Language Teaching (2020), **53**, 3–28
doi:10.1017/S0261444819000375

CAMBRIDGE UNIVERSITY PRESS

STATE-OF-THE-ART ARTICLE

Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges

Hossein Nassaji

University of Victoria, British Columbia, Canada
E-mail: nassaji@uvic.ca

Abstract

How to correct learner errors has long been of interest to both language teachers and second language acquisition (SLA) researchers. One way of doing so is through interactional feedback, which refers to feedback provided on learners' erroneous utterances during conversational interaction. Various theoretical claims have been made regarding the beneficial effects of interactional feedback, and over the years a considerable body of research has examined its effectiveness. In this context, a central and challenging question has always been how to determine whether such feedback is effective for language learning. Studies investigating the role of feedback have used various measures to assess its usefulness. In this paper, I will begin with a brief overview of the recent studies examining interactional feedback, with a focus on how its effectiveness has been assessed. I will then examine the various measures used in both descriptive and experimental research and discuss the issues associated with such measures. I will conclude with what continues to pose us a challenge in assessing the role of feedback and offer some recommendations to inform future research in this area.

Introduction

Interactional feedback is a kind of corrective feedback that learners receive on their erroneous utterances in the course of conversational interaction. Such feedback occurs through various forms of negotiation and modification strategies such as reformulation, repetitions, clarification requests, and the like, when students are engaged in communicating their meaning. The following provides an example of an interactional feedback exchange:

Example 1:

Teacher:	What did you do on the weekend?	
Student:	I watch a movie.	
Teacher:	You watched a movie.	*Feedback*
Student:	Yeah, I watched a movie	

The study of interactional feedback has received much attention in the field of SLA and second language (L2) instruction. A key question in this context has always been whether and how such feedback assists L2 acquisition. In this respect, opposing claims have been made from both theoretical and pedagogical perspectives. Theoretically, the argument for the role of feedback has related closely to whether or not there is any need for negative evidence in language acquisition. Negative evidence is the information that indicates to the learner that his or her target language use is somewhat incorrect. This has been contrasted with positive evidence that simply provides the correct model of the target language use. Nativists, for example, have claimed that language acquisition is driven mainly by some biologically innate linguistic principles called Universal Grammar (UG), and consequently,

© Cambridge University Press 2020

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

acquisition occurs mainly as a result of exposure to correct input or positive evidence (e.g., Flynn, 1988; Schwartz, 1993). Interactionists, on the other hand, have argued that language acquisition occurs as a result of interaction and the feedback learners receive during interaction. They have contended that negative evidence exists in interaction between parents and children and also between native speakers-nonnative speakers and that it also plays a critical role in language acquisition (e.g., Long, 1996; Gass, 1997, 2003; Pica, 2002).

Pedagogically, the argument has concerned the extent to which corrective feedback helps learners improve their language accuracy. The strong versions of communicative and task-based language teaching argue that the goal of language instruction is the development of communicative competence. Therefore, instruction should be exclusively meaning-focused. Error correction also interrupts the communication process, and thus, the use of corrective feedback should be minimized. The weaker version of the task-based instruction (e.g., Skehan, 1998; Ellis, 2003) and also the focus on form (FonF) approach (e.g., Long, 1996; Doughty & Williams, 1998; Long & Robinson, 1998), on the other hand, hold that purely meaning-focused approaches are inadequate for the development of L2 accuracy particularly among adult L2 learners, and thus, besides opportunities for meaning-focused communication, learners also need opportunities for corrective feedback and attention to form (e.g., Ellis, 2001; Doughty, 2003; Williams, 2005).

Given the various claims about the role of corrective feedback, a considerable body of research has explored its use and potential benefits (see Nassaji, 2015; Nassaji & Kartchava, 2017). This research has been both descriptive and experimental and has been conducted in various contexts including both laboratory and classroom contexts. Earlier research was mostly descriptive, aiming to document the presence of feedback in communicative interaction (e.g., Chaudron, 1977; Pica, 1987; Pica, Young, & Doughty, 1987; Gass & Varonis, 1994). More recent research has moved beyond description, attempting to examine the link between feedback and L2 acquisition (e.g., Doughty & Varela, 1998; Mackey & Philp, 1998; Ishida, 2004; Lyster, 2004; Ammar & Spada, 2006; Nassaji, 2017). It has also examined the various mechanisms and factors that may influence feedback effectiveness (see Nassaji, 2015, 2016 for a review).

Assessing feedback effectiveness

A key question in feedback research has always been whether, and to what extent, feedback is effective. Studies have used various measures to assess feedback effectiveness ranging from various forms of learner responses to feedback (including uptake and learner repair) and individualized tailor-made tests in descriptive research to different kinds of pretest-posttest measures in experimental research. The question that arises in this context is what these tests measure and to what extent they provide evidence for L2 learning. In the subsequent sections, I will discuss some of the most widely used outcome measures in corrective feedback research and the issues associated with them. I will begin with measures used in descriptive research and then those in experimental research.

Measures used in descriptive research

Uptake and repair

In much of the research on interactional feedback, particularly descriptive research, the effectiveness of feedback has been assessed through learner immediate responses to feedback or what has been called 'uptake' (e.g., Lyster & Ranta, 1997; Lyster, 1998; Ellis, Basturkmen, & Loewen, 2001; Panova & Lyster, 2002; Lyster & Mori, 2006; Sheen, 2006; Fu & Nassaji, 2016; Hawkes & Nassaji, 2016). The term 'uptake' was initially used in classroom research by Allwright (1984) and Slimani (1989, 1992) to describe events in classroom interaction and the opportunities they provide for learning. Lyster and Ranta (1997) then used it to refer to learner responses following feedback. According to these researchers, if there is no uptake, there is topic continuation. Thus, uptake could be any response,

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

including *yes*, *no* and even acknowledgment of the feedback. The following provides an example of a feedback move followed by uptake and another followed by topic continuation.

Example 2:

S:	She is walk on the beach.	
T:	She is walking on the beach?	*Feedback*
S:	Yeah, she is walking on the beach.	*Uptake*

Example 3:

T:	I have a white cat with white ear.	
S:	With white ears?	*Feedback*
T:	It is very cute.	*Topic continuation*

(Nassaji, 2015)

Descriptive research has generally shown that learners produce uptake in response to feedback but these responses vary depending on the type of feedback and other factors (e.g., Lyster & Ranta, 1997; Lyster, 1998; Ellis et al., 2001; Lyster & Mori, 2006; Sheen, 2006; Fu & Nassaji, 2016; Hawkes & Nassaji, 2016). However, there are a number of issues with the notion of uptake and how it has been used in feedback research. I will discuss two main issues: THE DEFINITION/INTERPRETATION ISSUE and THE LEARNING-RELATIONSHIP ISSUE.

Issues with uptake

The definition/interpretation issue. One issue with uptake concerns the definition and/or the interpretation of uptake. The actual meaning of uptake is taking up or making use of something. Furthermore, as noted earlier, the term was originally used to refer to what learners report they have learned from a language lesson. This is, however, different from the meaning used in most corrective feedback research in which the term has been used most often to refer to any responses following feedback (e.g., Lyster & Ranta, 1997; Lyster, 1998; Ellis et al., 2001; Panova & Lyster, 2002; Lyster & Mori, 2006). This discrepancy between the actual meaning of uptake and the way it has been used in feedback research can be problematic and at times confusing because when a study reports that a particular feedback type has led to uptake, the reader might assume that the learner has 'uptaken' the feedback in one way or another. However, that is not necessarily the case. For example, both Lyster and Ranta (1997) and Panova and Lyster (2002) reported that all the instances of elicitation led to uptake whereas this meant that all the instances of these feedback moves were followed by a learner response. Lyster and Ranta (1997) categorized uptake into learner repair and needs repair. This categorization identified the nature of the learner response. However, needs repair was further defined as utterances that included responses such as learner acknowledgments and hesitations. Needs repair defined as such cannot be a subcategory of uptake because such responses cannot necessarily indicate that the learner has noticed or understood the corrective force of the feedback. Ellis et al. (2001) classified uptake into successful and not successful uptake. This classification can also be problematic as any uptake should logically involve some degree of success; otherwise, there is no uptake. If there is uptake in any real sense of the term, it cannot be fully unsuccessful. It seems contradictory to state that there is uptake but at the same time saying that it is unsuccessful.

Some researchers have defined uptake as a learner response that involves some modification of the learner's output following feedback (Mackey & Philp, 1998; McDonough, 2005). In some of my work, I have used the term 'learner repair' instead of uptake (Nassaji, 2007a, 2011). Defining uptake as

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

modified output or repair provides a more sensible definition of uptake as the presence of modified output or repair can provide an indication that the learner has made use of the feedback in some way. Learners' modification of their output may also indicate that the learner has noticed the gap between their non-target-like output and the correct form (e.g., Ellis & Sheen, 2006).

However, while the presence of repair or modified output may indicate that the learner has noticed the feedback, learner repair cannot still be interpreted as a measure of L2 learning. Students may notice or even learn from the feedback but may still not repair their output in response to feedback. This can happen in cases when the feedback does not provide an opportunity for modified output (e.g., Oliver, 2002). Learners may also repair their output but the repair can be simply a mechanical repetition of the feedback (Nassaji, 2011). These possibilities then lead us to the question of the relationship between uptake/repair and acquisition, which I will discuss below.

The learning-relationship issue. Although it can be assumed that there is a relationship between uptake as modified output or learner repair and L2 acquisition, the exact role of learner repair or modified output in language learning is far from clear. It might be assumed that modified output may indicate that the learner has learned from the feedback. However, studies examining the relationship of this with learning have not been able to find a conclusive result. While some have shown a positive relationship (e.g., Nabei & Swain, 2002; McDonough, 2005), others have either failed to show a connection (Mackey & Philp, 1998; McDonough & Mackey, 2006) or have found that the relationship depends on the nature of the modified output or repair (Nassaji, 2011).

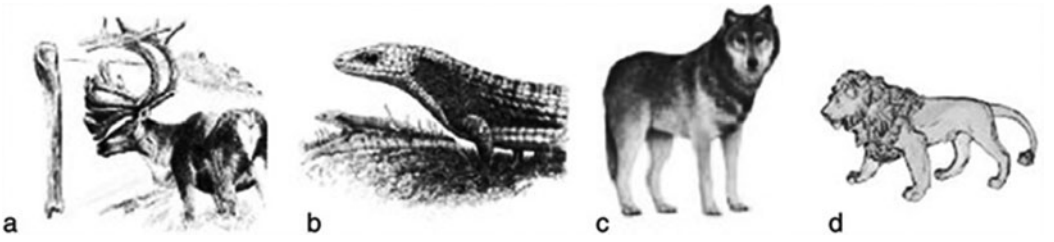
Mackey and Philp (1998), for example, found that at least 95% of the learners' responses to recasts were not modified. However, learners still showed evidence of development. Given such findings, they stated that responses to recasts could be red herrings as they do not indicate that the learner has made any use of the feedback. Nassaji (2011) identified two types of repair: repair that involved repetition of the feedback and repair that involved the incorporation of the correct form into a longer and new utterance. He found a positive relationship between incorporation of the feedback and learning but not between repetition of the feedback and learning (see also Nassaji, 2007b). Gurzynski-Weiss and Baralt (2015) found that learners' noticing of the feedback was differentially related to the type of modified output following feedback with partial modified output being the strongest predictor of feedback noticing. These findings suggest that the relationship between repair or output modification and learning is not straightforward and depends at least on the nature of the repair (see also the next section).

