2), 28–34.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*(3), 392–399. doi:10.1037//0022-0663.81.3.392

Goda, Y. (2004). Feedback timing and learners' response confidence on learning English as a foreign language (EFL): Examining the effects of a computer-based feedback and assessment environment on EFL students' language acquisition. (Doctoral dissertation). Available from ProQuest Dissertation and Theses Database. (UMI No. 3123098).

Goo, J. (2011). Corrective feedback, individual variation in cognitive capacities, and L2 development: Recasts vs. metalinguistic feedback. (Doctoral dissertation). Available from ProQuest Dissertation and Theses Database. (UMI Number: 3450729).

Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, *35*(1), 127–165. doi:10.1017/S0272263112000708

Guzmán-Muñoz, F. J., & Johnson, A. (2008). Error feedback and the acquisition of geographical representations. *Applied Cognitive Psychology*, *22*(7), 979–995. doi:10.1002/acp

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly*, *44*(1), 84–109. doi:10.5054/tq.2010.213781

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Hayes, S. C., Barnes-Holmes, D., & Roche, B. (Eds.). (2001). *Relational frame theory: A post-Skinnerian account of human language and cognition*. Hingham, MA: Kluwer.

Heift, T. (2001). Error-specific and individualised feedback in a Web-based language tutoring system: Do they read it? *ReCALL*, *13*(1), 99–109.

Heift, T. (2003). Drag or type, but don't click: A study on the effectiveness of different CALL exercise types. *Canadian Journal of Applied Linguistics*, *6*(1), 69–85.

Heift, T. (2004). Corrective feedback and learner uptake in CALL. *ReCALL*, *16*(2), 416–431. doi:10.1017/S0958344004001120

Heift, T. (2006). Context-sensitive help in CALL. *Computer Assisted Language Learning*, *19*(2–3), 243–259. doi:10.1080/09588220600821552

Heift, T. (2010). Developing an intelligent language tutor. *CALICO Journal*, *27*(3), 443–459.

Henshaw, F. (2011). Effect of feedback timing in SLA: A computer-assisted study on the Spanish subjunctive. In C. Sanz & R. P. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA and bilingualism* (pp. 98–113). Washington, DC: Georgetown University Press.

Jaehnig, W., & Miller, M. L. (2007). Feedback types in programmed instruction: A systematic review. *The Psychological Record*, *57*(2), 219–232.

Kane-lturrioz, R. (2008). Computer-based language assessment: A formative approach. *ReCALL*, *9*(1), 15–21. doi:10.1017/S0958344000004584

King, P. E., Young, M. J., & Behnke, R. R. (2000). Public speaking performance improvement as a function of information processing in immediate and delayed feedback interventions. *Communication Education*, *49*(4), 37–41.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.

Kregar, S. (2011). *Relative effectiveness of corrective feedback types in computer-assisted language learning*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3477247).

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 68*(5), 505–512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279–308.

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology*, *10*, 285–291.

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97. doi:10.2307/1170349

Lai, C., Fei, F., & Roots, R. (2008). The contingency of recasts and noticing. *CALICO Journal*, *26*(1), 70–90.

Lan, Y.-J., Sung, Y.-T., & Chang, K.-E. (2007). A mobile-device-supported peer-assisted learning system for collaborative early EFL reading. *Language Learning & Technology, 11*(3), 130–151.

Lantolf, J. P. (2008). Dynamic assessment: The dialectic integration of instruction and assessment. *Language Teaching*, *42*(3), 355–368. doi:10.1017/S0261444808005569

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*(1), 49–72. doi:10.1558/japl.1.1.49.55872

Lavolette, E. (2013). *Feedback timing effects on ESL students' learning of articles rules*. Unpublished manuscript.

Lavolette, E., Polio, C., & Kahng, J. M. (2013). *The usefulness of computer-mediated feedback in essay revision*. Manuscript submitted for publication.

