2022

Can Language Learners Hear Their Own Errors? The Identification of Grammaticality in One's Own Production

James Hunter

System, Vol. 111, No. 102933 (2022): 1-38. https://doi.org/10.1016/j.system.2022.102933.

Can Language Learners Hear Their Own Errors? The Identification of Grammaticality in One's Own Production

James Hunter
Gonzaga University

Abstract

This exploratory study investigated whether learners can correctly identify the grammaticality of items drawn from corrective feedback (CF) on their own oral production or on that of their peers. It was hypothesized that participants would judge less well-established items more slowly, and conversely that entrenched items, whether target-like or not, would be judged more quickly. 20 learners at two proficiency levels judged audio recordings of themselves reformulating errors they had made in small-group conversations. Items had been categorized according to reformulation accuracy and fluency, and the analysis investigated whether judgment accuracy and speed mirrored these categories. Results indicate clear parallels in reformulation and judgment accuracy, but a weak relationship between fluency of production and recognition. The categorization of errors occurring in both production and recognition, perhaps representing "attempts" at meaning-making (Edge, 1989; Willis, 2003), is proposed as the focus of future pedagogical research investigation. To this end, a pedagogical application of the self-judgment methodology is described.

Keywords

Complexity, Accuracy, Fluency, Individual differences, Error analysis, Computer assisted language learning (CALL), Delayed corrective feedback

INTRODUCTION

Suppose an intermediate learner is presented with a number of her own utterances, produced when she was a beginner, half of them judged (by an expert, e.g., the teacher) to be grammatical and the other half ungrammatical. If the now intermediate learner can correctly identify which items are grammatical and which are not, she is demonstrating progress: Her implicit and/or explicit knowledge has now developed to the point where she can accurately judge the grammaticality of her earlier production in spite of the errors it manifested. The same learner, presented with a selection of items from her *current* output, again half grammatical and half not, might well be expected to judge more of them grammatical, especially if they were communicatively effective. Hawkes and Nassaji (2016) found that intermediate learners could identify errors in their own speech on both a stimulated correction task and a written task with 65% and 83% accuracy, respectively. However, they also found that learners tended to identify errors where there were none. The issue of error detection, then, has relevance both pedagogically, in developing learners' perception of and confidence in their production, and theoretically, in developing our understanding of individual differences in language acquisition.

However, the conceptual issues behind grammaticality or acceptability judgments on one's own second-language production are inchoate, since this is an area that has received scant attention in SLA research. The

literature on the use of grammaticality judgments in SLA research offers very few examples of learners judging their own errors (see Chaudron, 1983, pp. 358–361; Plonsky et al., 2019 for overviews), and even fewer in which learners' own oral production has been used. In the few studies of this kind that have been attempted, results have been mixed. White (1977) found that her participants were able to identify approximately 60% of their target-language (TL) ungrammatical written production, with advanced learners showing no superior ability over intermediate learners. Cohen and Robbins (1976) took a qualitative approach, investigating learners' reactions to and explanations of their own written errors, but did not look at the extent to which learners were successful in correcting their errors. Gass (1983, p. 281) found that intermediate learners correctly judged the accuracy of items taken from their written production at a rate of 71%, and advanced learners at 68%, with intermediate learners slightly better at identifying the grammatical items than the ungrammatical. She also found that intermediate learners slightly overestimated their accuracy, judging a total of 53% of their items as grammatical, and advanced learners slightly underestimated it, judging a total of 48% of their items as grammatical. (For each group, 48% of the items were grammatical "from the perspective of standard English" (p. 279) but both groups judged some of the grammatical items as ungrammatical and vice versa.)

More recently, Hawkes and Nassaji (2016) used the innovative technique of showing learners a series of individualized video clips of themselves in conversational interaction. The clips included correct items and erroneous items with or without interlocutor recasts, with the goal of identifying which type would elicit the greater degree of own-error detection. They found that interlocutor recasts facilitated successful detection of errors on both the stimulated correction test and a written test, with the latter reaching statistical significance. Wouters and John (2020) investigated learners' ability to perceive their own and peers' pronunciation errors in audio recordings of the participants completing a read-aloud task. Their results indicate considerable overall success in detecting pronunciation errors (88%, with no significant difference between detection of own and peers' errors) but considerable variation in accuracy depending on the target phonemes. These studies are noteworthy for their attempt to measure own-error detection in a field of research that on the whole has sidelined both individual differences in language production and learner production itself as an object of study.