Individualized posttests

To provide more direct evidence for the relationship between feedback and learning in descriptive research, some studies have used what is called INDIVIDUALIZED POSTTESTS. Such tests are tailor-made and designed based on the feedback each learner receives on an erroneous target form and are then administered to the same learner after the interaction. The aim is to examine the effects of feedback on forms targeted incidentally during interaction. Individualized posttests are constructed based on the focus on form episodes (FFEs) in which a target language form has occurred. The construction of individualized posttests begins with the identification of such episodes and then the creation of the test items based on those episodes. The following from Nassaji (2010, p. 919) provides an example of an FFE and a test item developed based on that episode.

Example 4:
Focus on form episode (FFE)
S1: I don't know wolf.
S2: Wolf is a . . . wolf is a [gesturing with hands] . . .
T: Wolf [opening and closing hand like the mouth of a dog]. Ah . . . similar to a dog. A wild dog is a wolf.
Individualized posttest item based on FFE
Which animal is a wolf?

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press



Several studies have used individual posttests to assess the effectiveness of interactional feedback and FonF (Loewen, 2002, 2005; Nabei & Swain, 2002; Nassaji, 2010; Kim & Nassaji, 2018). Loewen (2005), for example, used such tests to investigate the effectiveness of incidental feedback in communicative English as a second language (ESL) classrooms. He found a significant relationship between feedback and the learners' correction of the targeted form after the interaction. Nassaji (2010) used individualized posttests to examine the effects of preemptive and reactive FonF including interactional feedback. He found higher individualized posttest scores for preemptive FonF than reactive FonF (see also Kim & Nassaji, 2018).

Hawkes and Nassaji (2016) used an innovative form of individualized posttests (i.e., video-based stimulated correction tests) to assess the effectiveness of incidental recasts. The method involved identification and extraction of oral episodes in which a student had made an error. Two types of such episodes were identified (i.e., +recast and −recast episodes). These episodes were then edited to produce clips ending with the student's erroneous utterance. After the interaction, each student was shown both types of clips on a computer screen and was asked to identify and orally correct his or her own errors. The tests were individualized since each student was shown exchanges that contained his or her own utterances. The results showed that students successfully modified a greater proportion of their errors on +recast episodes than on –recast episodes.

Issues with individualized posttests

The advantage of individualized posttests is that they allow for the assessment of incidental feedback. They also enable the researcher to analyze the extent to which each individual learner has benefited from the feedback. One limitation, however, is the lack of pretests (Loewen & Philp, 2006). This is problematic because, without a pretest, it is hard to know whether the learner has really learned the targeted form. Although the error made by the learner during the interaction can be considered as an indication that the learner did not know the correct form, it is still unclear whether the error is due to a lack of knowledge, partial knowledge, or simply a slip of the tongue during spontaneous interaction (Nassaji, 2009). Another issue is that the test simply assesses the learner's recall of the targeted form involving receptive skills, and not learners' production. Individualized posttests are also often written whereas the interaction is oral. Thus, if learners do not provide a correct response on the test, it could also be partly due to the mismatch between the modality of the test and the context of the feedback (this can apply to all written tests used to assess oral interaction). One way of addressing this issue would be through using tests in which students both hear the stimuli and then respond orally such as in Hawkes and Nassaji's (2016) aural-oral individualized posttests.

Individualized posttests are often created based on errors made in oral spontaneous discourse. Although the occurrence of such errors may be because the learner has not acquired the form, as noted earlier, it may also be due to random mistakes in the production of known forms. Thus, relying on spontaneous oral errors to measure L2 acquisition can be problematic (see Nassaji, 2009). The fact

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Table 1. An example of a learner's pre-interaction description, during-interaction description, teacher feedback, and post-interaction modification

a. Learner's pre-interaction description	'Lisa was a lucky guy because there was an old woman who helped Lisa to get her money back.' Error*: guy* Error type: *Lexical*
b. Learner's during-interaction description	'Although, I think Lisa is a lucky guy . . . Error: *guy* Error type: *Lexical*
c. Feedback	'a lucky girl?' Feedback: *Isolated recast*
d. Learner's post-interaction correction	'Lisa was a lucky because there was an *girl* old woman who helped Lisa to get her money back.'

(Nassaji, 2009, p. 431)

that the learner makes oral errors could of course be also because the learner has declarative and conscious knowledge about the form but the form has not yet been fully internalized or proceduralized.

To address the pretest limitation of individualized posttests, Nassaji (2009) used a pretest-posttest design for individualized tests, which attempted to assess the learners' prior knowledge of the forms that were targeted by the feedback during interaction. The data involved adult ESL learners who participated in task-based dyadic interactions and received various forms of interactional feedback. To provide a measure of the learner's previous knowledge, in addition to the dyadic oral interaction, the research design included three other components: a pre-interaction scenario description component, an immediate post-interaction error identification/correction component, and a delayed error identification/correction component. Initially the pre-interaction scenario descriptions were examined and the learners' erroneous utterances were identified. The erroneous utterances were then tracked and compared to those occurring during oral interaction to identify errors in common. Finally, the post interaction scenarios were examined to determine the extent to which learners were able to correct the errors that occurred prior to interaction and then received feedback during interaction. Table 1 provides an example of such an analysis. In this example, in the pre-interaction scenario description task, the learner produced an utterance with a lexical error: *guy* for *girl*. Then during the dyadic interaction, the learner produced the same error and received a recast. When analyzing the scenario description after the interaction, the learner identified and successfully corrected the error by changing the word *guy* to *girl*. This correction was then taken as evidence for the effect of feedback during the interaction.

Language related episodes (LREs)

Related to the notion of FFEs discussed earlier is also that of LANGUAGE RELATED EPISODES. LREs have been defined as 'any part of the dialogue where learners talk about the language they are producing, question their language use, or correct themselves or others' (Swain & Lapkin, 2002, p. 292). LREs do not necessarily involve providing/receiving feedback. However, they have been used in interaction research to identify if or when feedback occurs incidentally during interaction (e.g., Williams, 2001; Nabei & Swain, 2002).

Because of their purpose, LREs have often been categorized based on what the learner has focused on during interaction, including meaning (lexical or word choice) or various features of grammar (word order, tense, morphology, syntax, preposition, etc.). They have also been sometimes classified depending on other variables such as discourse, content, or the learners' language proficiency (e.g., Williams, 1999). The following from Williams (2001, p. 329) provides an example of an LRE from a pairwork activity in which two students are interacting. When one of the students makes an error, the teacher notices it and provides feedback on the learner's erroneous utterance.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Example 5:

Learner:	Do you know what time. how often. Can you tell me where food… fast food restaurant is?
Other Learner:	uh…
Learner (trigger):	**the** good fast food restaurant is? **The** good fast food restaurant… restaurant is?
Teacher (feedback):	mmhm and since it's **a** good fast food restaurant, you could say a good uhuh, cause we haven't said which one. If you said which one, you'd say **the**. When you don't really know which, you say **a**.
Learner (uptake/repair):	MacDonald's. Do you know where that is? Is there **a** MacDonald's?

LREs have been used not only to identify moments in interaction when learners focus on language form but also to track development. Lapkin, Swain, and Smith (2002), for example, analyzed transcripts of learner talks during task-based collaborative dialogue and trace the development of the language features in French (e.g., pronominal verbs) that were the focus of LREs, from the learners' initial collaborative draft to their posttests. This enabled them to find evidence for the learners' progress in the correct use of the targeted form. Williams (2001) identified LREs that learners were engaged in during collaborative activities. She then analyzed the relationship between the items targeted in the LREs and the learners' performance on learner-specific posttests.

Issues with LREs

LREs are useful techniques for not only identifying what kind of language form the learner has attended to but also their distributions and the nature and types of turns. One issue is the time-consuming nature of their identification as the researcher needs to transcribe and code all the interaction data to determine where the LREs have occurred and what features they have focused on. The reliability of LREs is another issue. The issue of reliability is not specific to LREs and relates to all assessment tools that involve interpretation and making inference about learners' performance. The coder reliability can usually be addressed by using more than one coder. However, although the researcher or the coders may decide that certain aspects have been attended to during the interaction, this interpretation may not necessarily be in line with the learner's intention. In addition, whereas each LRE may be coded as dealing with one certain linguistic form, there is also a possibility that multiple target structures or feedback types may take place in the same LRE. There is also a possibility of the learners' variable responses to the same feedback in an LRE. These different things going on during an LRE make it difficult to determine what contributes to the learners' learning as a result of participation in an LRE.

Experimental pretest-posttest research

To assess feedback effectiveness more directly, many studies have adopted an experimental research design using various pretest-posttest measures (Doughty & Varela, 1998; Han, 2002; Ishida, 2004; Lyster, 2004; McDonough, 2005; Ammar & Spada, 2006; Ellis, Loewen, & Erlam, 2006; Ellis, 2007; Loewen & Nabei, 2007; Yang & Lyster, 2010; Saito & Lyster, 2012; Nassaji, 2017). In experimental research, usually a particular type of feedback is selected and used as a treatment and its effects are assessed on learning some preselected target forms. Learners' pretest performance is then compared to their posttest performance and any increase is often taken to be due to the treatment. The findings of experimental pretest-posttest research have been reported in several narrative reviews and meta-analytic studies that combine results across studies (Russell & Spada, 2006; Mackey & Goo,

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

2007; Li, 2010; Lyster & Saito, 2010; Lyster, Saito, & Sato, 2012; Goo & Mackey, 2013; Nassaji, 2015, 2016; Brown, 2016). The findings of these studies have indicated that feedback in general has a positive effect on L2 learning. Table 2 summarizes the findings of some of these meta-analyses, all of which

Table 2. Overall effectiveness of feedback in different meta-analyses

Russell and Spada (2006)	15 corrective feedback studies including 10 oral feedback studies and 5 written feedback studies	A large overall effect (the mean effect size was 1.16) with oral feedback having a smaller effect size than written feedback
Mackey and Goo (2007)	28 interaction studies with 20 oral feedback studies including classroom and laboratory studies	A medium effect in immediate posttests (0.71) and a large effect in delayed posttests (1.09)
Lyster and Saito (2010)	15 oral classroom studies	An overall medium effect size (0.74)
Li (2010)	33 oral classroom and laboratory studies	A medium effect size (0.64)

report a medium to a large effect size for feedback.

