CrossMark

Studies in Second Language Acquisition, **44** (2022), 2–34
doi:10.1017/S0272263120000388

Research Article

THE EFFECTS OF IMMEDIATE AND DELAYED CORRECTIVE FEEDBACK ON L2 DEVELOPMENT

Mengxia Fu

Guangdong University of Foreign Studies and The University of Auckland

Shaofeng Li

Florida State University

Abstract

This article reports on a study exploring the differential effects of immediate and delayed corrective feedback (CF) on the acquisition of the English past tense. One hundred and forty-five seventh-grade EFL learners were assigned to four groups: Immediate CF, Delayed CF, Task Only, and Control. Each experimental group performed six focused communicative tasks, two each in three treatment sessions, eliciting the use of the English past tense. The Immediate CF group received feedback on their erroneous use of the target structure in the first session, the Delayed CF group received feedback in the final session, and the Task Only group performed the communicative tasks without receiving any feedback. The Control group only took the achievement tests. The effects of the feedback treatments were measured through an untimed grammaticality judgment test and an elicited imitation test. Mixed-effects analyses examining the influence of both fixed and random factors demonstrated that immediate CF was more facilitative of L2 development than delayed CF. The results suggest the importance of addressing linguistic errors before they are proceduralized in the interlanguage.

INTRODUCTION

Over the past two decades, corrective feedback (CF) has been at the forefront of second language research due to its theoretical and pedagogical significance. The popularity of CF is evidenced by the large amount of research examining various

The study was supported by the Chinese Scholarship Council and the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies. We would like to thank Rosemary Erlam and Yalu Wen for their assistance with various aspects of the study and Shuirong Deng and Zhuangwei Tang for their support with data collection and participant recruitment. We are solely responsible for any errors.
Correspondence concerning this article should be addressed to Shaofeng Li. E-mail: sli9@fsu.edu

© *The Author(s), 2020. Published by Cambridge University Press.*

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

issues surrounding CF including what types of CF teachers use in the classroom (Brown, 2016; Choi & Li, 2012); how learners respond to CF or whether CF is followed by uptake (Loewen, 2005; Lyster & Ranta, 1997); how learners perceive CF (Egi, 2007; Mackey et al., 2000; Philp, 2003); what types of CF are more effective than other CF types (Ammar & Spada, 2006; Ellis et al., 2006; Sheen, 2010); what beliefs and attitudes teachers and learners have toward CF (Li, 2017); whether CF fares differently under simple and complex task conditions (Baralt, 2013; Révész, 2009); and how CF or different types of CF interact with learner-internal variables such as language aptitude (Li, 2014; Yilmaz, 2013), working memory (Goo, 2012; Kim et al., 2015), and anxiety (Sheen, 2008). The large amount of research has been synthesized in a number of narrative reviews (Li & Vuono, 2019; Nassaji, 2016) and meta-analyses (Li, 2010; Lyster & Saito, 2010). In fact, Plonsky and Brown (2015) identified 18 meta-analyses relating to CF, testifying to its significance in SLA and the intensity of interest in this instructional device.

Despite the large amount of research, one area awaiting further empirical investigation is the timing of CF or the ideal time to provide feedback. The timing of CF can be operationalized in two ways. One is to distinguish whether CF is provided on errors learners make while performing a communicative task or after the task is completed, in which case the former can be called online feedback and the latter offline feedback (Li, 2020). This method of operationalization involves manipulating the context of feedback. Although SLA theories do not make overt claims about the context of CF, they all point to the superiority of online feedback to offline feedback (Li, 2020). This is primarily because online feedback is provided during communication and might therefore cater to communicative competence or implicit knowledge, whereas offline feedback constitutes isolated focus on form that likely facilitates explicit knowledge or knowledge that may not be easily accessible in spontaneous production (Krashen, 1981; Spada et al., 2014).[1] However, from a pedagogical perspective, offline or posttask feedback has an advantage because online feedback may disrupt the flow of communication and potentially has a negative impact on fluency (Harmer, 2007; Willis & Willis, 2007). Another way of operationalizing CF timing is to manipulate the schedule of feedback, that is, immediate feedback is provided at an early stage of an instructional cycle such as immediately after learners receive explicit instruction, and delayed feedback is offered after learners engage in some practice activities. However, in CF literature, the context and schedule of feedback are conflated, and immediate and delayed feedback are used as labels for the two types of feedback in both distinctions.[2] Here in this study, we would like to tease out the two distinctions by referring to feedback provided in different contexts as online and offline feedback and feedback based on different schedules as immediate and delayed feedback. In SLA research, there have been a few studies comparing the effects of online and offline feedback (Arroyo & Yilmaz, 2018; Li et al., 2016b; Quinn, 2014). However, there has been no research on the comparative effectiveness of immediate and delayed feedback. This study seeks to fill this significant gap by investigating whether immediate and delayed feedback has differential effects on the acquisition of the English past tense by young EFL learners.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

LITERATURE REVIEW

THEORIES OF FEEDBACK TIMING

The optimal timing of feedback has attracted theoretical debates in both psychology and SLA. In psychology, theories supporting immediate feedback are represented by behaviorism, which views feedback as a device that (a) corrects errors and (b) reinforces correct behaviors (Skinner, 1953). Behaviorism views learning as habit formation, and errors must be corrected immediately before they turn into bad habits. Also, good habits are formed by constant reinforcement of correct responses using repetition and feedback that confirms correct responses. Theories supporting the superiority of delayed feedback include Perseveration-Interference Theory and Spacing Theory. According to Perseveration-Interference Theory (Kulhavey & Anderson, 1972), delayed feedback is more effective than immediate feedback because if errors are corrected immediately, incorrect responses cause interference with correct responses, which hinders learning. In contrast, delaying feedback allows errors to dissipate or to be forgotten, and there will be no interference when only the correct answers are provided at a later time. Spacing Theory emphasizes the importance of learning from feedback to correct responses, claiming that feedback provided immediately after correct responses represents massed presentation while it constitutes spaced presentation if it is delayed (Smith & Kimball, 2010). Spaced instruction is more effective than massed instruction because of the lighter cognitive burden imposed on the learner.

It must be clarified that in psychological research, feedback is typically provided after both correct and incorrect answers, hence the dual function of feedback of reinforcing correct responses and eradicating errors. In essence, the research on feedback in psychology is derived from, or inspired by, behaviorism, which conceives learning as a process of stimulus response and habit formation. Accordingly, the learning tasks normally take the form of discrete item practice such as multiple-choice tests and paired-associate tasks (alternatively called "list learning") where learners memorize nonsense words and their meanings (Kulik & Kulik, 1988). In SLA research, however, the primary function of feedback is corrective rather than reinforcing, and feedback is typically provided in meaning-oriented communicative tasks. Therefore, the claims of psychological theories on the influence of the timing of feedback on learning outcomes are not entirely applicable to SLA. Nevertheless, psychological theories provide useful implications for the mechanism of feedback in SLA.

Feedback figures prominently in SLA theories. According to the currently popular Interaction Hypothesis (Long, 2015), the optimal time to address linguistic problems is through CF during negotiated interaction. This way of attending to linguistic forms is called focus on form, and it has a number of benefits. First, it occurs during meaning-oriented tasks and likely caters to the acquisition of implicit knowledge, the type of knowledge drawn upon in spontaneous L2 production. Second, it is reactive rather than preemptive, thus overcoming the limitation of traditional approaches where learners are imposed on forms that they are not ready for and increasing the likelihood of synchronizing form-focused instruction with the learner's internal timetable. Third, CF in the form of recasts, which reformulate erroneous utterances, is particularly useful because (a) it juxtaposes the correct and incorrect forms so learners may make a cognitive comparison between the two, and (b) it reformulates the utterance with the meaning

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

intact, thus freeing up the learner's cognitive resources for the linguistic target in question. Fourth, focus on form happens at a propitious moment when the learner is grappling with linguistic forms, and the need to resolve communicative problems provides an incentive for learning. Thus, based on the Interaction Hypothesis, feedback must be immediate, occur during communicative tasks, and be contiguous with errors. Providing feedback during the posttask stage separates feedback from errors and makes feedback decontextualized. Therefore, posttask feedback is not ideal. Essentially, Long's theory concerns online and offline feedback, that is, whether feedback should be provided during or after task performance, not whether feedback should be postponed until a later time. In our study, both feedback types are online feedback, and they are distinguished in terms of whether there is an interval between the initial instruction and feedback or whether learners are allowed to engage in some communicative practice before receiving feedback on their errors.

Another theory that feedback research draws on is Skill Acquisition Theory (Lyster, 2004), according to which feedback in the form of prompts is ideal because they withhold the correct form and encourage self-repairs. Self-repairs constitute practice of linguistic knowledge stored in long-term memory and facilitate the proceduralization and automatization of L2 knowledge. Two aspects of the theory are relevant to the current study. First, because this theory emphasizes the importance of feedback in activating and proceduralizing partially acquired knowledge, in the studies conducted in this theoretical framework such as Ammar and Spada (2006) and Sato (2013), learners typically receive explicit instruction on the target structure before engaging in communicative tasks. In this study, learners also received explicit instruction before communicative practice to maximize the effects of feedback. Second, with regard to the timing of feedback, which is the independent variable of this study, DeKeyser, a proponent of the theory, claims that CF "should not be delayed too much" (2007, p. 4) because delayed feedback may result in error fossilization. However, DeKeyser did not elaborate on the ideal length of the delay.

Notwithstanding the lack of a specific agenda for feedback delay, Skill Acquisition Theory suggests an advantage of immediate feedback based on the following grounds. First, immediate feedback embedded in communicative practice after explicit instruction makes it possible for the instruction and feedback to jointly contribute to the proceduralization of declarative knowledge. Immediate feedback occurs at a time when declarative knowledge can be easily triggered, reinforced, and applied and when the effects of feedback can be enhanced by the declarative knowledge that is still in an active state. Delaying feedback separates explicit instruction from feedback, making them work independently and mitigating the effects of both. Second, errors can be proceduralized or even automatized if they are not rectified at a stage where the interlanguage can be easily modified. This might be especially true of nonsalient structures such as the past tense or the third-person singular form, which are semantically redundant in many contexts.