DIFFERENTIATING BETWEEN "ERRORS", "MISTAKES", AND "ATTEMPTS"

Pedagogically, teachers require information that will enable them to target items that require focused attention, and conversely to ignore items that seem to need no further remediation. This refers as much to systematic errors no longer made as to "mistakes," a distinction still widely used and largely unchallenged empirically, despite Corder's (1967, p.167) caution:

> Mistakes are of no significance to the process of language learning. However, the problem of determining what is a learner's mistake and what a learner's error is one of some difficulty and involves a much more sophisticated study and analysis of errors than is usually accorded them.

The utterances learners produce in genuine communicative interaction are likely to be an amalgam of acquired forms, partially acquired forms, and guesswork, and a reasonable assumption is that adult L2 learners will rely on the highly automatized forms (even in translation) of their L1 when TL forms are not available. To exemplify, the following is one item from a learner in the present study (see Materials, p. 8) showing the development prompted by the corrective feedback (CF) cycle:

Item 1–38

Original error *(from small-group conversation)*	Feb 15	*I wish I can communicate another country people.

Teacher reformulation *(see Methodology)*	Feb 17	I hope I will be able to communicate with people from other countries.
Timed reformulation attempt 1	Mar 12	*I hope I would be able to communicate…people from another countries.
Timed reformulation attempt 2	May 25	*I hope I will comm…I hope…I wish I comm…I can communicate another country people.

The teacher reformulation[1] indicates four errors: 1) the context requires an expression of future rather than present desires; 2) "communicate" is intransitive and is therefore followed by a PP not a DO; 3) complex modifiers cannot be used attributively in English as in Japanese (this learner's L1), so "from other countries" must be used post-nominally (Payne, 2010); 4) the speaker is not referring to a single country but countries in general. The multiplicity of error types makes this a challenging item, and her second reformulation attempt indicates that the student is confused to the point of giving up. Clearly, this is not a simple case of recognizing a "mistake." However, the lexical choice of hope over wish does seem to have been consistently corrected and might therefore have been a "mistake" originally, and indicates the teacher's CF was effective.

Distinctions such as Edge's (1989, pp. 9–11) between "slips," which learners can self-correct, "errors" (which they cannot), and "attempts," which are "a guess or when neither the intended meaning nor the structure is clear to the teacher" are a useful elaboration on Corder's simple dichotomy, since the evidence presented here is that learners can correctly reformulate items and yet still judge them as ungrammatical, and vice versa. One is reminded of Willis' claim that "it is the learners' attempts to mean that pave the way for learning" (Willis 2003, pp. 110–111) and of Skehan and Foster's (1999, pp. 96–97) characterization of *complexity* as involving "fewer controlled language subsystems" and correlating with "a greater likelihood of restructuring." The challenge for teachers and researchers alike is to identify these attempts.

DYNAMIC ASSESSMENT OF FLUENCY AND ACCURACY

One response to this challenge has been presented by interactionist Dynamic Assessment (DA) approaches in language acquisition research (Lantolf & Poehner, 2004; Lantolf & Poehner, 2011; Poehner & Wang, 2021)[2]. Grounded in the Vygotskian principle that learning occurs through mediation with teachers or more advanced peers (Vygotsky, 1978), DA attempts to evaluate the gap between what a learner can do with and without such mediation. The techniques elaborated (Lantolf and Poehner 2011, p. 20), which progress from highly implicit ("pause [to indicate something is amiss]") to highly explicit ("teacher explains why [the correct answer is correct]") are identical to many that are well known in classroom-based CF research (Lyster, 2004; Lyster & Ranta, 1997), but crucially, DA claims to foreground the process of linguistic development with a view to promoting autonomy, or "self-regulation" (Lantolf and Poehner 2011, p. 17), i.e. self-correction, and not simply the *product* of uptake of the target forms. For this reason, the degree of mediation required is central to understanding the learner's abilities and for promoting development during the assessment process itself, as it highlights the degree of potential unassisted language production.