Lee, L. (2008). Focus-on-form through collaborative scaffolding in expert-to-novice online interaction. *Language Learning & Technology*, *12*(3), 53–72. Retrieved from

http://llt.msu.edu/vol12num3/vol12num3.pdf?q=microsoft-word-700-mhz-faq-final#page=60

Lewis, M. W., & Anderson, J. R. (1985). Discrimination of operator schemata in problem solving: Learning from examples. *Cognitive Psychology*, *17*(1), 26–65. doi:10.1016/0010-0285(85)90003-9

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, *60*(2), 309–365. doi:10.1111/j.1467-9922.2010.00561.x

Lin, J.-W., Lai, Y.-C., & Chuang, Y.-S. (2013). Timely diagnostic feedback for database concept learning. *Educational Technology & Society*, *16*(2), 228–242.

Loewen, S. (2004). Uptake in incidental focus on form in meaning-focused ESL lessons. *Language Learning*, *54*(1), 153–188.

Loewen, S., & Erlam, R. (2006). Corrective feedback in the chatroom: An experimental study. *Computer Assisted Language Learning*, *19*(1), 1–14. doi:10.1080/09588220600803311

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York, NY: Academic.

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

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302. doi:10.1017/S0272263109990520

Mackey, A., Gass, S., & Mcdonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*(4), 471–497.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford, England: Oxford University Press.

Mandernach, B. J. (2005). Relative effectiveness of computer-based and human feedback for enhancing student learning. *The Journal of Educators Online*, *2*(1), 1–17.

Master, P. (2002). Information structure and English article pedagogy. *System*, *30*(3), 331–348. doi:10.1016/S0346-251X(02)00018-0

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*(8), 1077–1087. doi:10.3758/MC.37.8.1077

Moreno, N. (2007). *The effects of type of task and type of feedback on L2 development in CALL*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3302088).

Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*(4), 617–673.

Murphy, P. (2010). Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer-mediated communication. *ReCALL*, *22*(2), 112–134. doi:10.1017/S0958344010000030

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *Modern Language Journal*, *77*(3), 330–339.

Nagata, N. (1996). Computer vs. workbook instruction in second language acquisition. *CALICO Journal*, *14*(1), 53–75.

Nagata, N. (1997). The effectiveness of computer-assisted metalinguistic instruction: A case study in Japanese. *Foreign Language Annals*, *30*(2), 187–200.

Nagata, N. (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals*, *32*(4), 469–479. doi:10.1111/j.1944-9720.1999.tb00876.x

Nagata, N., & Swisher, M. (1995). A study of consciousness-raising by computer: The effect of metalinguistic feedback on second language learning. *Foreign Language Annals*, *28*(3), 337–347.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. doi:10.1111/0023-8333.00136

Opitz, B., Ferdinand, N. K., & Mecklinger, A. (2011). Timing matters: The impact of immediate and delayed feedback on artificial language learning. *Frontiers in Human Neuroscience*, *5*, 1–9. doi:10.3389/fnhum.2011.00008

Park, O., & Gittelman, S. S. (1992). Selective use of animation and feedback in computer-based instruction. *Educational Technology Research & Development*, *40*(4), 27–38.

Peeck, T., & Tillema, H. H. (1978). Delay of feedback and retention of correct and incorrect responses. *The Journal of Experimental Education*, *47*(2), 171–178.

Phye, G. D., & Andre, T. (1989). Delayed retention effect: Attention, perseveration, or both? *Contemporary Educational Psychology*, *14*, 173–185.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, *44*(3), 493–527.

Pienemann, M. (1998). An introduction to Processability Theory. In *Language processing and second language development: Processability theory* (pp. 1–73). Amsterdam, The Netherlands: John Benjamins.

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal*, *91*(3), 323–340.

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development*. New York, NY: Springer Science+Business Media.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research*, *17*(3), 323–342. doi:10.1177/1362168813482935

Pujolà, J.-T. (2001). Did CALL feedback feed back? Researching learners' use of feedback. *ReCALL*, *13*(1), 79–98.