However, there are reasons to favor delayed feedback rather than immediate feedback. First, if we accept the argument that interlanguage is an evolving system where errors are inevitable and developmental (Ellis, 2015; Ortega, 2009), then errors should not be corrected immediately. As learners proceed to more advanced stages, some errors will possibly evaporate as a result of more input and practice, and feedback, if necessary, should be delayed and only target remaining errors that cannot be dealt with otherwise.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Second, drawing on Spacing Theory, following instruction with immediate feedback constitutes massed instruction whereas delaying feedback or separating feedback from instruction can be considered spaced instruction (see Quinn & Nakata, 2017 for a more detailed discussion). Massed instruction imposes more cognitive burden on learners than spaced instruction and is therefore less ideal (Nakata & Suzuki, 2019; Rogers, 2017). Third, delayed feedback may be more motivating than immediate feedback because in the former case, feedback is provided after learners struggle with the linguistic target during practice activities. This argument is similar to the Preparatory Attention theory, which claimed that learners will pay more attention to further instruction after they encounter difficulty in the tasks they perform (see Quinn, 2014 for a more detailed discussion). Fourth, delayed feedback can be more easily noticed, processed, and internalized than immediate feedback because learners will possess more knowledge about the linguistic structure and will therefore be developmentally more ready to take advantage of feedback (Mackey & Philp, 1998).

To conclude this section, we would like to cast the two distinctions about feedback timing in the two major SLA theories we discussed: Interaction Hypothesis and Skill Acquisition Theory. The Interaction Hypothesis makes clear claims about online and offline feedback, advocating the superiority of online feedback to offline feedback based on the reasons discussed in the preceding text. Skill Acquisition Theory provides a convenient framework for the conceptualization and operationalization of immediate and delayed feedback. Within this framework, feedback can be provided during tasks performed immediately after explicit instruction or postponed until a later time after learners complete some communicative practice. Skill Acquisition Theory also concerns "fairly early stages of learning" (DeKeyser, 2015, p. 101), where it is easier to minimize the influence of previous knowledge on learning outcomes and tease out the effects of feedback. Grounded in Skill Acquisition Theory, our study investigates the differential effects of immediate and delayed feedback on young learners' L2 development.

RESEARCH ON FEEDBACK TIMING

Psychological Feedback

In psychological research, the most comprehensive synthesis of the empirical studies on feedback timing is the meta-analysis conducted by Kulik and Kulik (1988). Based on 53 studies, the meta-analysis identified three types of research: applied studies, studies on the acquisition of test content, and studies on list-learning. Applied studies were carried out in the classroom, and in such studies students were typically given quizzes on studied materials and received feedback either immediately after responding to test items or one or two days after the quiz. Applied studies showed superior effects for immediate feedback compared with delayed feedback. In the studies on the acquisition of test content, learners responded to multiple-choice items, and answers to test items were the target of learning. These studies were conducted in laboratory settings and showed larger effects for delayed feedback than immediate feedback. In studies on list-learning (or paired associates), learners were required to study the associations between words and their meanings. These studies showed mixed effects, which, according to the meta-analysts, seem to be caused by some methodological features. For example, when immediate and delayed feedback

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

were fused in one session (e.g., distinguished in terms of whether feedback was given after each item or after several items), immediate feedback was more effective. However, when immediate and delayed feedback were given in separate sessions, delayed feedback worked better. Overall the studies seem to show that immediate feedback worked better in the classroom while delayed feedback fared better in laboratory settings. What caused the disparity between the findings of classroom and laboratory studies? One explanation derived from subsequent research is that in classroom settings, delayed feedback is not processed properly (i.e., learners do not take it seriously) when quiz answers are reviewed. When feedback is equally attended to in immediate and delayed feedback conditions, delayed feedback normally works better (Butler et al., 2007; Mullet et al., 2014), which is explainable by the Interference-Perseveration Hypothesis (i.e., errors are forgotten during the delay and therefore cause less interference) and the Spacing Theory (delayed feedback constitutes spaced instruction).

SLA Research

Online versus Offline Research. In SLA, several studies have examined the effectiveness of online and offline feedback, namely whether feedback is provided while learners are preforming a communicative task or after it is completed. The two feedback types, however, were referred to as immediate and delayed feedback in the studies. To maintain consistency with our conceptualization of feedback timing, we will refer to the two feedback types as online and offline feedback in our discussion of the representative studies on the effectiveness of the two types of feedback. Arroyo and Yilmaz's study (2018) investigated the effects of online and offline feedback on L2 Spanish learners' acquisition of the Spanish noun-adjective gender agreement at a U.S. university. The study involved 45 learners divided into three groups who engaged in dyadic interaction with a native speaker using Skype text messaging in synchronous computerized communication. The online group received recasts on their wrong use of the target structure during text chat. The offline group did not receive any feedback during the chat, but upon completion of the task, they received a list of errors they made and the correct versions of the sentences and were asked to study the corrections for 5 minutes. The control group only performed the communicative tasks without receiving any feedback. The results showed that online feedback was more effective than offline feedback on an oral production test, but there was no difference on a grammaticality judgment test.

Li et al.'s study (2016b) was conducted in a classroom setting with 120 eighth-grade EFL learners at a Chinese middle school. The learners were divided into four groups: online feedback, offline feedback, task only, and control. The experimental groups attended a 2-hour treatment where they performed two dictogloss tasks following the same steps. During each task, the teacher presented a narrative orally, then the students worked in pairs to research the narrative, and finally they were called on to retell the narrative to the rest of the class. During the reporting stage, the online feedback group received corrective recasts (teacher repetition of error + reformulation in the absence of self-correction) on their wrong use of the English passive voice; the offline feedback group received corrections on their errors after the learners completed the two tasks; the task only group did not receive any feedback; and the control group only took the pretests and posttests. The study showed an advantage for online feedback over offline feedback

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

on a grammaticality judgment test, but the difference was not evident on an elicited imitation test (EIT).

Quinn (2014) reported a study conducted in face-to-face dyadic interaction involving 90 adult ESL learners from three private language schools in Canada. The learners were assigned to three groups: online feedback, offline feedback, and control (task only). All three groups received a 10-minute grammar lesson about the target structure—the English passive voice—before completing three oral tasks (information gap, story-retell, and role play). The online feedback group received error correction during their task performance, the offline feedback group received feedback after completing each task, and the control group only performed the tasks without receiving any feedback. The feedback consists of a prompt followed by a recast, which was provided regardless of whether the prompted response was correct. The learner was asked to repeat the recast provided by the interlocutor. The study failed to find any differences among the treatment groups on any of the tests used to measure treatment effects.

What do the three studies tell us about online and offline feedback, and in what way are they related and different from each other? Methodologically, the three studies spanned laboratory and classroom contexts; involved both face-to-face and computer-mediated interaction; and examined both foreign language and second language settings, and investigated both adult and child learners. Therefore, they are representative of the different types of research typically seen in this substantive domain, and the empirical evidence, albeit limited, is of theoretical and pedagogical significance. Both Arroyo and Yilmaz (2018) and Li et al. (2016b) showed that online feedback was more effective than offline feedback, and the researchers provided the following explanations for the superior effects of online feedback. First, in online feedback, the short interval between feedback and errors makes it possible for learners to make a direct comparison between the correct and wrong forms. In offline feedback (after a task is completed), however, feedback is provided separately from errors made during task performance, in which case many learners were unable to identify their errors or recall the contexts in which the errors were made. Second, the knowledge learned through online feedback is applied immediately in subsequent, ongoing speech production while the knowledge learned through offline feedback is not applied in any production activity. In other words, there are no practice opportunities to proceduralize the learned knowledge. Third, online feedback is motivating on the grounds that (a) it is provided at a propitious moment when the learner is struggling for the right form to express meaning, and (b) the learner needs to display their learned knowledge in further language use during the ongoing task. Offline feedback, however, is not as motivating because there is a lack of participatory demand. The learner does not need to show the correct use in further task production, and in fact the learner does not even need to attend to feedback.

Although Quinn's study did not detect any significant differences between the two feedback types, the findings provide valuable implications and lend support to some of the interpretations provided by Arroyo and Yilmaz (2018) and Li et al. (2016b) for the superior effects of online feedback. One explanation for the lesser effects of offline feedback in the two studies is that there were no practice opportunities after learners received feedback in the posttask stage. This explanation is partially confirmed by Quinn's study where learners in the offline feedback group received feedback during the intervals between the three tasks, which means that the learners had opportunities to

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

apply or proceduralize the declarative knowledge learned from the feedback prior to task performance. Performing a task after receiving feedback may have also incentivized the learners to attend to and process feedback. However, Quinn's interpretations of his results are supported by the other two studies. For example, Quinn suspected that the lack of between-group difference was probably due to the mini grammar lesson learners received prior to the task-based treatments that may have neutralized the effects of feedback. This speculation receives indirect support from the other two studies, where learners did not receive any instruction. The lack of treatment effects in Quinn's study, which were obtained in the other two studies, is also subject to methodological differences between the studies in other dimensions. For example, the learners in Quinn's study were adult ESL learners who had a substantial amount of previous knowledge about the target structure. These learners also had heterogeneous linguistic and cultural backgrounds. In contrast, the other two studies investigated learners with limited previous knowledge and homogeneous backgrounds.

Immediate versus Delayed Feedback. The previously mentioned studies examined feedback timing operationalized as online and offline feedback, and no research has investigated the effects of immediate and delayed feedback, which refers to whether learners receive feedback immediately after explicit instruction or after completing some communicative practice following the initial instruction. We argued in previous sections that feedback timing operationalized as online and offline feedback has theoretical roots in the Interaction Hypothesis, and that feedback timing conceptualized as immediate and delayed feedback can find a theoretical basis in Skill Acquisition Theory. According to Skill Acquisition Theory, an instructional cycle starts with explicit instruction, followed by skill-specific practice activities. Learners in an immediate feedback condition receive feedback during tasks performed immediately after receiving explicit instruction, and learners in a delayed feedback condition do not receive feedback until a later stage after completing some communicative practice. To date, studies couched in Skill Acquisition Theory have invariably examined immediate feedback. These studies are typically carried out in classroom settings, and they compare the effects of prompts and recasts embedded in communicative tasks preceded by explicit instruction (Ammar & Spada, 2006; Lyster, 2004; Sato, 2013). The purpose of the instruction is to ensure that (a) learners have equivalent existing knowledge about the target structure and (b) learners have a certain amount of knowledge to draw on when processing feedback. Overall the studies showed larger effects for prompts than recasts, especially in classroom settings where the corrective force of recasts is not as salient as prompts.