[1] Reformulations are offered based on the context in which errors are made, so the learner in this instance was not expressing regret (*I wish I could…)* so much as a goal for learning English. This is a representative example of how CF of this kind can introduce greater complexity/range than learners currently command.

[2] I am grateful to an anonymous reviewer for drawing my attention to research in the field of Dynamic Assessment and pushing me to clarify the differences between interactionally-mediated language production and (delayed) corrective feedback.

In the context of the present study, it is precisely this unassisted language production that is of interest. Learners were engaged in regular student-led, small-group discussion sessions in which the teacher did not take part (see Hunter (2012) for a description of the process and rationale). In contrast, both CF and DA research has, with few exceptions, examined feedback provided by teachers during teacher-fronted activities, in which the teacher controls the activity

itself as well as the type and quantity of language produced. For example, the example above (Item 1–38) emerged in the negotiation of meaning between students discussing why they were learning English. It represents how the learner expressed her meaning in the moment, and it was understood by her listeners. Had a teacher intervened at that point to provide mediated feedback, it is difficult to specify with any confidence what level of mediation would have been required, since as we have seen there were multiple errors. What we can predict with more confidence is that the teacher's act of providing feedback would have taken the floor from the speaker and interrupted the flow of communication. This is a central justification for delayed CF: In short, the value of *un*assisted or *un*mediated language production is that students reach further than when a teacher, however gifted, guides the conversation. However, while these "attempts to mean" promote growth in complexity, they nonetheless require CF. Without it, they retain meaning-making capacity but remain non-target-like and have the potential to become proceduralized, and more fluently produced, in the same manner as target-like forms.

FLUENCY IN PRODUCTION AND RECOGNITION

Traditionally, *fluency* is a concept that has been associated with productive skills rather than receptive (Lintunen, Mutta, & Peltonen, 2019), and the relationship between fluency in production and recognition is still poorly understood. Gatbonton and Segalowitz summarize as follows:

> In one sense, [fluency] refers to the speed and ease of handling utterances; the greater the automaticity, the faster the *recognition* and production of grammatically correct and communicatively appropriate utterances. (Gatbonton & Segalowitz, 1988, p. 474; emphasis added)

This is an intriguing notion, and one that goes to the heart of the question of how (second) language is produced and comprehended. The assumption in much SLA research is that fluency is a measure of automaticity, of proceduralized knowledge (McLaughlin, 1987; Towell et al., 1996, Suzuki, 2017), or of "acquisition" (Krashen, 1981, 1982, 1985; Schwartz, 1986). It has been suggested that less fluent production requires the use of explicit knowledge ("learned" knowledge, in Krashen's Monitor Model), in other words the conscious application of rules, but that this process can become automatized and more implicit with time and practice (Suzuki, 2017; Suzuki & DeKeyser, 2015; 2017). For example, recent research (Hui & Godfroid, 2021; Hui, 2020) has investigated the indexing of fluency using the coefficient of variability (CV) proposed by Segalowitz and Segalowitz (1993) in the area of word recognition. The CV values, obtained by dividing the standard deviation of an individual participant's judgment reaction times (RT) by their mean RT, thus "reveal RT variability in Mean RT in the same individual's processing while correcting for their processing speed" (Hui & Godfroid 2021, pp. 1042–3). Hui (2020), for example, investigated intentional and incidental word learning and suggests that CV values can provide an indication of processing automaticity in learners, but only after a requisite level of declarative knowledge has been attained, i.e., when judgment accuracy is approximately 90% (p. 352).