Although experimental research has generally shown positive effects for feedback, there are a number of issues/limitations with pretest-posttest experimental research. One is that most of these studies have focused mainly on the changes in the pretest and posttest scores with much less concern with the details of what goes on during interaction. Some experimental studies have analyzed learner repair during interaction (e.g., Yang & Lyster, 2010). However, these studies have often reported the overall frequency of the repair without showing who, or how many of the students who received the feedback, produced the repair. This can be problematic if all or most instances of the repair reported come from only a few of the participants.

Experimental studies have been mostly dealing with planned feedback or feedback provided intensively and repeatedly on certain predetermined target forms. The study of preplanned feedback has provided important insights into the nature of effective feedback. However, because of their controlled focus, it is hard to know whether the effects shown are because of the feedback, its intensity, or their combination. For example, the positive effects shown for recasts in experimental research could be partly due to the intensity of the feedback on specific target forms (Nicholas, Lightbown, & Spada, 2001; Han, 2002; Lyster et al., 2012).

There is also the problem of ecological validity of intensive feedback in experimental research as teachers often provide feedback broadly on a wide range of forms instead of certain preselected forms. For example, in the context of discussing the role of recasts, Ellis and Sheen (2006) argued that although intensive recasts (i.e., recasts that occur repeatedly in response to preplanned forms) may be viewed to be more effective than incidental recasts, such claims are 'of little practical significance for teachers because they are not in a position to know whether their learners are developmentally ready or whether the targeted feature is one amenable to treatment through recasts' (p. 597). These researchers believe that recasts that occur incidentally and in response to a wide range of errors might have greater practical significance.

Issues with assessment tasks and testing instruments

A variety of measures have been used in experimental research to determine the effects of feedback or instruction. Norris and Ortega (2000) classified them into four types: (a) free constructed responses, (b) constrained-constructed responses, (c) selected responses, and (d) metalinguistic judgments. Free constructed-response measures are those in which learners produce the linguistic feature in tasks that involve free production. Constrained-constructed response measures require learners to produce a target form in controlled tasks such as in fill-in-the-blank tests. Selected-response measures need learners

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

to select the correct response among a number of alternatives such as in multiple-choice tests. In metalinguistic judgments, learners should decide whether an utterance is grammatically correct or not. Norris and Ortega's classification is mainly based on the test format. For the purpose of this article, I will categorize them based on their function or the kind of knowledge measured. In doing so, I distinguish four broad categories: (a) explicit knowledge tests, (b) implicit knowledge tests, (c) functional/pragmatic knowledge tests, and (d) measures of noticing, perception, and L2 processes.

Explicit knowledge tests

In the field of SLA, a distinction has been made between explicit and implicit knowledge. Williams (2009, p. 319) defined explicit knowledge as 'knowledge that we know that we know . . . and that we are aware of using.' Ellis (2004) defined it as declarative knowledge about language and numerated a number of characteristics for it such as being held conscious, learnable, verbalizable, and typically accessed through controlled processing when learners experience some kind of linguistic difficulty in using the L2. Implicit knowledge, on the other hand, involves little conscious awareness. It is also procedural and 'is available for use in rapid, fluent communication' (Ellis, 2005b, p. 214).

Although some researchers do not posit an essential role for explicit knowledge in SLA (e.g., Krashen, 1985; Reber, 1993), many believe that explicit knowledge contributes importantly to L2 acquisition (e.g., Sharwood Smith, 1981; Anderson, 1983; Hulstijn, 2002; Ellis, 2004, 2005a; DeKeyser, 2005). For example, using information processing theories and skills acquisition models of learning, some researchers have argued that language learners learn language initially through processing information available in the input and then accessing it for subsequent comprehension and production (Ellis, 2008). Skills acquisition theories (see Anderson, 1982, 1983) claim that learning is a movement from declarative knowledge (i.e., explicit knowledge of rules and systems) to procedural knowledge (i.e., knowledge of how to use the system). This perspective proposes that learners can learn the new target rules and structures first through the development of conscious knowledge and then practice them to gain control of them (e.g., McLaughlin, 1990; DeKeyser, 2005). A variety of measures have been used to assess explicit knowledge both in SLA and L2 feedback research. Three most commonly used measures are metalinguistic knowledge tests, untimed grammaticality judgment tests, and error correction tests.

Metalinguistic knowledge tests (MKTs). MKTs assess learners' metalinguistic knowledge. The operationalization of the notion of metalinguistic knowledge has varied across studies. However, this kind of knowledge has often been defined as the learners' ability to describe and explain errors using their knowledge of grammatical rules, with an important component of this knowledge being learners' analytic ability (Bialystok, 1982, 1991). Roehr (2007, p. 174) defined metalinguistic knowledge as 'explicit knowledge about [grammatical] categories as well as explicit knowledge about relations between categories'. MKTs can take different forms but in general learners are presented with a series of ungrammatical sentences and are asked to explain their incorrectness using metalinguistic terms. For example, in a study of explicit and implicit corrective feedback, Ellis et al. (2006) presented learners with five ungrammatical sentences containing the target structure (i.e., English past *–ed*) and asked learners to correct the errors and explain why the sentence was ungrammatical. Since such tasks require the learner to analyze the sentence and also use grammatical concepts, they are considered to encourage the use of explicit knowledge.

Untimed grammaticality judgment tests (GJTs). GJTs assess learners' ability to judge the grammaticality of sentences. Such tests are often composed of a series of grammatically correct and incorrect sentences and require participants to indicate which is correct and which is incorrect. Learners' responses are often in the form of categorical judgments but learners may also sometimes be given choices such as 'clearly grammatical', 'probably grammatical', 'probably ungrammatical', and 'clearly ungrammatical' (Ellis & Barkhuizen, 2005).

GJTs can tap into either explicit or implicit knowledge depending on the time condition or the grammaticality of the sentences. For example, untimed GJTs are considered to provide a measure

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

of explicit knowledge (e.g., Ellis, 2005a; Ellis & Barkhuizen, 2005; Ellis et al., 2006; Erlam, 2009) as learners have time to monitor their language using their conscious knowledge. Several feedback studies have used untimed GJTs to examine the effect of feedback on the development of explicit knowledge (e.g., Ellis et al., 2006; Loewen & Erlam, 2006; Loewen & Nabei, 2007; Kamiya, 2015; Nassaji, 2017; Thomas, 2018). For example, in a recent study, Nassaji (2017) used written grammaticality judgment tasks to examine the effects of extensive versus intensive recasts. The task consisted of sentences that provided obligatory contexts for the use of the target structure (i.e., English articles). Half of the sentences were grammatical and the other half were ungrammatical. Learners were asked to read each sentence and decide whether it contained an error and also correct the error.

At times, learners may be asked not only to judge and correct the sentence, but also to say what their judgment was based on (e.g., feel, rule, guess). In Ellis et al.'s (2006) study, for example, in addition to judging the grammaticality of the sentences, learners were asked to indicate the degree of their certainty and also indicate whether they used a language rule or their intuition to make a judgment. The latter requirements were used to determine to what extent learners were sure of their judgments or relied on their explicit versus implicit knowledge to make a decision.

Error correction tests (ECTs). ECTs measure explicit knowledge by asking learners to correct erroneous sentences (see Muranoi, 2000; Sheen, 2008). The difference between ECTs and GJTs is that in the latter learners need to judge what sentence is grammatical and what sentence is not but in the former the erroneous part is identified and the learner needs to correct it. The format is often written and the learner is asked to correct the error in writing. Since such tests require learners to correct errors, they may pay attention to the language structure in the tests using their conscious language knowledge. Also, since learners can perform the task under no time pressure, they are assumed to have time to monitor their performance using their explicit analytic knowledge. In a study of the relationship of recasts and learner anxiety, Sheen (2008), for example, used an ECT consisting of two related statements, one of which contained an error, which had been underlined. The learners were required to correct the underlined erroneous part. The study found no effects of recasts on learners' performance on this test, which was interpreted as being possibly due to the difference between the nature of the treatment (involving oral input) and the nature of the test (involving written analytic knowledge).

Issues with explicit knowledge tests

Explicit knowledge tests have been frequently used in feedback and instructed L2 research. In their meta-analysis, Norris and Ortega (2000) found that 90% of the studies on feedback and instruction had used such tests, including multiple-choice tests. The situation has changed since then and studies have begun to use tests claimed to measure implicit knowledge. I will examine some of these measures commonly used in feedback research in the next section. However, before doing so, I will discuss some of the issues with explicit knowledge tests.

First, explicit knowledge tests such as GJTs and MKTs are assumed to measure learners' conscious awareness of grammatical forms. However, we do not yet know what consciousness exactly is and how and when learners access and use their conscious knowledge. Furthermore, the assumption underlying most GJTs is that grammaticality is a dichotomous notion as sentences are often judged as being grammatical or ungrammatical. However, linguists have debated whether grammaticality should be treated as a dichotomous or gradient notion (Tremblay, 2005). Sorace and Keller (2005) argued that it is sometimes difficult to make a clear-cut line between what is grammatically well-formed and what is not as there are varying degrees of well-formedness. Therefore, they suggested the use of a grammaticality scale when developing and using GJTs (see also Sorace, 1996).

It has been assumed that when learners perform GJTs under no time pressure, they have time to pay attention to language forms and also monitor their language use, and in doing so they may rely on their conscious knowledge (e.g., Ellis, 2005a). However, although attention and consciousness are two related concepts (Posner, 1994; Velmans, 1996), the notion of attention is different from awareness or

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

consciousness (Kentridge, Heywood, & Weiskrantz, 1999; Lamme, 2003; Koch & Tsuchiya, 2007) Attention is a cognitive resource which can be allocated to any task whereas consciousness is a state of the mind. Attention may enhance detection and hence may bring an object to our consciousness. However, we may also become conscious of an event or an object without explicit attention (Koch & Tsuchiya, 2007).

In addition, it is quite possible to respond to an error correction test or an untimed GJT item successfully using implicit knowledge as correcting ungrammatical sentences does not necessarily require declarative conscious grammatical knowledge. Native speakers, for example, may have little explicit or declarative knowledge of how the language system works, but they are still able to perform well on GJTs. The kind of knowledge used during a GJT also depends on learners' level of language proficiency (Roehr, 2007). The more proficient the learner, the more likely they use their implicit knowledge.

The validity of GJTs has also been found questionable on several other grounds (Birdsong, 1989; Hedgcock, 1993; Munnich, Flynn, & Martohardjono, 1994; Sorace, 1996; Ellis, 2004; Tremblay, 2005). Hedgcock (1993) argued that grammaticality judgments are affected by extra-linguistic factors such as processing and cognitive constraints (e.g., Hedgcock, 1993). Munnich et al. (1994) pointed out that such tests do not simply elicit linguistic behavior but a response indicating learners' beliefs about the L2 grammar. Learners' sentence processing strategies, the context of the sentence, and the language learning environment are also factors that can influence grammaticality judgments (Sorace, 1985).