Because explicit instruction is an important component of the instructional treatment in this approach, a few studies have examined the influence of explicit instruction on feedback effects and task performance. It has been found that pretask explicit instruction reinforces the effects of feedback in that feedback preceded by instruction showed larger effects than feedback alone (Li et al., 2016a; Saito, 2013a, 2013b; Sato & Loewen, 2018). Other, "weaker" forms of pretask form-focused instruction include guided planning where learners' attention is directed toward a linguistic target during the planning stage of an oral task, and pretask modeling where learners are presented with a sample performance before completing a task. Mochizuki and Ortega (2008) reported that guided planning in the form of a handout explaining the formation of the English relative clause

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

led to more accurate production of the linguistic target without affecting fluency and complexity. Kim (2013) found that pretask modeling led to more incidence of language-related episodes, and van de Guchte et al. (2015) showed that meaning-oriented and form-oriented pretask modeling led to a trade-off between meaning and form: The former enhanced overall complexity while the latter prompted the learners to use the target structure more frequently. Overall, the limited research shows that pretask explicit instruction reinforces the effects of feedback and leads to more frequent use of the linguistic target.

As can be seen, to date no SLA research has explored the timing effects of CF through the manipulation of feedback scheduling, that is, whether feedback is offered immediately after learners receive explicit instruction or with a delay after a certain amount of practice. This study seeks to fill the gap, and it was guided by the following research question:

Do immediate and delayed feedback in the form of corrective recasts have differential effects on the acquisition of the English past tense?

METHODS

PARTICIPANTS

The participants were 145 seventh-grade EFL students at a private school in Eastern China. They were aged between 12 and 14 (M = 12.94) with a relatively balanced gender ratio (n = 70 for females and n = 75 for males). All the students were beginners having learned English for 4 to 11 years with an average length of 6.5 years before the commencement of the study. According to the curriculum of the school, the participants received about 4.5 hours of English instruction every week, delivered in six lessons, with each lesson lasting 45 minutes. Five of the six lessons incorporated all skills, and one focused on speaking. Although the learners' English teachers mainly adopted traditional approaches with a heavy dose of explicit instruction, the learners had opportunities to engage in spontaneous communication during their oral English classes taught by a native English speaker.

The participants were recruited from five classes and assigned to four groups: Immediate CF (n = 39), Delayed CF (n = 41), Task Only (n = 34), and Control (n = 31). One-way ANOVA did not detect significant differences between the four groups in terms of their overall L2 achievements represented by their mid-term examination scores, F (3, 139) = 1.09, p = .36, and their previous knowledge about the target structure (the past tense) reflected in their pretest scores on an untimed grammaticality judgment test (UGJT; a test of explicit knowledge), F (3, 137) = 2.29, p = .08, and an elicited imitation test (EIT; a test of implicit knowledge), F (3, 130) = 1.46, p = .23.

TARGET STRUCTURE

The target structure of this study was the English simple past tense, which was selected based on two reasons. First, Chinese EFL learners have difficulty using the past tense accurately due to the absence of tense morphology in their L1 (Yang & Lyster, 2010). Second, this linguistic structure fits the objective of this study. The study aims to examine whether immediate and delayed feedback has differential effects on the learning of a linguistic structure learners have limited knowledge about. The past tense meets the

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

TABLE 1. The target verbs seeded in the three practice sessions.

	Regular verbs			
	id	t	d	Irregular verbs
Session 1 (Tasks 1 & 2)	start; plant; count; harvest	clap; watch; walk; notice	move; listen; spoon; change	get; say; tell; break; go; begin; take; throw; find; become; eat; make
Session 2 (Tasks 3 & 4)	defeat; invite; shout; attempt	snatch; ask; drop; pass	yell; play; grab; pull	get; say; tell; break; go; begin; take; throw; find; become; eat; make
Session 3 (Tasks 5 & 6)	hand; decide; collect; pat	rip; laugh; stop; flash	follow; happen; return; answer	get; say; tell; break; go; begin; take; throw; find; become; eat; make

requirements in this regard because it had not been taught to the learners prior to this study,[3] and the learners had limited previous knowledge about this structure, which is reflected by their low pretest scores on the UGJT (20%) and the EIT (18%).

The target verbs of each practice session are displayed in Table 1. Forty-eight verbs, including 36 regular and 12 irregular verbs, were embedded in the instructional treatment, which consisted of six tasks evenly distributed among three practice sessions. To ensure that learners had equal opportunities to use regular and irregular verbs, each task included six regular and six irregular verbs. Thus each practice session involved 12 regular verbs and 12 irregular verbs. Whereas the 12 irregular verbs were recycled among the three sessions because irregular verb forms were expected to be acquired as items, the 12 regular verbs were different across treatment sessions because the learners were expected to acquire the rule of the regular past tense. To minimize the influence of the phonological features of past-tense morphemes on the acquisition of this linguistic structure, the three types of regular past-tense morphemes (/t/, /d/, /id/) were distributed equally among the three sessions, that is, each of the three morphemes appeared in four of the 12 target regular verbs. To control for the influence of word frequency on the acquisition of the English past tense, the frequency of the target verbs in the CF treatment session (i.e., tasks 1 & 2) was measured using the Web VP Classic (available at https://www.lextutor.ca/vp/eng/). It was found that 23 out of the 24 target verbs were high-frequency words. The only low-frequency word (clap) had appeared in English songs and classroom instructions with which the learners were familiar (e.g., "Clap your hands").

INSTRUCTIONAL TREATMENT

The instructional cycle of this study consisted of three practice sessions and lasted approximately 4.67 hours in total. The duration of Session 1 was 1.67 hours including a 10-minute grammar lesson and two 45-minute tasks; it was slightly longer than Sessions 2 and 3 (both 1.5 hours) where the learners only performed communicative tasks without receiving explicit instruction. The treatment included three components: (a) a mini

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

grammar lesson; (b) six communicative tasks; and (c) CF in the form of corrective recasts provided during task performance under immediate and delayed feedback conditions. The three components are explained in the following sections.

Mini Grammar Lesson

At the beginning of Session 1, the participants received a 10-minute grammar lesson on the target structure. The brief grammar lesson aimed to encourage learners to generate self-corrections, which would have been impossible without a certain amount of knowledge of the target structure[4] (Ammar & Spada, 2006; Lyster, 2004). The grammar lesson followed the three steps of VanPatten's (2004) Processing Instruction. First, the researcher verbally introduced the metalinguistic knowledge about the English past tense. Second, the researcher asked the learners to pay attention to the differences between English and Chinese in terms of verb inflections. This step sought to encourage the participants to abandon their default processing strategy—focusing on content words instead of verb inflections—given that verb inflections are absent in the participants' L1 (Ellis & Shintani, 2014). Finally, the learners were asked to complete a structured input activity, which sought to help them establish the form-meaning mapping of this linguistic structure.

Tasks

After the grammar lesson, the learners performed six communicative tasks lasting 4.5 hours, two each in three treatment sessions. Three task types were used: dictogloss, spot-the-difference, and sequencing. To eliminate the influence of task type on treatment effects, the two CF groups performed the same tasks during the CF treatments. The tasks followed similar procedures to those utilized in Li et al. (2018):

(1) Lead-in. Before each task started, the learners sang a song or watched a short video related to the topic of the task (e.g., friendship). Then the researcher asked several brainstorming questions about the topic (e.g., who is your best friend?) to arouse learners' interest in the task and activate their schemata.
(2) Providing input. At the beginning of each task, students got familiar with some new vocabulary by repeating some keywords after the teacher. They then engaged in input-based activities providing opportunities for exposure to the linguistic target. During the dictogloss task, the researcher read a gapped narrative aided with picture illustrations on 10 PowerPoint slides (nine picture slides and one blank slide at the end) three times. The teacher first read the narrative at a normal speed to allow students to establish an initial impression. During the second reading, the teacher presented the narrative at a slower speed, pausing for about 5 seconds between slides to allow the learners sufficient time to process the input. The teacher also placed some prosodic emphasis on the past-tense forms built in the narrative. Finally, the teacher presented the narrative again to solidify the learners' understanding. During the spot-the-difference and sequencing tasks, after the initial vocabulary learning, the learners were grouped into pairs, and each pair received two envelopes, each containing a set of pictures and a text describing the pictures. The learners were then asked to study their materials for 5 minutes, and to maintain the

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

information gap within pairs, they were not allowed to look at each other's materials during group work.

(3) Group work. For each dictogloss task, after listening to a story three times, the learners were given 10 jumbled pictures that had appeared on the PowerPoint slides. They then worked in groups of four for 10 minutes. During group work, they put the pictures into the order presented by the researcher, practiced retelling the story, and made up a plot for the blank picture, which served as the story ending. The target verbs were listed below each picture, and the learners were informed that the listed words were essential for their group presentations during the reporting stage (described in the following text). For the spot-the-difference task and the sequencing task, the materials were collected by the teacher after the learners studied them for 5 minutes. The learners were then given 10 minutes to identify 12 differences between their pictures in the case of a spot-the-difference task,[5] or reach an agreement regarding the sequence of the pictures in the case of a sequencing task.

(4) Public reporting. For each dictogloss task, each group (four students) was invited to present a complete story with an invented ending, with each student describing two to three pictures before passing the speaker's role to the next group member. The presenters showed the audience their pictures through a projector during their speeches. For the spot-the-difference task and the sequencing task, each pair presented the identified differences and the sequenced story to the rest of the class. The two feedback groups received individualized feedback on their wrong use of the English past tense during the reporting stages of the dictogloss tasks in the first (immediate feedback) or final (delayed feedback) session. They did not receive feedback during the performance of the spot-the-difference task and the sequencing task.

(5) Wrap-up. At the end of each dictogloss task, the students voted for the most interesting story ending. The sequencing task ended by asking students to select the most reasonable story sequence. For the spot-the-difference task, the group that had successfully identified all the differences was given a prize after the task was completed.

The topics of the four dictogloss tasks were related to the students' daily life at school or at home. The written texts and pictures of the four dictogloss tasks were adapted from four stories published in *School Journal*—an instructional series published by the Ministry of Education of New Zealand for Year 4 to Year 8 students. The pictures used in the sequencing task were the screenshots of a popular Mr. Bean episode (Mr. Bean's Holiday) and the written texts of the task were composed by the researchers. The researchers also created the written texts and pictures of the spot-the-difference task, which was about a bad day for two characters selected from the learners' textbook. Before finalizing all the task materials, the researchers consulted the participants' teachers to confirm the appropriacy of the written texts and the task pictures.