Furthermore, there is growing evidence (Bardovi-Harlig, 2002; Ellis et al., 2008; Eskildsen, 2009; Kormos & Dénes, 2004; Pawley & Syder, 1983; Wray, 2000; 2002; 2018) that formulaic language plays an integral role in the language acquisition process, in that more proficient learners make greater use of formulaic constructions than do novices. Thus the prevailing view among SLA researchers is that linguistic forms, explicitly focused on,

thought about, practiced, and gradually automatized, can contribute to the acquisition of implicit knowledge (DeKeyser, 2007; Gatbonton & Segalowitz, 2005; Suzuki, 2017; Suzuki &DeKeyser, 2017). We may not be able to directly examine the resulting implicit knowledge, but this does not mean it cannot derive from what was once in our awareness (Hulstijn & Schmidt, 1994; McLaughlin, 1990; Schmidt, 1990).

Tapping into learners' awareness of grammaticality is challenging, however. Ellis (1991) and Han (1996, 2000) have pointed out the lack of reliability in learners' grammaticality judgments, and the lack of reliability in the judgments of naïve judges is well documented (Birdsong, 1989; Schütze, 1996; Spinner & Gass, 2019). In the case of second-language learners, it is hardly surprising that there are elements (especially less frequent or salient elements) of the TL about which they have no intuitions, and hence the use of grammaticality judgment tests (GJTs) in second-language research often makes little sense. However, when the goal is to establish the nature and form of the learner's grammatical knowledge, the use of GJTs seems justified. This is not to say that the method is without complications, as discussed below, but these are no more insurmountable than many others in psycholinguistic research.

THE PRESENT STUDY

The present investigation focuses on learners' awareness of the well-formedness of their own and peers' oral production, and was motivated by three questions, which have hitherto not been addressed in SLA research:

1. To what extent can learners identify the grammaticality of their own and peers' spoken language?
2. What is the relationship between accuracy in production and accuracy in judgment of grammaticality of one's own errors?
3. What is the relationship between fluency in production and reaction time (RT) in judgment of grammaticality of one's own errors?

The first of these in particular is a deceptively simple question since the ontological status of a systematic "error" is problematic. If a particular structure of interlanguage is a true representation of the competence or mental grammar of a learner, then logically it cannot be erroneous within that system; (Bley-Vroman, 1983; Corder, 1967). According to Corder, if learners are able to judge their own production as inaccurate, this production must be a non-systematic "mistake" in performance; any other production, whether conforming to TL norms or not (i.e., an "error"), must be "grammatical" in their IL. Another view, proposed here, is that learners do systematically produce non-target-like forms but may be able to recognize that they are not well formed. If that were the case, the relationship between accuracy in production and in judgment of grammaticality should be weak or non-existent. Conversely, if both fluent production and intuitions of grammaticality reflect automatized explicit knowledge or implicit knowledge, there should be a strong association between measures of the two. Thus, RT in judging grammaticality should correlate closely with measures of production fluency.

PARTICIPANTS

Participants for the study came from two intact classes of ESL students in different English for Academic Purposes programs. The first group was an advanced class (approximately IELTS 6.0) in the English Language Center of a small, private university in the Northwest of the USA (N = 11; henceforth "ELC"). The second group was an intermediate class of students from a Japanese women's university in a 14-week Intensive English Program in the Northwest of the USA, where that university has a branch campus (N = 9; henceforth "IEP"). The average TOEIC score for this group was 407 (approximately IELTS 4.0, CEFR B1), with individual scores ranging from 370 to 435. Both groups used small group conversations as the main methodology to develop oral accuracy and fluency as part of their ongoing coursework and were aware of and consented to the use of their data in this research. Two groups were included not for the purposes of comparison but to establish whether the effects

observed for one group would parallel those for the other, and if so to counteract the threats to external validity and generalizability of the ecological research approach adopted (Kramsch and Steffensen 2008). Although intact classes were used for the research, the two groups differed in several important ways in addition to having different classroom teachers, as summarized in Table 1. In essence, the learning context of the IEP group can be thought of as more akin to English as a Foreign Language, in terms of being monolingual, monocultural, and having limited daily exposure to the target language.