Implicit knowledge tests

Although explicit knowledge can be helpful and, as assumed by some theories, can become implicit knowledge through practice, most researchers agree that the kind of knowledge underlying spontaneous and fluent language use is implicit knowledge because accessing explicit declarative knowledge is a slow process whereas accessing implicit knowledge is fast. As noted earlier, explicit knowledge is declarative and involves conscious awareness whereas implicit knowledge involves little conscious awareness, is proceduralized, and can be accessed fluently and spontaneously. It has been argued that the ultimate goal of feedback and instruction should be the development of implicit knowledge, and therefore, research in this area should also include measures that tap into implicit knowledge (Ellis, 2004). To assess implicit knowledge, a number of measures have been used in SLA and feedback research (e.g., Ellis et al., 2006, 2009; Loewen & Erlam, 2006; Loewen & Nabei, 2007; Révész, 2012; Kamiya, 2015; Nassaji, 2017). Four most widely used ones are timed GJTs; oral elicited imitation tests (OEITs), speeded dictation (SD), and free oral production tests (FOPTs). In what follows, I will discuss these measures and the issues associated with them.

Timed GJTs. As noted earlier, the GJT is composed of a series of grammatically correct and incorrect sentences and the learners are asked to judge which of them is grammatical and which is ungrammatical. These tests are suggested to measure either explicit or implicit knowledge depending on the time condition or the grammaticality of the sentences (i.e., whether the learners' responses are to grammatical or ungrammatical sentences). As discussed previously, untimed GJTs are assumed to assess explicit knowledge. Timed GJTs, however, are assumed to measure implicit knowledge because the time pressure in such tests gives the learner little opportunity to rely on explicit knowledge (e.g., Ellis, 2005a; Ellis & Barkhuizen, 2005; Ellis et al., 2006; Erlam, 2009). Responses to grammatical sentences of GJTs are considered to tap into implicit knowledge whereas responses to ungrammatical sentences tap into explicit knowledge (Ellis, 2004, 2005a; Ellis et al., 2006). Ellis (2005a) showed some support for the latter by finding that while learners' scores on ungrammatical sentences were more correlated with tests of explicit knowledge (e.g., metalinguistic knowledge tests), responses on grammatical sentences were more associated with tests assumed to measure implicit knowledge, such as free oral production tests and timed GJTs (see also Gutiérrez, 2013). GJTs can be written or aural. Aural GJTs are more cognitively demanding than written tests as the learners need to process the

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

language under real operating conditions. Therefore, they might provide a better measurement of implicit knowledge (Kim & Nam, 2017).

OEIT. The OEIT is a particular elicitation technique used initially in first language (L1) acquisition research to assess L1 development (see Slobin & Welsh, 1973, for a review). It has also received considerable attention in L2 research as a measure of general L2 proficiency (e.g., Hamayan, Saegert, & Larudee, 1977; Wu & Ortega, 2013; Tracy-Ventura, McManus, Norris, & Ortega, 2014) and also implicit knowledge (Ellis, 2005a; Ellis et al., 2006; Erlam, 2009; Erlam & Loewen, 2010; Spada, Shiu, & Tomita, 2015). The technique consists of asking learners to listen to an utterance that usually includes a target structure and then repeat it as exactly and correctly as possible. The assumption is that learners are able to repeat an utterance correctly if they have internalized the grammar and the lexicon needed to comprehend, parse, and reproduce the utterance (Rebuschat & Mackey, 2013). Such tasks are assumed to be constructive in that learners need to process the input and then reconstruct it using their own grammatical knowledge (Erlam, 2009). The OEITs that ask learners to repeat correctly ungrammatical sentences are suggested to be more constructive because in order to correctly and spontaneously reproduce ungrammatical sentences, learners must have already internalized the target structure as part of their linguistic system (Munnich et al., 1994). A number of studies have used OEITs to investigate the role of feedback in the development of implicit knowledge. Ellis et al. (2006), for example, used an OEIT to investigate the effects of recasts and metalinguistic feedback on the acquisition of the implicit knowledge of English past tense *–ed.* Results showed an advantage for metalinguistic feedback.

Speeded dictation tests (SDTs). SDTs present learners with a series of sentences containing the target structure and ask learners to write down quickly the sentences read to them by the teacher or the researcher. Because of the time pressure and their reconstructing nature, these tasks are also assumed to minimize learners' ability to use their explicit knowledge and therefore are considered to tap into implicit knowledge. Sheen (2008), for example, used an SDT to examine the effects of recasts among low- and high-anxiety learners. The test consisted of 15 sentences containing the target structure (English definite and indefinite articles). Learners were asked to listen to each sentence and write it down as quickly as they could and as exactly as they heard it. The study found that the low-anxiety learners outperformed the high-anxiety learners on this test, which was taken to indicate that recasts had an effect on the development of implicit knowledge among these learners.

Free production tasks (FPTs). Other tests used to measure implicit knowledge are FPTs. These include various kinds of unplanned oral or written tasks that provide opportunities for free and meaningful language production, such as story-telling tasks (e.g., Sheen, 2008; Nassaji, 2017), picture description tasks (e.g., Nassaji, 2017), narrative tasks (see Yang & Lyster, 2010, for an example of a written task and Ellis, 2005a and Nassaji, 2009, for examples of oral tasks), and story-retelling tasks, in which student read and rewrite a story (e.g., Sheen, 2008) or retell a story orally (e.g., Yang & Lyster, 2010). Various forms of information gap tasks such as spot-the-difference tasks, picture-sequencing tasks, and picture-drawing tasks have all been used to elicit and also create a context for interaction and meaning-focused language production.

Since FPTs require learners to produce the language without any preplanning while their focus is on meaning, they are considered to tap into implicit knowledge (Ellis, 2005a; Ellis et al., 2006; Révész & Han, 2006; Révész, 2012; Nassaji, 2017). The assumption is that no planning time will prevent the activation of explicit knowledge and will trigger implicit knowledge.

Issues with implicit knowledge tests

The implicit knowledge tests discussed can provide useful information about the degree to which students have developed implicit knowledge. However, there are a number of issues and limitations with such tests, which need to be considered.

Issues with timed GTTs: As noted earlier, one of the tests used in feedback research as a measure of implicit knowledge is timed GJTs. The assumption is that the time pressure does not allow learners to

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

draw on their explicit knowledge, and therefore, they are pushed to use their implicit knowledge. Some recent validation studies have provided empirical support for such tests as measures of implicit knowledge by showing a correlation between these tests and other tests assumed to measure implicit knowledge (Ellis, 2005a; Ellis et al., 2006; Ellis & Loewen, 2007; Bowles, 2011; Godfroid et al., 2015). Ellis (2005a), for example, examined the validity of several tests assumed to tap into either explicit or implicit knowledge, including an oral narrative test, written untimed and timed GJTs, an OEIT, and an MKT. Using factor analysis, he found that these tests loaded on two different factors. The timed GJT, the oral narrative test, and the OEIT were loaded on one factor. The MKT and the untimed GJT were loaded on a separate factor. Ellis interpreted these two factors as implicit and explicit knowledge. A similar pattern was also replicated in another study by Ellis and Loewen (2007) that reanalyzed the data using confirmatory factor analysis. While these findings are encouraging, there are a number of issues that challenge the validity of timed GJTs.

One has to do with what kind of knowledge, conscious or unconscious, learners actually use in such tests. As noted earlier, one important criterion used to distinguish implicit from explicit knowledge is the use of conscious knowledge. The involvement of conscious knowledge is often seen as a characteristic of explicit knowledge. However, it is hard to rule out the possibility of using conscious knowledge even when the GJT is timed. After all, whether timed or untimed, GJTs are still form-focused and hence they may direct learners' consciousness to grammatical forms (Suzuki, 2017).

As far as time pressure and sentence grammaticality are concerned, in a recent validation study, Vafaee, Suzuki, and Kachisnke (2017) showed that manipulating these two factors would not make GJTs measures of implicit knowledge (see also Gutiérrez, 2013). Their study raised the possibility that such timed tests might measure automatized explicit knowledge rather than implicit knowledge. Both types of knowledge are accessed quickly, but the latter involves conscious awareness. They also reviewed research evidence suggesting that explicit automatized knowledge can also account for fluent and automatic use of language (see Suzuki & DeKeyser, 2015). In an eye-tracking study involving GJTs, Godfroid et al. (2015, p. 270) also concluded that their findings 'point to a difference in the levels of automatic and controlled processing involved in responding to the timed and untimed tests'.

Of course, if we assume that implicit knowledge is developed when controlled processing turns into automatic processing, it can be argued that the speed with which learners perform a task would reflect access to implicit knowledge. However, if automatized explicit knowledge can also be drawn on fast, time pressure cannot be taken as a reliable condition for implicit knowledge. Indeed, in a recent study, Suzuki (2017) tested this issue by comparing timed GJTs with three comprehension real-time tests that were assumed to measure implicit knowledge (a word-monitoring task, a visual-world task, and a self-paced reading task. These tasks assess online grammatical sensitivity through measuring reaction time (see Suzuki & DeKeyser, 2015, 2017; Suzuki, 2017; Vafaee et al., 2017), and have been frequently used in psycholinguistic research (e.g., Jiang, 2004, 2007). The word-monitoring task requires participants to listen to a sentence for a target word. They are then asked to press a button as soon as they see the word while their attention is on meaning. In the visual-world task, participants are presented with a scene of a series of pictures on a computer screen for a short amount of time. Their eye movements are then tracked while they listen to sentences containing the target structure. The self-paced reading task requires participants to read a sentence word by word and as fast as possible while their attention is focused on meaning (see Suzuki, 2017, for detail). The research provided evidence that these real time tasks provide a better measure of implicit knowledge than the timed GJTs. Such tasks, however, have not been employed in feedback research, although they have the potential of being used in such research too.

Issues with OEITs: One issue with OEITs is that it is not clear whether they measure L2 knowledge or are simply a test of working memory (Bley-Vroman & Chaudron, 1994; Vinther, 2002). The ability to repeat a sentence verbatim depends on its length, including the number of words and the difficulty of the utterance. It may be possible to design OEITs that minimize reliance on memory by carefully controlling the length of the utterances. However, one challenge is determining the optimal length. What an optimal length of a sentence should be depends on a number of variables including the

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

complexity of the target structure as well as individual learner characteristics such as their level of language proficiency (Munnich et al., 1994). Another memory-related concern is the recency effect. Learners are often more likely to member the last item in a series of words than the rest of the utterance (Rundus, 1971). This poses a problem when learners are pushed to repeat particularly a long utterance. Learners' responses to OEITs are also influenced by other variables including listening skills, 'decoding, matching the perceived elements to their function and meaning, recalling, and producing the same linguistic elements' (Vinther, 2002, p. 62).