Corrective Feedback

The Immediate and Delayed CF groups received CF treatments in the first and final practice session, respectively, during group presentations. The CF treatments were operationalized as a type of hybrid feedback consisting of a prompt followed by a recast. This type of feedback was first introduced in Doughty and Varela (1998) and subsequently utilized in several studies (e.g., Erlam & Loewen, 2010; Li et al., 2016b; Quinn, 2014; Zhao, 2015). Li et al. (2016b) pointed out that corrective recasts are a "potentially ideal CF strategy" (p. 280) because they integrate output-prompting and input-providing feedback, taking advantage of the strengths and overcoming the limitations of both

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

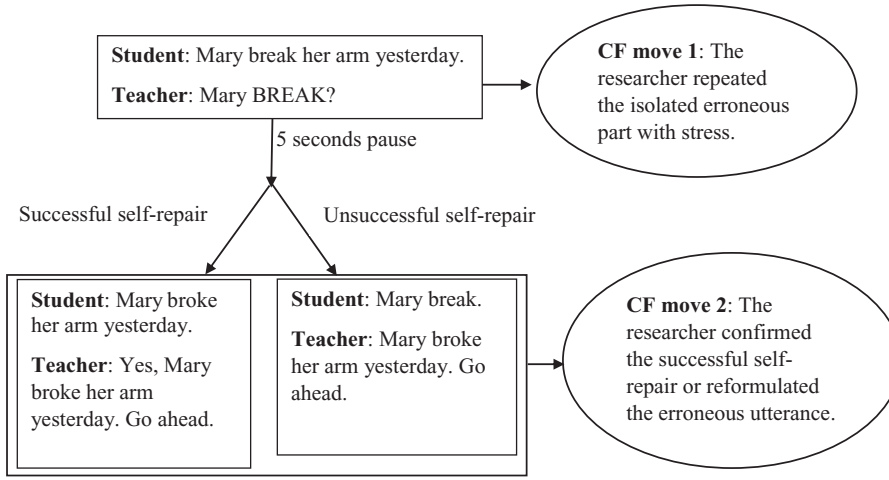


FIGURE 1. The implementation of corrective recasts in this study.

feedback types (Goo & Mackey, 2013; Lyster & Ranta, 2013). The feedback was provided adhering to the following steps. First, when a learner produced an erroneous past form, the teacher repeated the wrong form with a prosodic emphasis and a rising tone to draw the presenter's attention to the error and to elicit a self-correction. Then, the teacher paused for 5 seconds for a self-correction. If the self-correction was successful, the researcher confirmed it by repeating the correction[6] and signaled the presenter to continue his/her speech. If the presenter was unable to self-correct, the researcher reformulated the erroneous output by replacing the wrong form with the correct form. It is worth noting that, although the CF treatments were directed at individual presenters, they were considered to be beneficial for both presenters and listeners because previous research (e.g., Kim & Han, 2007) has shown that exposure to CF treatments (e.g., listening to corrections) also promoted L2 acquisition. The steps of the feedback provided in this study are illustrated in Figure 1. Information about the number of errors, corrections, and presenters in the two CF groups is provided in Appendix A.

MEASUREMENT

Treatment effects were measured by means of an UGJT and an EIT,[7] intended to measure learners' explicit and implicit knowledge of the linguistic target, respectively (Ellis, 2005; Li et al., 2019). The tests targeted the same 24 verbs, among which half were regular verbs and half irregular verbs. The verbs were evenly selected from the six treatment tasks. Five different versions of each of the two test types were created, each administered at a different time point during the study: pretest, posttest 1, posttest 2, posttest 3, and posttest 4 (see the "Procedure" section for the timing of the tests). Different versions of the same test were created by changing the sequence in which the test items were presented.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Untimed Grammaticality Judgment Test

During the test, the learners received a booklet with 34 sentences. They were asked to judge the grammaticality of each sentence and correct the ungrammatical sentences. Among the 34 test items, 24 were target items and 10 were distractors, and all the target items were ungrammatical. The full score for this test was 24. One point was given when a learner provided the correct form for a target item. Partial credits were given to answers with spelling errors (e.g., "pland" for "planted"). No credit was given if learners failed to correct the mistake regardless of the accuracy of their judgments. This is because in the absence of the correct form, it is impossible to know whether the learner's judgment was based on their knowledge about the linguistic target or random guessing, which has a 50% chance of success given the dichotomous nature of the answer (correct or incorrect). About 20% of the testing papers (one test for each group) were marked by a research assistant and the interrater reliability of the UGJTs was .98.

Elicited Imitation Test

This test aimed to elicit learners' use of the target structure in oral production. The test was delivered using DMDX, a free software program used in psychological studies to present visual stimuli and record learners' reaction time and oral output. For each test time, the learners were required to listen to the audio recording of a sentence, decide if the statement was true of them, and repeat the sentence in correct English. To impose some time pressure on the participants (Suzuki & DeKeyser, 2015), they were instructed to reproduce each sentence as quickly as possible. Each version of the EIT included 24 sentences, with one target verb per sentence. Half the target stimuli (n = 12) were grammatical and half ungrammatical (Erlam, 2006). The full score for each EIT was 24. Each correctly elicited past form received one point. Partial credits were allocated to answers containing pronunciation errors, such as "god" for "got" and "becane" for "became." One-fifth of the audio recordings were transcribed and marked by a research assistant, and the interrater reliability was .93.

PROCEDURE

This quasiexperimental study adopted a pretest–posttest design. Prior to the experiment, the four groups of learners (three experimental groups and a control group) completed a battery of pretests, including a vocabulary test,[8] an UGJT, and an EIT. One week after the pretests, in the first treatment session (Session 1), the three experimental groups attended a mini-lesson on the target structure and performed two tasks. The Immediate CF group received CF during their task performance. Immediately after the completion of Session 1, all four groups completed the first posttest. In Session 2, which took place one week after Session 1, the three experimental groups performed two different tasks where the researcher refrained from providing any CF. One week after Session 2, in Session 3, all three experimental groups performed the final two tasks, during which the Delayed CF group received CF. Immediately after the completion of Session 3, all four groups took the second posttest. The first and second posttests were administered after the first and final treatment sessions respectively to overcome a potential bias resulting from the timing of

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

feedback treatments. In other words, the effects of immediate feedback are likely larger on the first posttest because the test was administered at the end of the session in which the feedback was provided, and for the same reason, the effects of delayed feedback are likely larger on the second posttest. Two weeks after posttest 2, all four groups were given the third posttest. Two weeks after the third posttest, all four groups completed the last posttest. To prevent the influence of the UGJT on the EIT, the EIT was always administered before the UGJT. The Control group only took the achievement tests to tease apart testing effects, that is, learners' scores may improve as a result of repeating the same test.

ANALYSIS

To investigate the differential effects of immediate and delayed feedback, the data were analyzed through mixed-effects models. The dependent variable of each mixed-effects model was the four groups' posttest scores on the UGJTs or EITs.[9] The independent variables of each model included random and fixed-effects parameters. "Subject" was treated as a random-effects parameter because the four posttest scores of the same participant were clustered. A random intercept was added to "subject" to control for the differences among the four posttesting points resulting from the random variabilities across the participants. Fixed-effects parameters were the independent variables manipulated in the study, including group, time (the timing of the tests), pretest scores, and the interaction between group and time.[10] Pretest scores were included as a covariate to control for their influence on UGJT and EIT posttest scores.

Two mixed-effects models were built, one for the UGJT scores and one for the EIT scores. Prior to data analysis, outliers were identified and removed, and the assumptions of mixed-effects models were checked, including linearity, absence of collinearity, homoscedasticity, and normality. Then a mixed-effects analysis was carried out using the function lmer of R package lme 4 (Bates, 2010). The models were fit using the default model fit method—the restricted maximum likelihood technique. The overall effects of the fixed-effects variables were examined using the anova function of R package lmerTest, followed by post-hoc comparisons using the emmeans function of R package emmeans to investigate group differences at each posttesting point. The influence of the learners' pretest scores on their posttest performance was controlled in pairwise comparisons. To minimize Type I errors, Bonferroni corrections were used to adjust the p values of the pairwise comparisons. The magnitudes of the between-group differences were measured by means of calculating the Cohen's d value for each comparison. Following Plonsky and Oswald's (2014) criteria for assessing the magnitude of effect sizes for between-group comparisons, .4, .7, and 1.0 were interpreted as small, medium, and large effects, respectively.

RESULTS

DESCRIPTIVE STATISTICS

Table 2 presents the means and standard deviations of the UGJT scores (the full mark was 24 for each test) at the five testing points, which were the pretest (before the first practice session), posttest 1 (after the first practice session), posttest 2 (after the final practice

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

TABLE 2. Means and SDs of the UGJT scores.

			Immediate CF	Delayed CF	Task only	Control
Pretest	Mean		3.05	2.98	4.36	1.80
	95% CI[a]	Upper	1.79	1.74	2.85	.48
		Lower	4.31	4.21	5.88	3.12
	SD		3.78	3.92	4.27	3.54
	n		37	41	33	30
Posttest 1	Mean		9.51	8.13	9.26	4.21
	95% CI	Upper	7.84	6.62	6.99	2.11
		Lower	11.19	9.63	11.53	6.30
	SD		5.02	4.70	6.19	5.50
	n		37	40	31	29
Posttest 2	Mean		11.14	9.01	11.05	4.63
	95% CI	Upper	9.17	7.33	8.69	2.59
		Lower	13.11	10.70	13.40	6.68
	SD		5.83	5.12	6.53	5.48
	n		36	38	32	30
Posttest 3	Mean		11.49	9.01	10.80	4.85
	95% CI	Upper	9.43	7.28	8.31	2.71
		Lower	13.55	10.75	13.29	6.99
	SD		6.27	5.43	7.02	5.74
	n		38	40	33	30
Posttest 4	Mean		11.38	8.20	10.35	4.87
	95% CI	Upper	9.37	6.45	7.50	2.77
		Lower	13.39	9.94	13.21	6.97
	SD		6.12	5.53	7.21	5.62
	n		38	41	27	30

[a]95% confidence interval for mean.

session), posttest 3 (2 weeks after the final practice session), and posttest 4 (4 weeks after the final practice session). This test aimed to measure learners' explicit knowledge about the linguistic target. The mean pretest scores ranged from 1.80 to 4.76 out of 24, suggesting that the learners had limited explicit knowledge about the past tense. Overall, all four groups improved from the pretest to the posttests, but the experimental groups showed more improvements than the Control group. As shown in Figure 2, among the three treatment groups, the group that received immediate CF appeared to show a greater improvement than the group that only performed communicative tasks, and the group that received delayed feedback improved the least. Further analyses with the test scores of regular and irregular verbs found that Immediate CF was superior to the other two experimental groups regardless of whether the treatment verbs were regular or irregular (see Appendix C and D). The largest increase happened between the pretest and posttest 1, followed by an increase between posttest 1 and posttest 2, from which point Immediate CF continued to improve (until posttest 3) while Delayed CF and Task Only started to decline. Reliability analyses showed the internal reliability of the UGJTs represented by Cronbach's alpha, which was .92, .95, .94, .95, .94 for the pr-test and four posttests, respectively.