Quinn, P. (2013, March). The effects of altering the timing of corrective feedback. Paper presented at the meeting of the American Association of Applied Linguistics, Dallas, TX.

Rankin, R. J., & Trepper, T. (1978). Retention and delay of feedback in a computer-assisted instructional task. *The Journal of Experimental Education*, *46*(4), 67–70.

Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *Modern Language Journal*, *88*(2), 192–216.

Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for the acquisition of L2 grammar. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 147–178). Amsterdam, The Netherlands: John Benjamins.

Sakai, H. (2004). Roles of output and feedback for L2 learners' noticing. *JALT Journal*, *26*(1), 25–54.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning*, *54*(1), 35–78.

Sauro, S. (2009). Computer-mediated corrective feedback and the development of L2 grammar. *Language Learning & Technology*, *13*(1), 96–120.

Saxton, M. (1997). The contrast theory of negative input. *Journal of Child Language*, *24*(1), 139–161. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9154012

Saxton, M., Backley, P., & Gallaway, C. (2005). Negative input for grammatical errors: Effects after a lag of 12 weeks. *Journal of Child Language*, *32*, 643–672. doi:10.1017/S0305000905006999

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu, HI: University of Hawai'i Press.

Schooler, L., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 702–708.

Schroth, M. L. (1992). The effects of delay of feedback on a delayed concept formation transfer task. *Contemporary Educational Psychology*, *17*, 78–82.

Schroth, M. L. (1995). Variable delay of feedback procedures and subsequent concept formation transfer. *The Journal of General Psychology*, *122*(4), 393–399.

Schroth, M. L., & Lund, E. (1993). Role of delay of feedback on subsequent pattern recognition transfer tasks. *Contemporary Educational Psychology*, *18*, 15–22.

Schwartz, B. D., & Gubala-Ryzak, M. (1992). Learnability and grammar reorganization in L2A: Against negative evidence causing the unlearning of verb movement. *Second Language Research*, *8*(1), 1–38. doi:10.1177/026765839200800102

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 301–322). Oxford, England: Oxford University Press.

Sheen, Y. (2012, October). The timing of corrective feedback and L2 learning. Paper presented at the Second Language Research Forum, Pittsburgh, PA.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189.

Skinner, B. F. (1968). *The technology of teaching*. New York, NY: Appleton-Century-Crofts.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 80–95. doi:10.1037/a0017407

Smits, M. H. S. B., Boon, J., Sluijsmans, D. M. A., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments*, *16*, 183–193.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263–308. doi:10.1111/j.1467-9922.2010.00562.x

Sturges, P. T. (1978). Delay of informative feedback in computer-assisted testing. *Journal of Educational Psychology*, *70*(3), 378–387. doi:10.1037//0022-0663.70.3.378

Surber, J. R., & Anderson, R. C. (1975). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology*, *67*(2), 170–173.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Van der Kleij, F. M., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263–272. doi:10.1016/j.compedu.2011.07.020

Van der Linden, E. (1993). Does feedback enhance computer-assisted language learning? *Computers & Education*, *21*(1–2), 61–65. doi:10.1016/0360-1315(93)90048-N

Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Comtemporary Educational Psychology*, *19*, 251–265.

White, L. (1991). Adverb placement in second language acquisition: Some effects of positive and negative evidence in the classroom. *Second Language Research*, *7*(2), 133–161. doi:10.1177/026765839100700205

Whyte, M. M., Karolick, D. M., Nielsen, M. C., Elder, G. D., & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research*, *12*(2), 195–203. doi:10.2190/M2AV-GEHE-CM9G-J9P7

Yanguas, Í. (2010). Oral computer-mediated interaction between L2 learners: It's about time! *Language Learning & Technology*, *14*(3), 72–93. Retrieved from http://www.llt.msu.edu/issues/october2010/yanguas.pdf