Table 1: Summary of differences between ELC and IEP groups

		ELC	IEP
English proficiency level		advanced	intermediate
English study program:			
	Hours per week	18	18
	classes	Listening & Speaking, Writing, Reading, Grammar	Conversation, Writing, Reading, American Studies, Grammar
Diversity:			
	L1	Arabic, Spanish, Korean, Mandarin	Japanese
	Age	18-38	19-20
	Sex	Women & men	Women
	Socio-economic	Mixed (business and academic professionals and clergy from working-class and privileged backgrounds)	Middle-class; second-tier Japanese university
Purpose of study		Undergraduate or graduate study in US	Compulsory part of English BA degree in Japan
Exposure to target language outside of class		Unlimited (housed with American roommates or other L1 speakers; unrestricted access to local community)	Limited (housed with classmates; one-weekend with American family; very restricted access to local community owing to safety concerns)

METHODOLOGY

TEST ITEMS

During the semester, each group was given two elicited reformulation tests (ERTs) as follows: Students received a list of all of their own and a selection of peers' spoken errors that had been transcribed during small-group conversations, on which they had received CF in the form of audio recorded reformulations made available to all via the learning management system, and which they had been given ample time to practice (see Materials, p. 1–5 for example ERT test materials). They were then recorded as they reformulated the erroneous sentences within a two-minute time limit, with students averaging 10.4 reformulations per minute. The recordings were made using the digital language laboratory platform DiLL (https://www.swifteducation.com/).

The ERT is a central component of the delayed CF methodology used in these programs, and its use in this study therefore had a threefold benefit: first, it provided material for the timed grammaticality judgment

test (TGJT) described below; second, it permitted measurement of spoken accuracy and fluency for each student, against which the accuracy and speed of their grammaticality judgments could be compared; third, it ensured a greater degree of face (and ecological) validity than would be possible had non-pedagogical measures been used: the participants knew and understood the purpose of the ERTs and were motivated to complete them as part of their regular coursework, which is not always the case in classroom-based research, let alone in experimental research. Finally, students are typically required to reformulate all of their own errors and a selection of peers' errors chosen by the classroom teacher to provide a shared context for incidental focus on form. This is fortuitous for the present study as it provides the opportunity to compare student performance on the two error sources.

The ERTs resulted in 40 two-minute mp3 files (two for each of the 20 participants), which were processed as follows: each was manually segmented in Audacity (Audacity Team 2018) into individual mp3 files for each item, with care being taken to trim leading and trailing silence, and was scored for accuracy and fluency by the researcher. For accuracy, any version that conveyed the intended meaning in standard English was considered grammatical. The accuracy of each item was compared to the score originally given by the classroom teacher, and reliability of the accuracy scores was found to be very high (□ = .96), indicating a strong consensus between the raters. Any items that were controversial were then eliminated from the subsequent analyses, resulting in the removal of 18 items of the 1070 items (1.68%). For fluency, the number of words per minute (WPM) for each item was used, after it was established that WPM correlates highly with phonation time ratio ($r = .682$, $p = .021$) and speech rate ($r = .758$, $p = .007$), two of the temporal fluency features that correlate most highly with fluency ratings given by human judges (e.g. Chambers, 1997; Kormos & Dénes, 2004; see Ellis & Barkhuizen, 2005, pp. 139–164 for an overview).

In the absence of a clear theoretical basis for categorization or these items as "errors," "mistakes," or "attempts," and so on, a simple, two-factor classification was used to categorize the reformulated items:

A. Fast and correct (FC): items that were correctly reformulated at above mean WPM (operationalized here as ≥1 SD above the mean for the individual), indicating a language item that is acquired, familiar, and easily reformulated. These items most likely correspond to "mistakes" made in the original small-group conversations, which are quickly and accurately reformulated on the ERT.
B. Fast and incorrect (FI): items that were incorrectly reformulated at ≥1 SD above the mean for the individual, suggesting acquired forms that are erroneous but entrenched (MacWhinney, 2018; Pulvermuller 2002) or possibly new "mistakes" introduced during the ERT itself
C. Moderate and correct (MC): items that were correctly reformulated at average fluency (<1 SD above or below the mean for the individual)
D. Moderate and incorrect (MI): items that were incorrectly reformulated at average fluency (<1 SD above or below the mean for the individual)
E. Slow and correct (SC): indicating an item that is not yet proceduralized and automatic, and that therefore needs to be considered more carefully, possibly with the assistance of explicit metalinguistic knowledge, L1 translation, analogy with acquired forms, and so on (≥1 SD below the mean for the individual)
F. Slow and incorrect (SI): indicating an item that is either completely unfamiliar or not familiar enough to benefit from the assistance of explicit metalinguistic knowledge, L1 translation, analogy with acquired forms, and so on (≥1 SD below the mean for the individual)

ADMINISTRATION OF THE TIMED GRAMMATICALITY JUDGEMENT TEST (TGJT)

Students in both groups were invited to take part in the TGJT phase of the research but it was emphasized that participation was on a strictly voluntary basis, and the TGJT was administered outside of class time. Eleven

students from the ELC and nine from the IEP group chose to take part, reducing the pool of items from 1052 to 530. The distribution of participants' own errors to peer errors on the ERTs for the TGJT participants (Table 2) indicates that for both groups, over half of the items reformulated were peer errors. Participants were told that they would hear items that they had recorded during the ERTs and that they simply had to decide, as fast as possible, whether the item was correct or incorrect. Before taking the TGJT, participants were asked to familiarize themselves with the procedure using a set of practice items (see Materials, p. 6–7).

Table 2: Count (percentage) of own and peers' errors for each group

	ELC	**IEP**
Own	163 (48%)	74 (39%)
Peers'	175 (52%)	118 (61%)

The reformulations elicited in the ERTs for each participant formed the set of test items for the same participant on the TGJT but were presented in a randomized order. No attempt was made to balance the number of incorrect and correct items, or to screen the items for specific structures or lexis. Each group of participants took the TGJT approximately two weeks after doing the ERTs, so that enough time would have elapsed for them not to remember exactly what they had recorded as reformulations, but not so much time that their ILs could have changed or stabilized (Reinders, 2005). During the TGJT, participants wore headphones, to maximize sound quality and isolation, and had no other materials in front of them. The tests took place in a language lab on iMac computers and were administered using an online interface. Each audio file was played to the participant once, and immediately after the audio finished playing, two buttons appeared on the screen, allowing for a choice of Correct or Incorrect. As soon as either button was clicked, the RT was displayed on the screen and sent to a MySQL database, along with the participant's name, her judgment, the date and time, and the ID of the item sound file. In the ERT stage, all items had already been assigned grammatical or ungrammatical status according to whether they conformed to TL norms. The TGJT judgment for an item was thus considered grammatical if it agreed with the corresponding ERT score and ungrammatical if it did not.

DATA ANALYSIS

All data were entered into SAS JMP 15, a statistical analysis software package (SAS Institute Inc., Cary, NC, 1989–2021) to provide descriptive and inferential statistical analysis of individual and group performance. Bivariate correlations were calculated for measures of accuracy and fluency on the TGJT and the ERT for each group. A comparison of means test was used to compare performance between groups, and within groups on own and peer items. For each participant, z-scores were calculated for RT and WPM for each item to permit comparison of performance within each task and between the two tasks (see Materials, p. 8–17). To facilitate the interpretation of data, the standardized scores for RT were inverted since greater RT values should be associated with less automaticity, and vice versa, whereas with WPM the converse is true.

FINDINGS

IDENTIFICATION OF GRAMMATICAL STATUS

Overall, participants accurately identified the grammatical status of the ERT items, with a global success rate of 80%. Their ability to accurately judge correct items (94%) was much greater than their accuracy with incorrect items (26% – Table 3). Accuracy scores ranged from 60–100% for the ELC group (M = 79%) and 59–89% for the IEP group (M = 77%).