The means and SDs of the EIT scores (the full mark was 24) are displayed in Table 3, and the developmental trajectory can be seen in Figure 3. The EIT was designed to test the

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

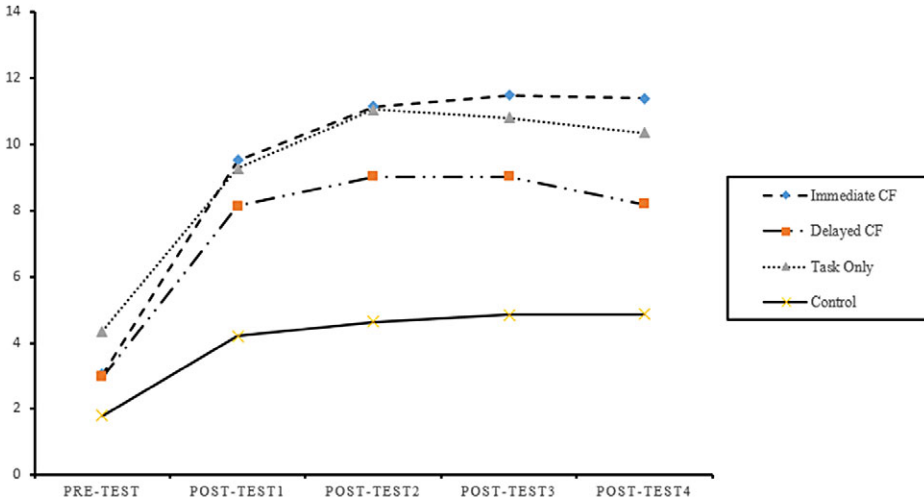


FIGURE 2. Group means of the UGJT scores at the five testing points.

TABLE 3. Means and SDs of the EIT scores.

			Immediate CF	Delayed CF	Task only	Control
Pretest	Mean		2.89	2.90	4.00	3.26
	95% CI[a]	Upper	2.09	1.99	3.10	2.36
		Lower	3.29	3.81	4.90	4.16
	SD		2.33	2.66	2.55	2.46
	n		35	35	33	31
Posttest 1	Mean		6.86	5.53	6.71	4.67
	95% CI	Upper	5.67	4.26	5.46	3.55
		Lower	8.06	6.81	7.95	5.79
	SD		3.38	3.47	3.27	3.00
	n		33	31	29	30
Posttest 2	Mean		8.92	7.67	8.13	4.37
	95% CI	Upper	7.77	6.34	6.83	3.19
		Lower	10.06	8.99	9.43	5.54
	SD		3.06	3.74	3.55	3.15
	n		30	33	31	30
Posttest 3	Mean		9.29	7.18	8.87	4.19
	95% CI	Upper	8.33	5.82	7.71	3.00
		Lower	10.25	8.54	10.04	5.39
	SD		2.61	3.83	3.18	3.25
	n		31	33	31	31
Posttest 4	Mean		8.18	6.69	7.83	5.32
	95% CI	Upper	6.93	5.31	6.63	3.93
		Lower	9.43	8.06	9.04	6.70
	SD		3.40	3.82	3.05	3.71
	n		31	32	27	30

[a]95% confidence interval for mean.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

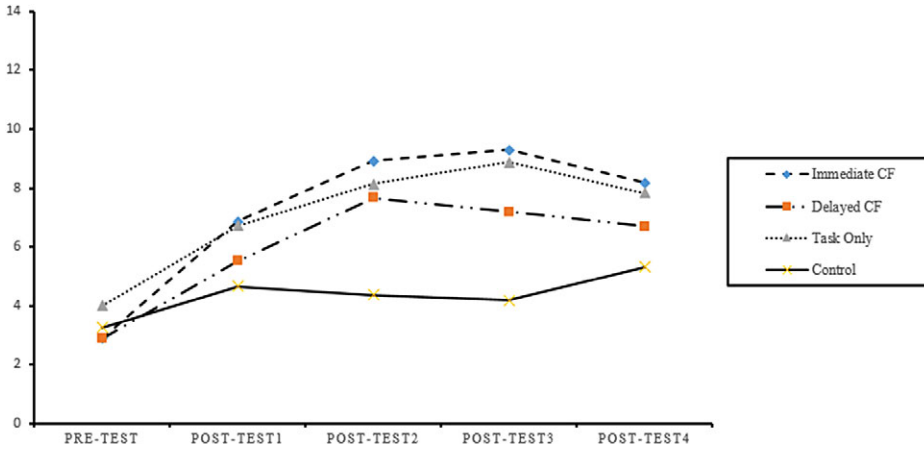


FIGURE 3. Group means of the EIT scores at the five testing points.

learners' implicit knowledge. Similar to the UGJT, the learners' mean EIT scores were quite low (< 4.5), demonstrating their limited previous knowledge about the target structure. Overall, the EIT posttest scores were lower than the UGJT scores, suggesting that implicit knowledge grew at a slower rate than explicit knowledge. Otherwise, the EIT scores showed similar patterns with the UGJT scores. First, among the three treatment groups, the immediate feedback group seemed to have improved the most, followed by the task-only group and the delayed feedback group, regardless of the features of the target structure (see Appendix C and D for the descriptive statistics of regular and irregular verbs). Second, the greatest increase was between the pretest and posttest 1, followed by another leap to posttest 2, and the scores remained relatively flat from posttest 2 onward. The internal reliability for the EITs, indexed by Cronbach's alpha, was .66 for the pretest and .62, .68, .76, .79 for the four posttests, respectively.

INFERENTIAL STATISTICS

Assumption Testing

The assumptions of mixed-effects analysis, including the absence of collinearity, linearity, homoscedasticity, and normality, were checked for the UGJT and EIT scores. For both datasets, multicollinearity among the three explanatory variables was checked using the variance influence factor (VIF) tests. The VIFs for all the explanatory variables were below 3, indicating that no collinearity existed. Linearity and homoscedasticity were examined through scatterplots, which did not show nonlinear or curvy patterns, suggesting these two assumptions were met. Normality was checked through Shapiro–Wilk tests, which showed that the learners' EIT posttest 1 scores were normally distributed but the remaining three EIT posttest scores and the four UGJT posttest scores were not. The EIT and UGJT data, however, were considered to be suitable for parametric tests given that (a) the results of Shapiro–Wilk tests can be significant in large samples ($N > 100$), such as in this study (Field, 2013) and (b) linear models are robust against the violation of normality (Winter, 2013).

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

UGJT Results

The omnibus test for the fixed-effects variables, including pretest scores, group, time, and the interaction between group and time, was conducted with the anova function of R package lmerTest. The results showed that the effects of pretest scores, $F(1, 135.80) = 63.36$, $p < .01$, group, $F(3, 135.89) = 7.19$, $p < .01$, and time, $F(3, 394.83) = 9.20$, $p < .01$, were all significant. The interaction effect between group and time, however, was not significant, $F(9, 394.79) = 1.32$, $p = .22$. Post-hoc analyses using the Bonferroni corrections were conducted to investigate group differences on the four UGJT posttests after excluding the influence of pretest scores. Table 4 displays the p values, effect sizes and mean effect sizes of each group comparison. The results showed that the Immediate CF group significantly outperformed the Control group on all four posttests with large effect sizes: $p = .02$, $d = 1.09$ for posttest 1; $p < .01$, $d = 1.42$ for posttest 2; $p < .01$, $d = 1.45$ for posttest 3; and $p < .01$, $d = 1.41$ for posttest 4. The differences between the Delayed CF group and the Control group were statistically significant on posttest 2, $p = .01$, $d = .84$, and posttest 3, $p = .04$, $d = .78$, both with a medium effect size. The Task Only group significantly outperformed the Control group on all three posttests, with small to medium effect sizes: $p < .01$, $d = .97$ for posttest 2; $p = .01$, $d = .85$ for posttest 3; and $p = .02$, $d = .74$ for posttest 4. Other between-group comparisons were not statistically significant. Overall the results showed that the effect sizes for Immediate CF (mean $d = 1.34$) were consistently larger than those for Delayed CF(mean $d = .73$) and Task Only (mean $d = .80$) and that the effects of Delayed CF and Task Only were similar. Therefore, we can conclude that with regard to the UGJT results, Immediate CF was the most effective and that the efficacy of Delayed CF and Task Only was comparable.

EIT Results

Findings of the omnibus test for the fixed-effects variables showed that the effects of the fixed-effects variables were all significant, $F(1, 157.92) = 177.07$, $p < .01$ for pretest scores; $F(3, 153.69) = 16.56$, $p < .01$ for group; $F(3, 350.15) = 23.12$, $p < .01$, for time; and $F(9, 350.06) = 4.13$, $p < .01$ for the interaction between group and time. The significant interaction effect between group and time indicated that the effects of the treatment types

TABLE 4. Group differences on UGJTs.

	Posttest 1		Posttest 2		Posttest 3		Posttest 4		
Group comparisons	p	d[a]	p	d	p	d	p	d	Md[b]
Immediate CF−Control	.02*	1.09$^{\#}$	<.01**	1.42$^{\#}$	<.01**	1.45$^{\#}$	<.01**	1.41$^{\#}$	1.34$^{\#}$
Delayed CF−Control	.07	.72	.01**	.84	.04*	.78	.27	.57	.73
Task Only−Control	.21	.62	<.01**	.97	.01*	.85	.02*	.74	.80
Immediate CF−Delayed CF	1.00	.34	.58	.53	.33	.62	.07	.80	.57
Immediate CF−Task Only	1.00	.38	1.00	.34	1.00	.49	1.00	.58	.45
Delayed CF−Task Only	1.00	.06	1.00	−.16	1.00	−.10	1.00	−.19	−.10

[a]Cohen's *d*.
[b]Mean effect size.
$^{*}p \le .05$; $^{**}p \le .01$; $^{\#}$ Cohen's $d \ge 1.00$.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

TABLE 5. Group differences on EITs.