In OEITs, researchers often score learners' repetition performance. However, we also need to make sure that learners have processed the sentence for meaning. To address this issue, Ellis et al. (2006) required learners to indicate their beliefs about each statement first and then to repeat it orally in correct English. The assumption was that when learners are asked to state their belief, the task becomes a meaning-focused task and not one relying on memory only. However, although this procedure may overcome the memory issue with OEITs, it introduces another factor, which is the possibility of the learners' performance being affected by their initial comprehension. Since the task requires learners to both comprehend and produce the utterance, it is unclear whether it is a test of comprehension or production.

Finally, the test involves out-of-context sentences, which may also include sentences that do not occur in real communicative situations. Therefore, they do not reflect learners' ability in real-life language use. This problem also exists with other sentence-level tests such as GJTs, which are also decontextualized. GJTs also involve receptive skills. Therefore, they do not show whether learners are able to produce the target structure in spontaneous discourse.

Issues with FPTs: The data obtained through FPTs are informative especially with regard to learners' ability to use the language in free production. However, there are a number of limitations with such measures. FPTs require learners to produce the language without any pre-planning while their attention is on producing the language in meaning-focused contexts. As noted earlier, when learners do not have time to plan, they may be unable to rely much on their explicit knowledge and hence may be more likely to use their implicit knowledge (Ellis, 2005a). However, although unplanned tasks may have this advantage, one challenge is the difficulty to design FPTs in ways to elicit certain target structures in a naturalistic manner (Loschky & Bley-Vroman, 1993). To elicit certain target structures, it is possible to use more focused and structure-based tasks in which learners are pushed (for example, by the use of prompts) to use the target structure when completing the task. Some studies have used such focused tasks (VanPatten & Cadierno, 1993; Adams, Nuevo, & Egi, 2011; Nassaji, 2017). For example, Adams et al. (2011) used prompts to elicit past tense in oral production tasks. Nassaji (2017) used pictured-cued tasks to elicit the use of definite and indefinite articles in obligatory contexts. These procedures may help the elicitation of a certain target structure. However, when learners are pushed to use a target structure in a specific context, their attention may be drawn to that language form. This would then make the task more consciousness raising, promoting the use of explicit knowledge (Nassaji & Fotos, 2010).

The degree to which FPTs measure implicit knowledge also depends on the modality of such tasks. Such tasks, for example, can be both oral and written. The use of written tasks as a valid measure of implicit knowledge should be treated with caution as learners have time to monitor their language use. Oral tasks may provide a better measure of implicit knowledge as learners have less time to rely on their explicit knowledge.

Functional/pragmatic knowledge tests

All the tests discussed so far focus on learners' accuracy of grammatical forms. Even when communicative tasks are used, their scoring focuses almost exclusively on linguistic accuracy. Functional/pragmatic tests emphasize not only learners' ability to produce grammatically accurate sentences but also utterances that are contextually or pragmatically appropriate. These tests are usually context-specific and their scoring considers both accuracy and contextual appropriacy. They have often been used

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

in pragmatic research to assess learners' pragmatic ability. Examples of such tests include role plays designed to examine how learners understand and produce language appropriately in social interaction. Examples of more controlled forms of such tasks are dialogue and discourse completion tasks (DCTs) used to examine learners' ability to use a particular speech act, such as requests, refusals, and apologies, in simulated communicative contexts (Hinkel, 1997; Rose & Kasper, 2001; Kasper & Blum-Kulka, 2003; Koike & Pearson, 2005; Martínez-Flor & Usó-Juan, 2010; Bardovi-Harlig & Vellenga, 2012). In DCTs, learners are provided with a description of a situation and are asked to describe what they would say in that situation. In a dialogue completion task, they are presented with an incomplete dialogue and are asked to complete it. These tasks can also be both written and oral. Their oral form may tap more into implicit knowledge whereas their written form may also encourage explicit knowledge. DCTs may also differ in the degree of control, ranging from those that are close-ended such as multiple-choice DCTs to those that require learners to say whatever they would say in a specific situation.

Several feedback studies have used various forms of pragmatic knowledge tests including DCTs or dialogue completion tasks to assess the effect of feedback on learners' socio-pragmatic skills (Fukuya & Zhang, 2002, 2006; Koike & Pearson, 2005; Takimoto, 2006; Ajabshir, 2014; Nguyen, Do, Pham, & Nguyen, 2017). Fukuya and Zhang (2002), for example, investigated the effects of implicit recasts on Chinese learners' learning of English requests. Using DCTs, they found a significant effect of recasts on learning this pragmatic feature.

Although functional/pragmatic tests such as DCTs might be an effective way of gathering socially appropriate responses from the learner, one limitation is that they do not elicit naturally occurring speech (Beebe & Cummings, 1996). Another is the difficulty of designing such tests in ways that can elicit only the speech act targeted.

Assessing noticing, perception, and L2 processes

While the role of learner attention or the extent to which learners need to be consciously aware of certain linguistic forms in order to learn them is a contentious issue in the field of SLA, many researchers believe that learning cannot occur without some degree of attention to form or noticing (Schmidt, 2001; Doughty, 2003; Robinson, 2003; Leow, 2015; Ellis, 2016b). SLA research also suggests that the effectiveness of corrective feedback depends largely on the degree to which learners successfully notice and process the feedback (Ellis & Sheen, 2006; Ellis, 2016a; Nassaji, 2016).

Mental activities including noticing or L2 processing are not subject to direct observation. It is impossible to directly observe what goes on in a learner's mind when performing a task. However, there are indirect ways of getting some evidence. To this end, research has used a number of tools, including, for example, various forms of retrospective and introspective methods. In the following section, I will discuss five of such methods which have been used in feedback research to find out what learners notice or how they process feedback during interaction. These include immediate recall, stimulated recall, think-aloud protocols, eye-tracking methods, and other retrospective methods such as questionnaires and interviews.

Immediate recall. Immediate recall is a retrospective technique in which learners are asked immediately after the feedback to recall what they noticed. The assumption is that such information is accessible for further recall. One of the studies that used immediate recall is Philp (2003) to examine learners' noticing of recasts in dyadic task-based interaction. Noticing was operationalized as remembering recasts following the feedback and was measured by cued immediate recall, where, after each recast, the learner's response was interrupted by the sound of two knocks to cue the learner to recall the feedback. The results showed that learners were able to notice a substantial number of the recasts (60–70%). The amount of noticing was also found to be affected by learners' level of language proficiency as well as the length and the number of changes in the recast. Learners with a higher level of L2 proficiency were more likely to notice recasts than lower-level learners, and shorter recasts were better noticed than longer recasts.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Stimulated recall. Another research tool used to collect noticing or processing data in feedback research is stimulated recall (e.g., Mackey, Gass, & McDonough, 2000; Mackey, 2006; Egi, 2007, 2010; Bao, Egi, & Han, 2011). Stimulated recall is a retrospective method in which learners are required to recall their thought processes after a task is completed. It is called stimulated recall because learners are presented, after the interaction, with a clue or stimulus to recall their thoughts at the time they were doing the task (see Gass & Mackey, 2017). In corrective feedback research, learners are usually shown a video clip of their feedback exchange and are asked to pause the clip when they hear the feedback move and recall what they were thinking at the time they received the feedback. One of the feedback studies using such a method is Mackey et al. (2000), which examined learners' noticing of the feedback in task-based dyadic interaction. The researchers found that the learners perceived the target of the feedback differently depending on error types and feedback types (see also Carpenter, Jeon, MacGregor, & Mackey, 2006; Egi, 2007, 2010).

Think-aloud. The think-aloud is a concurrent strategy used to capture learners' thoughts when completing a task (Ericsson & Simon, 1993). When doing a think-aloud, learners are often asked to verbalize their thought processes while completing the task. The think-aloud reports are then recorded, transcribed, and coded. Think-alouds might be difficult to use at the time of oral feedback. However, they can be used to assess the effect of feedback during a subsequent task. Nipaspong and Chinokul (2010), for example, employed think-aloud protocols during a succeeding discussion task to examine the effect of explicit and implicit feedback on learners' noticing of a pragmatic feature (English refusals). Their results showed an advantage for implicit over explicit feedback.

Eye-tracking methods. Eye-tracking has recently attracted increasing attention among L2 researchers as a tool to assess how L2 learners notice and process L2 input. This is an online technique using computer-based technology to gather real-time data about how and where learners allocate their attention by recording their eye movements and the length of their fixations using video-based eye trackers (see Winke, Godfroid, & Gass, 2013; Conklin & Pellicer-Sánchez, 2016, for a discussion). Since the technique does not interrupt the task performance, it does not have the limitations of introspective tools that require learners to do two tasks at the same time. Eye-tracking is used when learners are processing the input naturally and hence 'provides a very rich moment-to-moment data source' (Conklin & Pellicer-Sánchez, 2016, p. 453). This method has been used in a number of L2 studies to investigate text processing, including reading (Kang, 2014), vocabulary learning (Godfroid, Boers, & Housen, 2013; Godfroid et al., 2018), syntactic processing (Frenck-Mestre, 2005), and also written feedback (Shintani & Ellis, 2013). Relatively fewer studies have employed this technique to examine oral feedback, but some are emerging. For example, in a recent study of recasts, McDonough, Crowther, Kielstra, and Trofimovich (2015) examined the relationship of learners' eye-gaze with their responses to recasts. The eye gaze lengths of both the feedback provider and the learner were tracked. Their results showed a significant relationship between the learner's and the interlocutor's mutual eye gaze and the learner's target-like responses to recasts.

Retrospective questionnaires and interviews. Retrospective questionnaires and interviews are other means that can gather information about learners' perception of the feedback, though very indirectly. In a classroom study of interactional feedback, Fu and Nassaji (2016), for example, used a retrospective questionnaire in which students were asked to indicate how often they received such feedback. The findings showed that the perception of the feedback frequency reported was very close to the actual frequency of the feedback. Tools such as reflective journals and post-interaction uptake sheets have also been used to assess learners' perception of the feedback (see Basturkmen, Loewen, & Ellis, 2004; Sheen, 2008; Loewen et al., 2009; Junqueira & Kim, 2013, for examples).

Issues with introspective and retrospective reports

Although introspective and retrospective measures such as immediate and stimulated recall provide important information regarding how learners perceive the feedback, they have a number of

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

shortcomings. First of all, not all mental processes are conscious, and thus they are less responsive to verbalization whether immediate or delayed. Also, in stimulated recall, learners are often asked to report their thought processes after the interaction. Although care should be taken to minimize the length of the gap between the feedback and the recall sessions, because of the interval between the feedback and the recall session, learners' recall might be negatively affected as learners may not be able to accurately remember what happened when asked to comment after the task is completed. Immediate recall may be more advantageous than stimulated recall in that sense as the former requires the learner to recall their thought processes immediately after each feedback move. However, since the recall is prompted and occurs during the task, the flow of the communication can be interrupted. In addition, since the learner is asked to do a different task, such measures are also prone to interference effect.