Table 3: Count of items from ERT and TGJT by grammatical status, with percentage of accurate judgments in parentheses

	TGJT			
ERT		Incorrect	Correct	Total
	Incorrect	92	33 (26%)	125
	Correct	25 (5%)	380 (94%)	405
	Total	117	413 (80%)	530

This suggests that overall, the CF provided in the teachers' reformulations was effective in leading to production and recognition of target forms. It also suggests that for many participants, some incorrect items sounded correct and vice versa, with at least one participant performing barely above chance. A t-test comparison of means was performed to determine inter-group differences (Table 4) and while there is no significant difference in accuracy (again, it must be stressed that the linguistic proficiency of the two groups is not being compared here, only their ability to judge the grammaticality of their production), the ELC group was on average almost a second faster in their judgments than the IEP group. Possible reasons for this difference are addressed in the discussion section, but it is encouraging, from a CF perspective, to note that while more proficient speakers may be faster at judging grammaticality, the groups are largely equal in their ability to do so.

Table 4: T-test of TGJT measures for ELC and IEP groups

	ELC Mean (SD)	IEP Mean (SD)	Difference	df	t	sig.
Accuracy	0.79 (0.41)	0.77 (0.42)	0.02	492	.52	.602
RT	1364 (1698)	2334 (2150)	970	492	5.56	≤ .0001

RESPONSE BIAS

Three ELC participants and one IEP participant judged all their items as grammatical. In the absence of corroborating information, as would have been provided by a think-aloud review of the items such as was done in Ellis (1991), it is unwise to generalize from these findings, but the possibility that participants were simply being uncooperative by choosing only one response option (or randomly choosing either) was considered. It is likely that any participant who did so would have been ignoring accuracy to answer as fast (or slowly) as possible, which would likely be reflected in a fast (or slow) mean RT and small SD. Two of the participants (M = 709 ms, SD= 414; M = 910 ms, SD = 686) could possibly qualify, but another participant with similar data (M = 735 ms, SD = 483) used both options; and two others had M > 2000 and SD > 1500, which would indicate considered responses. It is thus more likely that all four simply thought that all the items sounded correct.

ELC GROUP

To investigate the degree of response bias, a Chi-square test was performed using grammatical status of the items as the expected frequencies, and it was found that overall, participants were indeed more likely overall to judge items grammatical (X^2 (1, N = 338) = 30.38, $p < .0001$) than ungrammatical.

IEP GROUP

Similarly, it was found that the less proficient IEP participants were more likely overall to judge items as grammatical (X^2 (1, N = 192) = 13.79, $p < .0001$) than ungrammatical. However, this response bias was not as pronounced as that found for the more proficient ELC group, which runs counter to the intuitive wisdom that with greater proficiency comes a greater ability to detect grammaticality. This point will be taken up in the discussion below.

DIFFERENTIAL REACTION TIME WHEN JUDGING ITEMS AS GRAMMATICAL AND UNGRAMMATICAL

ELC GROUP

To determine whether there is a measurable difference in RT when participants judge an item to be ungrammatical, a t-test of the ELC group's responses was performed. This showed a significant mean difference in RT when they judged an item as grammatical (M = 1248, SD = 1515) in contrast to ungrammatical (M = 2852, SD = 2897), t(1) = 2.57, p = .002. Thus, the group as a whole made their judgments approximately twice as fast when they thought items were grammatical than when they did not.

IEP GROUP

A t-test of the IEP group's responses also showed a significant mean difference in RT when they judged an item as grammatical (M = 2121, SD = 1976) in contrast to ungrammatical (M = 3384, SD = 2654), t(1) = 2.55, p = .003. Here again, the group on average made judgments approximately 1.5 times faster when they thought an item was grammatical. The findings from the two groups suggest that when an item "sounds right" (an intuitive or gestalt judgment), learners are faster in their judgments than when it "sounds wrong," in which case an analytical judgment must be performed, possibly using explicit knowledge, resulting in a slower judgment.

COMPARISON OF JUDGEMENTS OF OWN AND PEERS' ERRORS

ELC GROUP

For the ELC group, 52% of the items on the ERTs were errors originally made by peers. To determine if TGJT performance was affected by the source of the item, a t-test was performed on the data, which revealed no significant differences in accuracy or speed of judgments between their own and peers' errors (Table 5).