Group comparisons	Posttest 1		Posttest 2		Posttest 3		Posttest 4		
	p	*d*[a]	*p*	*d*	*p*	*d*	*p*	*d*	*Md*[b]
Immediate CF−Control	<.01**	1.06#	<.01**	2.03#	<.01**	2.26#	<.01**	1.33#	1.67#
Delayed CF−Control	.35	.47	< .01**	1.41#	<.01**	1.29#	.02*	.67	.96
Task Only−Control	.31	.51	< .01**	1.19#	<.01**	1.55#	.03	.70	.99
Immediate CF−Delayed CF	.60	.53	1.00	.50	.09	.84	.72	.59	.61
Immediate CF−Task Only	1.00	.51	.21	.77	.54	.62	.87	.59	.62
Delayed CF−Task Only	1.00	−.03	1.00	.24	1.00	−.22	1.00	−.02	−.01

[a]Cohen's *d*.
[b]Mean effect size.
*$p \leq .05$; **$p \leq .01$; # Cohen's $d \geq 1.00$.

in relation to each other varied significantly at the four posttesting time points. Bonferroni post-hoc analyses were carried out to identify group differences on the four EIT posttests after taking the influence of pretest scores into account. Table 5 presents the *p* values, the effect sizes and the mean effect sizes of the post-hoc comparisons. As can be seen, the Immediate CF group was the only group that significantly outperformed the Control group on all four posttests with large effect sizes, $p < .01$, $d = 1.06$ for posttest 1; $p < .01$, $d = 2.03$ for posttest 2; $p < .01$, $d = 2.26$ for posttest 3; and $p < .01$, $d = 1.33$ for posttest 4. The Delayed CF group significantly outperformed the Control group on posttest 2, $p < .01$, $d = 1.41$ (large); posttest 3, $p < .01$, $d = 1.29$ (large); and posttest 4, $p = .02$, $d = .67$ (small). The Task Only group significantly outperformed the Control group on posttest 2, $p < .01$, $d = 1.19$, and posttest 3, $p < .01$, $d = 1.55$, both with large effect sizes. Other between-group comparisons did not reach significance. Overall, the EIT results suggest that Immediate CF (mean $d = 1.67$) was more effective than Delayed CF (mean $d = .96$) and Task Only (mean $d = .99$), and that the effects of Delayed CF and Task Only were similar.

DISCUSSION

The study investigated if immediate and delayed feedback operationalized as corrective recasts showed differential effects on the acquisition of the English past tense. The answer to this question is affirmative. Results of this study showed that Immediate CF was the only group that performed significantly better than the Control group on all four posttests. Immediate CF also showed larger mean effect sizes on the UGJTs ($d = 1.34$) and the EITs ($d = 1.67$) than Delayed CF ($d = .73$ on the UGJTs and $d = .96$ on the EITs), and Task Only ($d = .80$ on the UGJTs and $d = .99$ on the EITs). Furthermore, Delayed CF and Task Only were more effective than Control, and they demonstrated similar effects on L2 development. Note that the effects of different treatment types were compared based on their mean effect sizes because the differences of the posttest scores between the three experimental groups were not statistically significant. The lack of statistical differences was probably due to the small group sizes and a larger sample may have made the pattern more unequivocal (Li et al., 2016b).

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

The larger effects of immediate feedback in comparison with delayed feedback are subject to multiple interpretations. First, it would seem that feedback is most effective before errors are proceduralized through communicative practice. In the immediate feedback condition, errors were corrected when they first occurred in the learners' interlanguage, and the adjusted, rectified system was reinforced and proceduralized in further practice activities. In contrast, the learners in the delayed feedback condition did not receive corrections on their errors until three weeks later. The incorrect knowledge of the target language may have been proceduralized and reinforced through repeated use of the wrong forms in subsequent communicative tasks. The learners in this condition, therefore, had to unlearn the proceduralized errors through feedback in the final session. Here, we would like to extend the notion of proceduralization, a term in Skill Acquisition Theory referring to the application and integration of linguistic knowledge or declarative knowledge in actual linguistic performance, such as using a grammar rule in oral production. Once proceduralized, linguistic knowledge is programmed into a system and is available for use "as a ready-made chunk" (DeKeyser, 2015, p. 95). Proceduralized knowledge may become automatized through repeated practice. We would like to argue that similar to correct linguistic knowledge, incorrect knowledge or errors can also be proceduralized and even automatized if they are not remedied at early stages of the interlanguage.

Second, it would seem that feedback may be particularly useful at a time when it is easy to be connected with the instruction learners receive about the linguistic target regardless of the nature of the linguistic target. In the case of the immediate feedback condition in this study, the feedback may have reinforced the effects of the brief grammar instruction because it was provided immediately after the instruction. Because of the proximity of the instruction to the feedback, the learners were able to make an association between the received feedback and the knowledge learned from the instruction, for both regular and irregular past-tense forms, while the knowledge was in an active state. However, in the case of delayed feedback, where feedback was provided three weeks after learners received the explicit instruction, feedback may have worked separately from the instruction. When the delayed feedback was provided, the learners were less likely or less able to link the feedback with the instruction, and the function of the feedback became purely corrective rather than reinforcing (i.e., increasing the effects of the instruction). Whether the interval between instruction and feedback influences the effectiveness of feedback is an empirical question that has not been examined in the research. However, previous research did show that the effects of feedback can be elevated by explicit instruction, and in those studies feedback and instruction are contiguous rather than separate, suggesting that bundling instruction and feedback may be more effective than keeping them apart. For example, a number of studies (Saito, 2013a, 2013b; Saito & Lyster, 2012; Sato & Loewen, 2018) reported that explicit information followed by CF was more effective than CF or explicit instruction alone. Li et al. (2016b) reported that explicit instruction followed by within-task feedback was the most effective among the various configurations of form-focusing instruction including pretask instruction, within-task feedback, and posttask feedback. Although these studies did not examine delayed feedback, they show that bundling instruction and feedback has powerful effects on learning gains and that tying them together may be more effective than keeping them separate—as what happened in the delayed feedback condition in this study.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Third, whether feedback is followed by practice activities may have contributed to the differential effects of immediate and delayed feedback (See Li et al., 2016a). In this study, whereas immediate CF was followed by two practice sessions (four communicative tasks) to reinforce the effects of feedback, delayed CF was provided in the final session, after which there was no communicative practice. It is possible that had there been practice activities after delayed feedback, the difference between delayed and immediate feedback may have disappeared. Previous studies found that "contextualized practice" (Sato & McDonough, 2019, p. 24) promoted L2 development. Although the importance of practice activities to strengthen and proceduralize the knowledge learned from feedback has not been examined in previous research, there has been indirect evidence supporting this argument. For example, Li et al. (2016a) and Quinn (2014) both compared the effects of online feedback (provided during a communicative task) and offline feedback (provided during the posttask stage). Whereas the former study showed an advantage for online feedback, the latter found no difference between the two types of feedback. One methodological difference between the two studies is that the offline feedback in Li et al.'s study was provided after learners completed both communicative tasks, whereas in Quinn's study, it was provided during the interval between the two tasks. In other words, the learners in Li et al.'s study had no practice opportunities after receiving feedback while those in Quinn's study did. Therefore, it is perhaps the unavailability of practice activities after feedback that caused the smaller effects of offline feedback in Li et al.'s study. By the same token, it is probably because of the lack of practice opportunities after feedback that caused delayed feedback to be less effective than immediate feedback in this study. Certainly, there is an obvious need for empirical investigation to verify this hypothesis.

Finally, we would like to highlight a finding that is often ignored in feedback research, that is, the group that only performed communicative tasks made substantial gains. They significantly outperformed the Control group on posttests 2–4 on the UGJT and on posttests 2 and 3 on the EIT. There were no significant differences between Task Only and Delayed CF, and their effect sizes were similar, d = .80 and .99 for the Task Only group and d = .73 and .96 for the Delayed CF group on UGJT and EIT, respectively.[11] These results suggest that performing communicative tasks was as effective as receiving delayed feedback, with the caveat that had there been practice opportunities, the effects of Delayed CF may have been greater. Certainly, it can be argued that the effects of the Task Only condition may be attributable to the instruction the learners received at the outset of the treatment. However, in other studies in which learners did not receive any instruction, the learners who only performed communicative tasks scored significantly higher on posttests than on pretests (Yang & Lyster, 2010) or in comparison with a control group who did not receive any treatment (Li et al., 2016a). However, it is noteworthy that in the studies that showed significant effects for Task Only, the treatment tasks (e.g., dictogloss) contained input materials providing models of the correct use of the linguistic structure. In contrast, in studies using purely output-based tasks such as picture description (Arroyo & Yilmaz, 2018; Ellis et al., 2006; Révész, 2009), the performance of communicative tasks without any form-focused instruction (either in the form of pretask instruction or CF) did not seem to generate any learning gains, at least in the short term and in terms of the acquisition of new linguistic knowledge. Therefore, we argue that for task-based instruction to facilitate the development of new linguistic knowledge, tasks that integrate input and output such as the dictogloss tasks used in this and other studies are necessary.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

CONCLUSION

This study is the first study to examine whether CF should be provided immediately after learners' initial exposure to a linguistic structure or after learners practice using the structure in communicative activities. The study fills an important gap and contributes significantly to the theory and pedagogy of second language learning. Most SLA theories make explicit claims about the importance of CF, albeit from different perspectives, but few elaborate the influence of the timing of feedback on L2 development. Pedagogically, teachers and students seem to diverge on whether errors should be corrected immediately in that while teachers are hesitant to correct students' errors immediately, students favor immediate feedback (Li, 2017). In line with teachers' attitudes, teacher guides also recommend that students must be given chances to solve problems independently and teachers should not "simply tell students about how language works" (Brown, 2007, p. 291). Our study examined this controversy empirically, and our findings supported the utility of immediate CF. Based on the findings and methodological features of this study, we recommend that feedback be provided immediately after learners' initial exposure to the linguistic feature and at the beginning of an instruction cycle. We argue that errors should be addressed before they are proceduralized in the L2 system. We also argue that should CF be accompanied by instruction, they should be delivered in tandem rather than separately, so as to achieve maximal effects for both. Finally, keeping in mind that empirical research is in order, we recommend providing practice opportunities after CF so learners have chances to apply the knowledge learned through CF.

Based on the preceding arguments, we would like to tentatively point out that the takeaway message based on the research on the timing of feedback is that feedback should be immediate, whether "immediate feedback" is operationalized as online feedback embedded within a communicative task (vs. offline feedback) or as feedback provided at the initial phase of an instructional cycle (vs. feedback after a period of delay). However, we realize that this conclusion is based on limited research, and we therefore call for more research on the timing of CF.