Issues have also been raised about the validity of online reports such as think-aloud measures (e.g., Rosa & O'Neill, 1999; Leow & Morgan-Short, 2004; Bowles, 2010; Leow, Grey, Marijuan, & Moorman, 2014). Two concerns are reactivity and veridicality (Ericsson, 2003). Reactivity refers to the idea that introspective verbalization could lead to disruption and change of cognitive and thinking processes. Veridicality concerns the extent to which verbal protocols accurately reflect cognitive processes. Of course, the degree to which concurrent verbalization is affected by these issues depends on the nature of the tasks. Leow and Morgan-Short (2004), for example, found no reactivity effect for the think-aloud when learners were performing a reading task. However, learners also vary in terms of their ability to verbalize. Therefore, they differ in the amount of information they can provide in introspective and retrospective reports. Researchers often train learners how to verbalize their thoughts, which is helpful. However, there is evidence that even trained and also highly proficient L2 learners may find it difficult to verbalize their thoughts concurrently while doing a task (see Bowles & Leow, 2005).

These limitations suggest that although introspective/retrospective measures can provide important insight into learners' thought processes, they cannot provide a valid measure of learners' noticing or L2 processes. In the case of feedback, if learners accurately recall or comment on the feedback, it can be assumed that noticing has occurred, but the reverse is not true.

Other issues and challenges

In addition to the issues discussed so far, there are a number of other issues and challenges when it comes to assessing the effectiveness of feedback. I will discuss these under the following two categories: ACQUISITION DEFINITION CHALLENGES and MEASUREMENT CHALLENGES.

Acquisition definition challenges. One of the challenges in L2 acquisition research including feedback research is how to define acquisition. This issue is important because if there is no consensus on how to conceptualize this construct, there will be no agreement on how to measure it. Ellis (2006b) provided three definitions for acquisition: (a) internalization of a new linguistic form, (b) increased control over a known form, and (c) progress along an acquisitional path. As he pointed out, most studies have examined acquisition in terms of increased accuracy over a language form. However, there are also studies that have defined acquisition in terms of sequence of stages (e.g., Mackey & Philp, 1998; Mackey, 1999; Mackey & Oliver, 2002; McDonough, 2005). Although these definitions suggest that acquisition is complex and multifaceted, it is clear that different conceptualizations of the same construct could lead to different research results, which makes comparison across studies difficult.

There are also different theoretical perspectives in SLA, which provide different views about what acquisition is and how it should be measured. Interactionists, for example, define acquisition in terms of learners' ability in using language in social interaction. Good outcome measures are then those that assess learners' skill to use the language in communicative interaction. Cognitivists distinguish between two types of knowledge: knowledge of what, or declarative knowledge and knowledge of how, or procedural knowledge. They then define acquisition as a movement from controlled to free production or declarative to procedural knowledge. From this perspective, good tests are those that

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

measure both types of knowledge and in particular learners' progress from controlled to automatized language use. Generativists consider language knowledge as idealized decontextualized knowledge residing in the learner's mind. They make a distinction between competence and performance and argue that performance is affected by extra-linguistic factors. Therefore, tests should measure competence not performance.

While these different views broaden our perspective of what acquisition is, they also make it difficult to agree upon approaches to research and data collection, and for the same reason make it hard to synthesize the results of different studies and arrive at a conclusive evaluation.

Measurement challenges. Another main challenge is that of measurement. As discussed before, there are theoretical distinctions among different types of knowledge and attempts to measure them. However, it is very difficult to know what kind of knowledge learners use to perform a task. It is also difficult to control the kind of knowledge learners use when doing a test. In addition, learners' performance on any test would vary depending on a number of variables, including the complexity of the test, learners' language proficiency, and the various strategies learners may use in doing the test. Some authors have also pointed out (Tarone, 1983, 1985) that language competence is not uniform but highly variable. For the same reason, depending on the social context in which learners operate and its complex interaction with other learner- and context-related variables, learners may perform differently on the same test or respond differently to the same test item. Ellis (1991), for example, found that learners' performance varied significantly when they were asked to do the same grammaticality judgment task a second time.

Conclusions and recommendations

In this paper, I have discussed the different assessment measures frequently used to evaluate the effectiveness of corrective and interactional feedback. Based on this, I would like to make some final remarks and offer some recommendations.

First, most measures, particularly those used in experimental research, have focused primarily on the product rather than the process of learning. That is, they have focused on what is learned rather than how it is learned. Testing the process of learning is also crucial as it helps us understand why L2 learners who are similar in many other respects learn differently from the same feedback type. It can also explain some of the variability seen in research results. As noted earlier, various retrospective and introspective verbal reports such as concurrent think-aloud, stimulated recall, and eye-tracking techniques have been used to capture learners' mental processes during a task. Although these measures have been used in feedback research, few studies have used a combination of both product- and process-oriented measures. Such a combination can provide more comprehensive information regarding the effect of feedback by showing not only what, but also how, learners have learned from the feedback.

As discussed earlier, acquisition is a complex and multidimensional construct. Thus, it is important to use multiple assessment measures. This could be not only tests of different types of knowledge (e.g., explicit and implicit) but also different skills (receptive and productive) as well as tests of accuracy and appropriacy. It is important to use multiple measures as they allow us not only to capture different levels of acquisition but also cross-validate findings across studies and measurements. Using multiple measures is also important because learners have been shown to display variability in their performance depending of the type of outcome measures used. Ellis (2006a), for example, found that learners exhibited different degrees of difficulty with the use of regular past tense (*–ed*) depending on the kind of tests used. For example, they showed a high score for this target form on explicit tests but a relatively low score for it on implicit tests. For adverb placement, however, they showed a moderately high implicit score but a low explicit score.

Although it might be difficult to use many measures in one study, we cannot conclude that learners have acquired a linguistic form successfully unless they are able to use it across a range of tasks and situations. Therefore, we need to be cautious when interpreting our results. In a study on the role of

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

task-based instruction, Shintani (2012, p. 269), for example, concluded that her results showed 'that the listen-and-do tasks led to the acquisition of both nouns and adjectives'. Such claims about acquisition are not warranted in the absence of tests that measure learners' ability in a variety of contexts. As Norris and Ortega (2003) highlighted, in assessing SLA, 'a multiplicity of behavioral observations' can assist 'and triangulate interpretations' (p. 731).

As reviewed earlier, feedback studies have started to use multiple measures including tests of explicit and implicit knowledge (e.g., Ellis et al., 2006; Erlam & Loewen, 2010; Révész, 2012; Nassaji, 2017). Such research should continue. Other worthwhile attempts would also be including tests of not only accuracy but also fluency and complexity in feedback research. It may be considered that with a focus on accuracy, corrective feedback may have the potential to disrupt the flow of communication and thus may negatively affect learner fluency during meaning-focused interaction. Little research, however, has assessed the impact of feedback on L2 fluency. There may also be a trade-off between fluency and accuracy, and also accuracy and complexity. Based on the notion that humans have limited cognitive resources and that these resources compete between attention to form and attention to meaning (e.g., Skehan, 2009; VanPatten, 2015), we can assume that the more the learner pays attention to producing a language form accurately, the less attention he or she may have to process the form fluently. It is also suggested that as learners become more fluent, an improvement may be noticed in the complexity of their production (Larsen-Freeman & Strom, 1977). Given these assumptions, it is useful to examine the extent to which corrective feedback with emphasis on accuracy would affect learner fluency and complexity.

As for the acquisitional effect of feedback, it is quite clear the question cannot be satisfactorily answered without valid and reliable measures. As we saw earlier, there are currently several validation studies to examine the construct validity of various explicit and implicit knowledge tests (Ellis, 2005a; Ellis & Loewen, 2007; Godfroid et al., 2015; Spada et al., 2015; Kim & Nam, 2017). There is clearly a need for more studies in this area. Also as Prutting, Gallagher, and Mulac (1975) pointed out, a true test of acquisition is a test that is representative of learners' ability in real-life situations. Thus, there is a need to systematically examine and show how well the tests claiming to measure acquisition meet this criterion. There is also a need for studies to examine the different factors that may affect what a test measures. In a recent study, for example, Kim and Nam (2017) offered evidence that OEITs and GJTs provided stronger measures of implicit knowledge when the OEIT was performed under time pressure and when the timed GJT used aural modality.

Finally, the various challenges associated with reliably assessing SLA underscore the significance of replication studies, including those on corrective feedback. The need for replication has recently been emphasized in several publications in the field of applied linguistics (see Porte, 2012, for a detailed discussion). There are many reasons why replication is needed, including confirming the outcomes of previous research, adding new perspectives on the results of the original studies, or examining the generalizability of findings in new contexts. Another reason could be the inadequacy of assessment measures. In this respect, replication informs us of the relevance of the outcome measures used and also increases our understanding of how accurately they assess the intended outcomes across different settings and learners.

Questions arising

1. This article began by examining various measures currently used to assess the effectiveness of incidental feedback in descriptive research (e.g., uptake, repair, modified output, individualized posttests). Several issues were discussed with these measures. How can these issues be resolved, and how can we better explore the relationship between incidental feedback and L2 learning? What other means or assessment tools could be used to examine more accurately how learners engage with and learn from incidental feedback?
2. I have discussed a number of challenges with defining and measuring explicit and implicit knowledge. What are other challenges in addition to those discussed, and how can they be overcome?

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

3. It has been proposed that when declarative knowledge becomes automatized through practice, this automatized knowledge can be explicit and can involve conscious awareness. Is it possible to develop assessment tools that can reliably differentiate automatized explicit knowledge from unconscious implicit knowledge? As reviewed, some current research suggests the use of real-time comprehension measures as tools to assess implicit knowledge. But can these tools prevent learners from using their awareness of linguistic features when performing the tests?
4. If we agree that even automatized knowledge that occurs in the classroom as a result of practice is explicit and that implicit knowledge occurs more in naturalistic SLA contexts, what is the relevance of the distinction between explicit and implicit knowledge for classroom SLA learning?
5. Studies have recently used various forms of retrospective and introspective methods to examine L2 processing. As discussed, the validity of these methods has been questioned. How can these validity issues be addressed? What other measures could be used to examine more accurately the internal processes involved in learning from corrective feedback?
6. Currently, the use of eye-tracking methods and other technological innovations has gained considerable attention in studying L2 processing in several L2 areas such as reading, vocabulary learning, and syntactic processing. Few studies have employed such techniques in corrective feedback research, particularly oral feedback research. How can future research on oral feedback make more use of such technological advances to assess feedback processing?
7. Implicit knowledge can be considered to develop as a result of either extensive exposure to meaningful language use or extensive practice. Given this, should we expect implicit knowledge to be developed as a result of corrective feedback during the short period of an L2 study? If not, what is the significance of research examining the effects of corrective feedback on the development of implicit knowledge, and if such short-term research shows positive effects on implicit knowledge, how could these effects be interpreted?
8. As discussed, experimental pretest-posttest studies have focused mainly on intensive feedback provided on single errors. Is it possible to conduct pretest-posttest research to examine the effectiveness of extensive feedback that occurs on a wide variety of forms? What are some of the methodological challenges to do such research and how can these challenges be overcome?
9. The effectiveness of any corrective feedback move could be due to not a single but multiple factors. How important is it to control all the relevant factors when conducting corrective feedback research? What are the implications of this multiplicity of factors for the development of reliable and valid assessment tools?
10. As noted earlier, we cannot claim that a learner has acquired a language form successfully unless he or she is able to use it both accurately and appropriately in a variety of contexts. It was suggested that one way of dealing with this could be using multiple measures. However, there is the problem of feasibility and the challenges of preparing well-validated tests. How, if at all, can SLA research address such problems?