Future research may address the following limitations of the study. First, the pretask grammar explicit instruction may have neutralized the influence of feedback type and made it less likely to distinguish the effects of different types of feedback. The impact of feedback timing may have been greater had the explicit instruction been reduced or left out. Second, feedback was provided in only one of six treatment tasks, and a larger dose of feedback may have resulted in larger differences between the effects of feedback treatments and the task-only instruction. Third, delayed feedback was not followed by further productive practice, which would have augmented the effects of delayed feedback and mitigated the advantage of immediate feedback. Last but not least, this study only included the target verbs in the achievement tests and it is unclear whether the learners will be able to apply the acquired grammatical rule of regular past tense to new verbs. Further studies may select both target verbs and new verbs to measure L2 development of rule-based grammatical structures.

NOTES

[1]One anonymous reviewer suggested that the distinction between declarative and procedural knowledge rather than explicit and implicit knowledge can be drawn on to describe L2 measures. However, we would like to adhere to the paradigm of explicit and implicit knowledge following Ellis et al. (2009), where the tests we used to measure treatment effects were validated as measures of explicit and implicit knowledge, not declarative and

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

procedural knowledge. The relationship between the two sets of terms is subtle and controversial, and some researchers consider them interchangeable depending the research paradigm one works in. For example, Dörnyei (2009) pointed out that "declarative and procedural knowledge correspond to explicit and implicit knowledge, respectively" (p. 147). From his point of view, the difference between declarative/procedural knowledge and explicit/implicit knowledge lies in the "scientific approach" (Dörnyei, 2009, p. 147) rather than the "content" (Dörnyei, 2009, p. 147), as the terms "declarative knowledge" and "procedural knowledge" grew out of artificial intelligence research, while the terms "explicit knowledge" and "implicit knowledge" were developed in the research paradigm in which explicit and implicit learning were explored.

[2]Certainly there is an interface between the context and schedule of CF: Both online and offline feedback can be immediate or delayed. The research has only examined immediate online and offline feedback, and this study investigates immediate online and delayed online feedback.

[3]According to the curriculum of the participating school, the learners were scheduled to learn the English past tense in the second semester of Year 7, which was two months after the completion of the study.

[4]One anonymous reviewer pointed out that the effect of feedback may have been conflated with that of the grammar instruction when the treatment groups were compared with the Control group. However, the three treatment groups all received the grammar instruction, so any difference between their learning gains must have been due to (a) the presence or absence of feedback (feedback treatment vs. task only) and (b) whether feedback was immediate or delayed (feedback timing).

[5]One anonymous reviewer questioned whether the spot-the-difference task creates a context for the natural use of the past tense. The authors would like to clarify that during the task, the learners received pictures and texts that contained past forms, studied the materials, and then compared their pictures without access to the texts. It would seem natural for the learners to use the past tense when they described their pictures in group work. Excerpts from the spot-the-difference task are provided in Appendix G.

[6]Doughty and Varela (1998) did not include this step in their corrective recasts.

[7]Ellis (2005) validated elicited imitation as a measure of implicit knowledge. However, Suzuki and DeKeyser (2015) found that elicited imitation was a measure of automatized explicit knowledge. One methodological difference between the two studies is that in Ellis's study learners were simply told to repeat sentences they were presented with in correct English, while in Suzuki and DeKeyser's study, learners were required to correct grammar errors while repeating the sentences, which may have increased learners' chances of accessing their explicit knowledge.

[8]Learners' knowledge of the target verbs was assessed through a vocabulary test. The new verbs identified through the test were taught by one of the researchers using flashcards and the learners were given a vocabulary list to study at home. Group leaders selected by their classroom teachers checked if their group members were familiar with the words prior to the experiment.

[9]The four posttest scores of the four groups on UGJT or EIT were included in each model at the same time.

[10]The interactions between pretest scores and the other two independent variables (group and time) were not significant (see Appendix B) and thus were not included in the models.

[11]One reviewer argued that the finding that task only worked equally well as delayed feedback contradicts the overall significant effects of CF. We would like to clarify that we do not mean to deny the effects of CF; rather the argument is that the performance of communicative tasks is facilitative of L2 development, unlike what is assumed by those who object to a task-based approach.

REFERENCES

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

Arroyo, D. C., & Yilmaz, Y. (2018). An open for replication study: The role of feedback timing in synchronous computer-mediated communication. *Language Learning*, *68*, 942–972.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*, 689–725.

Bates, D. (2010). *lme4: Mixed-effects modeling with R*. Springer.

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*, *20*, 436–458.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Brown, H. (2007). *Teaching by principles: Interactive language teaching methodology* (3rd ed.) Person Education.

Butler, A., Karpicke, J., & Roediger, H. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*, 273–281.

Choi, S., & Li, S. (2012). Corrective feedback and learner uptake in a child ESOL classroom. *The RELC Journal*, *43*, 331–251.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge University Press.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 94–112). Routledge.

Dörnyei, Z. (2009). *The psychology of second language acquisition*. Oxford University Press.

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

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

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

Ellis, R. (2015). *Understanding second language acquisition*. Oxford University Press.

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

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

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge.

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

Erlam, R., & Loewen, S. (2010). Implicit and explicit recasts in L2 oral French interaction. *The Canadian Modern Language Review*, *66*, 887–916.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*, 445–474.

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

Harmer, J. (2007). *The practice of English language teaching*. Pearson.

Kim, J., & Han, Z. (2007). Recasts in communicative EFL classes: Do teacher intent and learner interpretation overlap? In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 269–297). Oxford University Press.

Kim, Y. (2013). Effects of pretask modeling on attention to form and question development. *TESOL Quarterly*, *47*, 8–35.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory. *Studies in Second Language Acquisition*, *37*, 549–581.

Krashen, S. (1981). *Second Language Acquisition and Second Language Learning*. Pergamon.

Kulhavey, R., & Anderson, R. (1972). Delay-retention effect with multiple choice tests. *Journal of Educational Psychology*, *63*, 505–512.

Kulik, J., & Kulik, C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97.

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

Li, S. (2014). The differential roles of two aptitude components in mediating the effects of two types of feedback on the acquisition of an opaque linguistic structure. *AAUSC 2013 Volume-Issues in Language Program Direction*, 32–52.

Li, S. (2017). Student and teacher beliefs and attitudes about oral corrective feedback. In H. Nassaji & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 143–157). Routledge.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Li, S. (2020). What is the ideal time to provide corrective feedback? Replication of Li, Zhu & Ellis (2016) and Arroyo & Yilmaz (2018). *Language Teaching*, *53*, 96–108.
Li, S., Ellis, R., & Kim, J. (2020). The influence of pre-task grammar instruction on L2 learning: An experimental study. *Studies in English Education*, *23*, 831–857.
Li, S., Ellis, R., & Zhu, Y. (2016a). Task-based versus task-supported language instruction: An experimental study. *Annual Review of Applied Linguistics*, *36*, 205–229.
Li, S., Ellis, R., & Zhu, Y. (2019). The associations between cognitive ability and L2 development under five different instructional conditions. *Applied Psycholinguistics*, *40*, 693–722.
Li, S., & Vuono, A. (2019). Twenty-five years of research on oral and written corrective feedback in System. *System*, *84*, 93–109.
Li, S., Zhu, Y., & Ellis, R. (2016b). The effects of the timing of corrective feedback on the acquisition of a new linguistic structure. *The Modern Language Journal*, *100*, 276–295.
Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, *27*, 361–386.
Long, M. (2015). *Second language acquisition and task-based language teaching*. John Wiley & Sons.
Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*, 399–432.
Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19*, 37–66.
Lyster, R., & Ranta, L. (2013). Counterpoint piece: The case for variety in corrective feedback research. *Studies in Second Language Acquisition*, *35*, 167–184.
Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302.
Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*, 1–40.
Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive international feedback? *Studies in Second Language Acquisition*, *22*, 471–497.
Mackey, A., & Philp, J. (1998). Conversational Interaction and Second Language Development: Recasts, Responses, and Red Herrings? *Modern Language Journal*, *82*, 338–356.
Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classroom: A study of guided planning and relativization. *Language Teaching Research*, *12*, 11–37.
Mullet, H., Butler, A., Verdin, B., Von Borries, R., & Marsh, E. (2014). Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *Journal of Applied Research in Memory and Cognition*, *3*, 222–229.
Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*, 287–311.
Nassaji, H. (2016). Anniversary article Interactional feedback in second language teaching and learning: A synthesis and analysis of current research. *Language Teaching Research*, *20*, 535–562.
Ortega, L. (2009). *Understanding second language acquisition*. Hodder Education.
Philp, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, *25*, 99–126.
Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*, 267–278.
Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
Quinn, P. (2014). *Delayed versus immediate corrective feedback on orally produced passive errors in English* [Unpublished PhD dissertation, University of Toronto].
Quinn, P., & Nakata, T. (2017). The timing of oral corrective feedback. In H. Nassaji & E. Kartchava (Eds.), *Corrective feedback in second language teaching and learning: Research, theory, applications, implications* (pp. 35–47). Routledge.
Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*, 437–470.
Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*, 906–911.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Saito, K. (2013a). Reexamining effects of form-focused instruction on L2 pronunciation development: The role of explicit phonetic information. *Studies in Second Language Acquisition*, *35*, 1–29.

Saito, K. (2013b). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English r to adult Japanese learners. *Language Learning*, *63*, 499–529.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /r/ by Japanese learners of English. *Language Learning*, *62*, 595–633.

Sato, M. (2013). Beliefs about peer interaction and peer corrective feedback: Efficacy of classroom intervention. *Modern Language Journal*, *97*, 611–633.

Sato, M., & Loewen, S. (2018). Metacognitive instruction enhances the effectiveness of corrective feedback: Variable effects of feedback types and linguistic targets. *Language Learning*, *68*, 507–545.

Sato, M., & McDonough, K. (2019). Practice is important but how about its quality? Contextualized practice in the classroom. *Studies in Second Language Acquisition*, *41*, 1–28.

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

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

Skinner, B. (1953). *Science and Human Behavior*. The Free Press.

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

Spada, N., Jessop, L., Tomita, Y., Suzuki, W., & Valeo, A. (2014). Isoltaed and integrated form-vosued instruction: Effects on different types of L2 knowledge. *Language Teaching Research*, *18*, 453–473.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895.

van de Guchte, M., Rijlaarsdam, G., Braaksma, M., & Bimmel, P. (2015). Learning new grammatical structures in task-based language learning: The effects of recasts and prompts. *Modern Language Journal*, *99*, 246–262.

VanPatten, B. (2004). Input processing in SLA. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–31). Erlbaum.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. https://arxiv.org/abs/1308.5499

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

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory and language analytic ability. *Applied Linguistics*, *34*, 344–368.