References

Adams, R., Nuevo, A., & Egi, T. (2011). Explicit and implicit feedback, modified output, and SLA: Does explicit and implicit feedback promote learning and learner–learner interactions? *Modern Language Journal*, *95*, 42–63.

Ajabshir, Z. (2014). The effect of implicit and explicit types of feedback on learners' pragmatic development. *Procedia-Social and Behavioral Sciences*, *98*, 463–471.

Allwright, R. (1984). Why don't learners learn what teachers teach? The interaction hypothesis. In D. Singleton & D. Little (Eds.), *Language learning in formal and informal contexts* (pp. 3–18). Dublin, Ireland: IRAL.

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

Anderson, J. C. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–406.

Anderson, J. C. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Bao, M., Egi, T., & Han, Y. (2011). Classroom study on noticing and recast features: Capturing learner noticing with uptake and stimulated recall. *System*, *39*, 215–228.

Bardovi-Harlig, K., & Vellenga, H. (2012). The effect of instruction on conventional expressions in L2 pragmatics. *System*, *40*, 77–89.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Basturkmen, H., Loewen, S., & Ellis, R. (2004). Teachers' stated beliefs about incidental focus on form and their classroom practices. *Applied Linguistics*, *25*, 243–272.
Beebe, L., & Cummings, M. C. (1996). Natural speech act data versus written questionnaire data: How data collection method affects speech act performance. In S. Gass & J. Neu (Eds.), *Speech acts across cultures* (pp. 65–86). Berlin, Germany: Mouton.
Bialystok, E. (1982). On the relationship between knowing and using forms. *Applied Linguistics*, *3*, 181–206.
Bialystok, E. (1991). Achieving proficiency in a second language: A processing description. In R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith & M. Swain (Eds.), *Foreign/second language pedagogy research* (pp. 63–78). Bristol, UK: Multilingual Matters.
Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. Berlin, Germany: Springer.
Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 245–261). Hillsdale, NJ: Lawrence Erlbaum.
Bowles, M. (2010). *The think aloud controversy in second language research*. New York: Routledge.
Bowles, M. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition*, *33*, 247–271.
Bowles, M., & Leow, R. (2005). Reactivity and type of verbal report in SLA research methodology: Expanding the scope of investigation. *Studies in Second Language Acquisition*, *27*, 415–440.
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.
Carpenter, H., Jeon, K., MacGregor, D., & Mackey, A. (2006). Learners' interpretations of recasts. *Studies in Second Language Acquisition*, *28*, 209–236.
Chaudron, C. (1977). A descriptive model of discourse in the corrective treatment of learners' errors. *Language Learning*, *27*, 29–46.
Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, *32*, 453–467.
DeKeyser, R. (2005). Implicit and explicit learning. In E. Hinkel (Ed.), *Handbook on research in second language teaching and learning* (pp. 314–348). Mahwah, NJ: Lawrence Erlbaum Associates.
Doughty, C. (2003). Instructed SLA: Constraints, compensation, and enhancement. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 256–310). Oxford, UK: Blackwell.
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, UK: Cambridge University Press.
Doughty, C., & Williams, J. (1998). *Focus on form in classroom second language acquisition*. Cambridge, UK: Cambridge University Press.
Egi, T. (2007). Interpreting recasts as linguistic evidence: The role of linguistic target, length, and degree of change. *Studies in Second Language Acquisition*, *29*, 511–537.
Egi, T. (2010). Uptake, modified output, and learner perceptions of recasts: Learner responses as language awareness. *Modern Language Journal*, *94*, 1–21.
Ellis, R. (1991). Grammaticality judgments and learner variability. In H. Burmeister & P. Rounds (Eds.), *Variability in second language acquisition: Proceedings of the tenth meeting of the second language acquisition forum* (pp. 25–60). Eugene: University of Oregon, Department of Linguistics.
Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, *51*, 1–46.
Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, UK: Oxford University Press.
Ellis, R. (2004). The definition and measurement of L2 explicit knowledge. *Language Learning*, *54*, 227–275.
Ellis, R. (2005a). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.
Ellis, R. (2005b). Principles of instructed language learning. *System*, *33*, 209–224.
Ellis, R. (2006a). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied Linguistics*, *27*, 431–463.
Ellis, R. (2006b). Researching the effects of form-focussed instruction on L2 acquisition. *AILA Review*, *19*, 18–41.
Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 339–360). Oxford, UK: Oxford University Press.
Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford, UK: Oxford University Press.
Ellis, R. (2016a). Focus on form: A critical review. *Language Teaching Research*, *20*, 405–428.
Ellis, R. (2016b). Focus on form: A critical review. *Language Teaching Research*, *20*, 405–428.
Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, UK: Oxford University Press.
Ellis, R., Basturkmen, H., & Loewen, S. (2001). Learner uptake in communicative ESL lessons. *Language Learning*, *51*, 281–318.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Ellis, R., & Loewen, S. (2007). Confirming the operational definitions of explicit and implicit knowledge in Ellis (2005): Responding to Isemonger. *Studies in Second Language Acquisition*, *29*, 119–126.
Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (Eds.) (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol, UK: 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*, 339–369.
Ellis, R., & Sheen, Y. (2006). Reexamining the role of recasts in second language acquisition. *Studies in Second Language Acquisition*, *28*, 575–600.
Ericsson, A. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness Studies*, *10*, 1–18.
Ericsson, A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
Erlam, R. (2009). 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. *Canadian Modern Language Review*, *66*, 877–905.
Flynn, S. (1988). Nature of development in L2 acquisition and implications for theories of language acquisition in general. In S. Flynn & W. O'Neill (Eds.), *Linguistic theory in second language acquisition* (pp. 277–294). Dordrecht, Netherlands: Kluwer.
Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research*, *21*, 175–198.
Fu, T., & Nassaji, H. (2016). Corrective feedback, learner uptake, and feedback perception in a Chinese as a foreign language classroom. *Studies in Second Language Learning and Teaching*, *6*, 161–183.
Fukuya, Y., & Zhang, Y. (2002). Effects of recasts of EFL learners' acquisition of pragmalinguistic conventions of request. *Second Language Studies*, *21*, 1–47.
Fukuya, Y., & Zhang, Y. (2006). The effects of recasting on the production of pragmalinguistic conventions of request by Chinese learners of English. *Issues in Applied Linguistics*, *15*, 59–91.
Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Erlbaum.
Gass, S. (2003). Input and interaction. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 224–255). Oxford, UK: Blackwell.
Gass, S., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research*. New York, NY: Routledge.
Gass, S., & Varonis, E. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*, 283–302.
Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., … Yoon, H. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition*, *21*, 563–584.
Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*, 483–517.
Godfroid, A., Loewen, S., Jung, S., Park, J., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. *Studies in Second Language Acquisition*, *37*, 269–297.
Goo, J., & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, *35*, 127–165.
Gurzynski-Weiss, L., & Baralt, M. (2015). Does type of modified output correspond to learner noticing of feedback? A closer look in face-to-face and computer-mediated task-based interaction. *Applied Psycholinguistics*, *36*, 1393–1420.
Gutiérrez, X. (2013). The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition*, *35*, 423–449.
Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*, *20*, 86–97.
Han, Z. (2002). A study of the impact of recasts on tense consistency in L2 output. *TESOL Quarterly*, *36*, 543–572.
Hawkes, L., & Nassaji, H. (2016). The role of extensive recasts in error detection and correction by adult ESL students. *Studies in Second Language Learning and Teaching*, *6*, 19–41.
Hedgcock, J. (1993). Well-formed vs. ill-formed strings in L2 metalingual tasks: Specifying features of grammaticality judgements. *Second Language Research*, *9*, 1–21.
Hinkel, E. (1997). Appropriateness of advice: DCT and multiple choice data1. *Applied Linguistics*, *18*, 1–26.
Hulstijn, J. (2002). Towards a unified account of the representation, processng and acquisition of second language knowledge. *Second Language Research*, *18*, 193–223.
Ishida, M. (2004). Effects of recasts on the acquisition of the aspectual form -tei-(ru) by learners of Japanese as a foreign language. *Language Learning*, *54*, 311–394.
Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, *25*, 603–634.
Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, *57*, 1–33.
Junqueira, L., & Kim, Y. (2013). Exploring the relationship between training, beliefs, and teachers' corrective feedback practices: A case study of a novice and an experienced ESL teacher. *Canadian Modern Language Review*, *69*, 181–206.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Kamiya, N. (2015). The effectiveness of intensive and extensive recasts on L2 acquisition for implicit and explicit knowledge. *Linguistics and Education*, *29*, 59–72.
Kang, H. (2014). Understanding online reading through the eyes of first and second language readers: An exploratory study. *Computers & Education*, *73*, 1–8.
Kasper, G., & Blum-Kulka, S. (2003). *Interlanguage pragmatics*. Oxford, UK: Oxford University Press.
Kentridge, R., Heywood, C., & Weiskrantz, L. (1999). Attention without awareness in blindsight. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*, 1805–1811.
Kim, J., & Nam, H. (2017). Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition*, *39*, 431–457.
Kim, J., & Nassaji, H. (2018). Incidental focus on form and the role of learner extraversion. *Language Teaching Research*, *22*, 698–718.
Koch, C., & Tsuchiya, N. (2007). Attention and consciousness: Two distinct brain processes. *Trends in Cognitive Sciences*, *11*, 16–22.
Koike, D., & Pearson, L. (2005). The effect of instruction and feedback in the development of pragmatic competence. *System*, *33*, 481–501.
Krashen, S. (1985). *The input hypothesis: Issues and implications*. Oxford, UK: Pergamon Press.
Lamme, V. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences*, *7*, 12–18.
Lapkin, S., Swain, M., & Smith, M. (2002). Reformulation and the learning of French pronominal verbs in a Canadian French immersion context. *Modern Language Journal*, *86*, 485–507.
Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development 1. *Language Learning*, *27*, 123–134.
Leow, R. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge.
Leow, R., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, *30*, 111–127.
Leow, R., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, *26*, 35–57.
Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta analysis. *Language Learning*, *60*, 309–365.
Loewen, S. (2002). *The occurrence and effectiveness of incidental focus on form in meaning-based ESL lessons* (Unpublished doctoral dissertation). University of Auckland, Auckland, New Zealand.
Loewen, S. (2005). Incidental focus on form and second language learning. *Studies in Second Language Acquisition*, *27*, 361–386.
Loewen, S., & Erlam, R. (2006). Corrective feedback in the chatroom: An experimental study. *Computer Assisted Language Learning*, *19*, 1–14.
Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). Second language learners' beliefs about grammar instruction and error correction. *Modern Language Journal*, *93*, 91–104.
Loewen, S., & Nabei, T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A series of empirical studies* (pp. 361–378). Oxford, UK: Oxford University Press.
Loewen, S., & Philp, J. (2006). Recasts in adults English L2 classrooms: Characteristics, explicitness, and effectiveness. *Modern Language Journal*, *90*, 536–556.
Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.
Long, M., & Robinson, P. (1998). Focus on form: Theory, research and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom language acquisition* (pp. 15–41). Cambridge, UK: Cambridge University Press.
Loschky, L., & Bley-Vroman, R. (1993). Grammar and task based methodology. In G. Crookes & S. M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 123–167). Bristol, UK: Multilingual Matters.
Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *48*, 183–218.
Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition*, *26*, 399–432.
Lyster, R., & Mori, H. (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition*, *28*, 269–300.
Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition*, *19*, 37–66.
Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*, 265–302.
Lyster, R., Saito, K., & Sato, M. (2012). Oral corrective feedback in second language classrooms. *Language Teaching*, *46*, 1–40.
Mackey, A. (1999). Input, interaction and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557–587.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27*, 405–430.
Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.
Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford, UK: Oxford University Press.
Mackey, A., & Oliver, R. (2002). Interactional feedback and children's L2 development. *System*, *30*, 459–477.
Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *Modern Language Journal*, *82*, 338–356.
Martínez-Flor, A., & Usó-Juan, E. (2010). *Pragmatics and speech act performance*. Amsterdam, Netherlands: John Benjamins.
McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on ESL question development. *Studies in Second Language Acquisition*, *27*, 79–103.
McDonough, K., Crowther, D., Kielstra, P., & Trofimovich, P. (2015). Exploring the potential relationship between eye gaze and English L2 speakers' responses to recasts. *Second Language Research*, *31*, 563–575.
McDonough, K., & Mackey, A. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning*, *56*, 693–720.
McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, *11*, 113–128.
Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks: What they measure and how they relate to each other. In E. Tarone, S. Gass & A. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 227–243). Hove, UK: Lawrence Erlbaum.
Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, *50*, 617–673.
Nabei, T., & Swain, M. (2002). Learner awareness of recasts in classroom interaction: A case study of an adult EFL student's second language learning. *Language Awareness*, *11*, 43–63.
Nassaji, H. (2007a). Elicitation and reformulation and their relationship with learner repair in dyadic interaction. *Language Learning*, *57*, 511–548.
Nassaji, H. (2007b). Focus on form through recasts in dyadic student-teacher interaction: A case for recast enhancement. In C. Gascoigne (Ed.), *Assessing the impact of input enhancement in second language education* (pp. 53–69). Stillwater, OK: New Forums Press.
Nassaji, H. (2009). Effects of recasts and elicitations in dyadic interaction and the role of feedback explicitness. *Language Learning*, *59*, 411–452.
Nassaji, H. (2010). The occurrence and effectiveness of spontaneous focus on form in adult ESL classrooms. *Canadian Modern Language Review*, *66*, 907–933.
Nassaji, H. (2011). Immediate learner repair and its relationship with learning targeted forms in dyadic interaction. *System*, *39*, 17–29.
Nassaji, H. (2015). *Interactional feedback dimension in instructed second language learning*. London, UK: Bloomsbury Publishing.
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.
Nassaji, H. (2017). The effectiveness of extensive versus intensive recasts for learning L2 grammar. *Modern Language Journal*, *101*, 353–368.
Nassaji, H., & Fotos, S. (2010). *Teaching grammar in second language classrooms: Integrating form-focused instruction in communicative context*. London, UK: Routledge.
Nassaji, H., & Kartchava, E. (2017). *Corrective feedback in second language teaching and learning: Research, theory, applications, implications*. New York, NY: Routledge.
Nguyen, M., Do, H., Pham, T., & Nguyen, A. (2017). The effectiveness of corrective feedback for the acquisition of L2 pragmatics: An eight month investigation. *International Review of Applied Linguistics in Language Teaching*, *56*, 345–375.
Nicholas, H., Lightbown, P., & Spada, N. (2001). Recasts as feedback to language learners. *Language Learning*, *51*, 719–758.
Nipaspong, P., & Chinokul, S. (2010). The role of prompts and explicit feedback in raising EFL learners' pragmatic awareness. *University of Sydney Papers in TESOL*, *5*, 101–146.
Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.
Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 716–761). Oxford, UK: Blackwell.
Oliver, R. (2002). The patterns of negotiation for meaning in child interactions. *Modern Language Journal*, *86*, 97–111.
Panova, I., & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly*, *36*, 573–595.
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.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Pica, T. (1987). Second-language acquisition, social interaction, and the classroom. *Applied Linguistics*, *8*, 3–21.
Pica, T. (2002). Subject-matter content: How does it assist the interactional and linguistic needs of classroom language learners? *Modern Language Journal*, *86*, 1–19.
Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, *21*, 737–758.
Porte, G. (2012). *Replication research in applied linguistics*. Cambridge, UK: Cambridge University Press.
Posner, M. (1994). Attention: The mechanisms of consciousness. *Proceedings of the National Academy of Sciences*, *91*, 7398–7403.
Prutting, C., Gallagher, T., & Mulac, A. (1975). The expressive portion of the nsst compared to a spontaneous language sample. *Journal of Speech and Hearing Disorders*, *40*, 40–48.
Reber, A. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford, UK: Clarendon Press.
Rebuschat, P., & Mackey, A. (2013). Prompted production. In C. Chappelle (Ed.), *The encyclopedia of applied linguistics* (Vol. *5*). Oxford, UK: Wiley-Blackwell.
Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*, 93–132.
Révész, A., & Han, Z. (2006). Task content familiarity, task type and efficacy of recasts. *Language Awareness*, *15*, 160–179.
Robinson, P. (2003). Attention and memory during SLA. In C. Doughty, & M. Long (Eds.), *The handbook of second language acquisition* (pp. 631–678). Malden, MA: Blackwell Pub.
Roehr, K. (2007). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics*, *29*, 173–199.
Rosa, E., & O'Neill, M. (1999). Explicitness, intake, and the issue of awareness. *Studies in Second Language Acquisition*, *21*, 511–556.
Rose, K., & Kasper, G. (2001). *Pragmatics in language teaching*. New York, NY: Cambridge University Press.
Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*, 63–77.
Russell, J., & Spada, N. (2006). The effectiveness of corrective feedback for second language acquisition: A meta-analysis of the research. In J. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 131–164). Amsterdam, Netherlands: Benjamins.
Saito, K., & Lyster, R. (2012). Investigating the pedagogical potential of recasts for L2 vowel acquisition. *TESOL Quarterly*, *46*, 387–398.
Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.
Schwartz, B. (1993). On explicit and negative data effecting and affecting competence and linguistic behavior. *Studies in Second Language Acquisition*, *15*, 147–163.
Sharwood Smith, M. (1981). Consciousness-raising and second language acquisition theory. *Applied Linguistics*, *2*, 159–168.
Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research*, *10*, 361–392.
Sheen, Y. (2008). Recasts, language anxiety, modified output, and L2 learning. *Language Learning*, *58*, 835–874.
Shintani, N. (2012). Input-based tasks and the acquisition of vocabulary and grammar: A process-product study. *Language Teaching Research*, *16*, 253–279.
Shintani, N., & Ellis, R. (2013). The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, *22*, 286–306.
Skehan, P. (1998). Task-based instruction. *Annual Review of Applied Linguistics*, *18*, 268–286.
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.
Slimani, A. (1989). The role of topicalization in classroom language learning. *System*, *17*, 223–234.
Slimani, A. (1992). Evaluation of classroom interaction. In C. Anderson, & A. Beretta (Eds.), *Evaluating second language education* (pp. 197–221). Cambridge, UK: Cambridge University Press.
Slobin, D., & Welsh, C. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson, & D. Slobin (Eds.), *Studies of child language development* (pp. 485–489). New York, NY: Holt, Rinehart and Winston.
Sorace, A. (1985). Metalinguistic knowledge and language use in acquisition-poor environments *. *Applied Linguistics*, *6*, 239–254.
Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W. Ritchie, & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375–409). San Diego, CA: Academic Press.
Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*, 1497–1524.
Spada, N., Shiu, J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge. *Language Learning*, *65*, 723–751.
Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261.
Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895.
Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*, 747–790.