Zhao, Y. (2015). *The effects of explicit and implicit recasts on the acquisition of two grammatical structures and the mediating role of working memory* [Unpublished PhD dissertation, The University of Auckland, New Zealand].

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

APPENDIX A

METHODOLOGICAL DETAILS OF THE CF TREATMENTS IN THE TWO CF GROUPS

Group	Task	Number of past tense errors	Number of CF	Number of presenters
Immediate CF	Task 1	34	31 (R[1]:17; IR:14[2])	20
	Task 2	29	27 (R:15; IR:12)	16
Total		63	58 (R:32; IR:26)	36
Delayed CF	Task 1	22	20 (R:13; IR:7)	20
	Task 2	20	19 (R:12; IR:7)	18
Total		42	39 (R:25; IR:14)	38

[1]Number of corrected regular verbs.
[2]Number of corrected irregular verbs.

APPENDIX B

SUMMARY OF THE LIKELIHOOD RATIO TEST RESULTS FOR THE INTERACTIONS WITH PRETEST SCORES

	Likelihood ratio test		Test results	
Dataset	Test	Tested interaction effects	χ^2	p
UGJT	Test 1	Pretest score × Group × Time	$\chi^2(9) = 15.55$	.08
	Test 2	Pretest score × Group	$\chi^2(3) = 6.74$	.08
	Test 3	Pretest score × Time	$\chi^2(3) = 1.50$	.68
EIT	Test 1	Pretest score × Group × Time	$\chi^2(9) = 7.34$	.60
	Test 2	Pretest score × Group	$\chi^2(3) = 2.02$	.57
	Test 3	Pretest score × Time	$\chi^2(3) = 7.25$	.06

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

APPENDIX C

MEANS AND SDS OF UGJT AND EIT SCORES FOR REGULAR VERBS

		Immediate CF	Delayed CF	Task only	Control
UGJT Pretest	Mean	1.96	1.40	2.35	.95
	SD	2.90	2.22	2.58	2.13
	n	38	41	33	30
UGJT Posttest 1	Mean	5.39	4.90	5.53	2.62
	SD	3.18	2.72	3.79	3.47
	n	37	40	31	29
UGJT Posttest 2	Mean	6.17	5.09	6.36	2.83
	SD	3.35	2.99	3.67	3.30
	n	36	38	32	30
UGJT Posttest 3	Mean	6.70	5.54	6.36	2.92
	SD	3.53	3.37	4.37	3.63
	n	38	40	33	30
UGJT Posttest 4	Mean	6.68	4.93	5.81	3.08
	SD	3.47	3.43	4.28	3.61
	n	38	41	27	30
EIT Pretest	Mean	1.06	.87	1.12	1.13
	SD	1.34	1.17	1.10	1.37
	n	35	35	33	31
EIT Posttest 1	Mean	2.62	2.31	2.43	1.75
	SD	1.96	1.74	1.79	1.55
	n	33	31	29	30
EIT Posttest 2	Mean	3.63	3.08	3.14	1.65
	SD	1.79	2.03	2.15	1.93
	n	30	33	32	30
EIT Posttest 3	Mean	3.89	3.10	3.73	1.81
	SD	1.90	2.11	2.28	1.96
	n	31	34	32	31
EIT Posttest 4	Mean	3.53	2.77	3.57	2.30
	SD	1.77	1.93	1.91	2.02
	n	31	32	28	30

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

APPENDIX D

MEANS AND SDS OF UGJT AND EIT SCORES FOR IRREGULAR VERBS

		Immediate CF	Delayed CF	Task only	Control
UGJT Pretest	Mean	1.54	1.57	2.02	.85
	SD	1.90	1.91	2.02	1.61
	n	38	41	33	30
UGJT Posttest 1	Mean	4.12	3.23	3.58	1.59
	SD	2.32	2.28	2.45	2.20
	n	37	40	31	29
UGJT Posttest 2	Mean	4.97	3.92	4.69	1.80
	SD	3.01	2.62	3.13	2.42
	n	36	38	32	30
UGJT Posttest 3	Mean	4.79	3.48	4.44	1.93
	SD	3.14	2.34	3.00	2.32
	n	38	40	33	30
UGJT Posttest 4	Mean	4.70	3.27	4.54	1.78
	SD	2.89	2.43	3.14	2.27
	n	38	41	27	30
EIT Pretest	Mean	1.83	2.01	2.88	2.13
	SD	1.39	1.63	1.76	1.63
	n	35	35	33	31
EIT Posttest 1	Mean	4.24	3.23	4.28	2.92
	SD	1.78	2.00	1.83	1.78
	n	33	31	29	30
EIT Posttest 2	Mean	5.28	4.59	5.33	2.72
	SD	1.98	2.19	2.15	1.57
	n	30	33	32	30
EIT Posttest 3	Mean	5.55	4.40	5.36	2.39
	SD	1.69	2.33	1.57	1.65
	n	32	34	32	31
EIT Posttest 4	Mean	4.65	3.92	4.59	3.02
	SD	1.97	2.13	1.93	1.92
	n	31	32	28	30

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

APPENDIX E

UNTIMED GRAMMATICALITY JUDGMENT TEST

Class: Name:

Instructions: In this test, you are going to decide if the following 34 sentences are GRAMMATICALLY correct or incorrect (NOTE: the number of correct and incorrect sentences is not the same). If the sentence is grammatically correct, please choose "correct." If the sentence is ungrammatical, please correct the sentence in the brackets labeled "incorrect." There is NO time limit for the test so you can take as much time as you need to check your answers. Remember once you have made your judgment, you are NOT allowed to go back to revise your previous answer.

1. Sarah will go to America tomorrow.
   correct ( ) incorrect ( )
2. Linda gets up at 7 o'clock every day.
   correct ( ) incorrect ( )
3. Look! They are having breakfast.
   correct ( ) incorrect ( )
4. Jim get a new computer last Friday.
   correct ( ) incorrect ( )
5. John plant a tree last Sunday.
   correct ( ) incorrect ( )
6. We walks to a supermarket yesterday.
   correct ( ) incorrect ( )
7. Tim answers my questions last Monday.
   correct ( ) incorrect ( )
8. Amy decides to wash her clothes last Tuesday.
   correct ( ) incorrect ( )
9. Lily is changing her car last month.
   correct ( ) incorrect ( )
10. Tom will notice a change in Mary yesterday.
    correct ( ) incorrect ( )
11. Lily is listening to a story last night.
    correct ( ) incorrect ( )
12. Nancy will return a book this morning.
    correct ( ) incorrect ( )
13. Mike will start to learn Chinese yesterday.
    correct ( ) incorrect ( )
14. We usually have a school trip in autumn.
    correct ( ) incorrect ( )
15. Jenny hand in her art work two days ago.
    correct ( ) incorrect ( )
16. My father will say goodnight to me last night.
    correct ( ) incorrect ( )
17. They drink water every morning.
    correct ( ) incorrect ( )
18. The police is finding a baby last Tuesday.
    correct ( ) incorrect ( )

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

19. Boys always run fast in PE classes.
    correct ( ) incorrect ( )
20. Cindy throw her bag away three days ago.
    correct ( ) incorrect ( )
21. Tom become a father last week.
    correct ( ) incorrect ( )
22. Jim is laugh at me two days ago.
    correct ( ) incorrect ( )
23. Tom usually starts class at 8 o'clock.
    correct ( ) incorrect ( )
24. James stop his car in our school yesterday.
    correct ( ) incorrect ( )
25. Amy and Tom is going to school yesterday.
    correct ( ) incorrect ( )
26. Toby will break a window two weeks ago.
    correct ( ) incorrect ( )
27. Luke begin to play computer games last year.
    correct ( ) incorrect ( )
28. Betty is making a cake last week.
    correct ( ) incorrect ( )
29. Jack will have a new English teacher next week.
    correct ( ) incorrect ( )
30. My father eat a banana this morning.
    correct ( ) incorrect ( )
31. Listen! Some girls are singing now.
    correct ( ) incorrect ( )
32. Jack will take a shower this morning.
    correct ( ) incorrect ( )
33. My mother is making dinner now.
    correct ( ) incorrect ( )
34. Alice tell Tom a story last night.
    correct ( ) incorrect ( )

APPENDIX F

ELICITED IMITATION TEST

Instructions: In this test, you are going to hear 24 statements one by one. After listening to each statement, you should indicate whether you agree or disagree with the statement based on your situation. If you agree, press the LEFT shift button. If you disagree, press the RIGHT shift button. After making the judgment, you should reproduce the sentence in CORRECT English. Remember to make the judgment and reproduce each sentence AS SOON AS POSSIBLE. Now let's start with eight practice items. After doing the practice items, you are going to do the test.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press

Practice items	Test items
I am a boy.	I make (made) breakfast this morning.
I like reading books.	My mother take (took) a picture for me last Saturday.
I am a Year 6 student.	My teacher said "Good Morning" to me yesterday.
There are 45 students in my class.	My sister walked to her school yesterday.
We have five English lessons every week.	My sister stopped learning English last year.
I can draw pictures.	My father begin (began) to play football last year.
My mother is a teacher.	My friend go (went) to school by bus this morning.
My father is a doctor.	I return (returned) ten books to my friend last week.
	I changed my English name last year.
	I plant (planted) some flowers last Tuesday.
	I find (found) ten yuan on the ground last Sunday.
	My uncle decide (decided) to study in the USA last January.
	I got a new book two days ago.
	My teacher laugh (laughed) to me this morning.
	I start (started) to learn English six years ago.
	My brother broke his arm two weeks ago.
	I became a year 5 student last year.
	I tell (told) my friend an English story last week.
	I threw a book through a window yesterday.
	I handed in my homework this morning.
	I answered three English questions yesterday.
	My mother listened to an English song yesterday.
	I ate an orange this morning.
	I notice (noticed) a broken window yesterday.

APPENDIX G

EXCERPT FROM THE SPOT-THE-DIFFERENCE TASK

A: Wu Yifan defeat John a football game.
B: Oh, in my picture, John defeated Wu Yifan in a basketball game.
A: The teacher asked John difficult math questions.
B: The teacher ask Wu Yifan difficult English questions. Different.
A: John dropped his glasses.
B: John drobbed his iphone. I have his iPhone in my picture.
A: Wu Yifan threw John's iPhone on the ground.
B: Are you picture 7?
A: No, picture 8.
B: Picture 8 is John threw Wu Yifan's glasses. Yes, John threw Wu Yifan's glasses.

https://doi.org/10.1017/S0272263120000388 Published online by Cambridge University Press