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press

Swain, M., & Lapkin, S. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research*, *37*, 285–304.
Takimoto, M. (2006). The effects of explicit feedback and form–meaning processing on the development of pragmatic proficiency in consciousness-raising tasks. *System*, *34*, 601–614.
Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics*, *4*, 142–164.
Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning*, *35*, 373–403.
Thomas, K. (2018). Comparing explicit exemplar-based and rule-based corrective feedback: Introducing analogy-based corrective feedback. *Modern Language Journal*, *102*, 371–391.
Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Bristol, UK: MultilingualMatters.
Tremblay, A. (2005). Theoretical and methodological perspectives on the use of grammatically judgement tasks in linguistic theory. *University of Hawai'I Second Langauge Studies*, *24*, 129–167.
Vafaee, P., Suzuki, Y., & Kachisnke, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95.
VanPatten, B. (2015). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, *53*, 91–109.
VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*, 225–244.
Velmans, M. (1996). *The science of consciousness: Psychological, neuropsychological, and clinical reviews*. London: Routledge.
Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics*, *12*, 54–73.
Williams, J. (1999). *Learner-generated* attention to form. *Language Learning*, *49*, 583–625.
Williams, J. (2001). The effectiveness of spontaneous attention to form. *System*, *29*, 325–340.
Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook on research in second language teaching and learning* (pp. 673–691). Mahwah, NJ: Lawrence Erlbaum Associates.
Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319–353). Bingley, UK: Emerald.
Winke, P., Godfroid, A., & Gass, S. (2013). Introduction to the special issue: Eye-movement recordings in second language research. *Studies in Second Language Acquisition*, *35*, 205–212.
Wu, S., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*, 680–704.
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.

Hossein Nassaji is Professor of Applied Linguistics in the Department of Linguistics, University of Victoria, Canada, and is currently the co-editor of *Language Teaching Research*. His recent books are *Corrective feedback in second language teaching and learning* (2017, Routledge, with Eva Kartchava) and *Interactional feedback dimension in instructed second language learning* (2015, Bloomsbury).

Cite this article: Nassaji, H. (2020). Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges. *Language Teaching*, *53*(1), 3–28. https://doi.org/10.1017/S0261444819000375

https://doi.org/10.1017/S0261444819000375 Published online by Cambridge University Press