Table 5: t-test comparison of TGJT measures for own and peers' errors, ELC group

	Own Errors Mean (SD)	Peers' errors Mean (SD)	Difference	df	t	sig.
Accuracy	0.80 (0.04)	0.77 (0.03)	0.02	336	.46	.648
RT	1570 (1955)	1565 (1989)	-5.6	336	.003	.979

IEP GROUP

As with the ELC group, a comparison of means for judgments of participants' own items and those of peers revealed no significant differences in either accuracy or RT of judgments between their own and peers' errors (Table 6).

Table 6: t-test comparison of TGJT measures for own and peers' errors, IEP group

	Own Errors Mean (SD)	Peers' errors Mean (SD)	Difference	df	t	sig.
Accuracy	0.82 (0.038)	0.73 (0.45)	0.09	190	1.52	.130
RT	1931 (1351)	2172 (2078)	241	190	.976	.33

The finding that participants are no slower or less accurate with their own errors than with those of peers suggest that the practice of assigning peers' items for CF for pedagogical reasons is not in any way detrimental for the group as a whole, regardless of group composition (multilingual or monolingual) or proficiency.

THE RELATIONSHIP BETWEEN FLUENCY AND ACCURACY IN PRODUCTION AND RECOGNITION

To determine whether there were any significant effects for the six categories of errors (FC, FI, MC, MI, SC, SI) apparent in the TGJT data, a 2 x 4 ANOVA was performed with RT and accuracy as dependent variables.

Significant effects were found for both accuracy ($F(3, 526) = 91.7$, $p < 0.0001$, partial $\eta 2 = 0.47$) and RT ($F(3, 526) = 5.79$, $p < 0.0001$, partial $\eta 2 = 0.05$). In contrast, no significant interactions between these variables and group (ELC, IEP) were found, suggesting that the effects are not dependent on proficiency. The Standard Least Square Means for each category (Table 7) indicate that all incorrect ERT items (FI, MI, SI) were least accurately judged, with FI items being judged most rapidly, and SI items least rapidly but also highly inaccurately. The relative size of the Standard Error for the FI category, however, suggests that strong conclusions should not be drawn at this stage. For example, items in the FI category were judged faster than any others (M = 1308), which might suggest acquired, entrenched erroneous forms, as described above. This is supported by the accuracy score of 33%.

Table 7: TGJT Least Squares Means and Standard Error for RT and Accuracy, by ERT reformulation category

	N	Accuracy		RT (ms)	
		Least Sq Mean	Std Error	Least Sq Mean (ms)	Std Error
Fast correct (FC)	79	.94	0.033	1455	213
Fast incorrect (FI)	12	0.33	0.087	1308	546
Moderate correct (MC)	262	0.94	0.027	1557	170
Moderate incorrect (MI)	90	0.25	0.050	2471	319
Slow correct (SC)	64	0.94	0.021	1534	132
Slow incorrect (SI)	23	0.26	0.034	2680	213

Pairwise comparisons for RT (Table 8) show significant differences between FI and MI items, but not between FI and other categories. It must be remembered that the ERT items were all generated after participants had been given a reformulation by the teacher and had the opportunity to practice, so it is not surprising that there were relatively few "Fast incorrect" items.

Table 8: Pairwise comparisons and effect sizes for RT and Accuracy, by ERT reformulation category

	Accuracy (max=1)					RT (ms)			
Level	Diff.	*t*	*p*	*d*	Level	Diff.	*t*	*p*	*d*
SI ≤ MC	0.69	10.46	≤.0001	2.28	SI ≤ MC	827	2.01	0.04	0.44
SI ≤ FC	0.68	9.46	≤.0001	2.24	SI ≤ FC	926	2.06	.039	0.49
SI ≤ SC	0.64	8.81	≤.0001	2.14	SI ≤ SC			*ns*	
MI ≤ MC	0.69	18.76	≤.0001	2.29	MI ≤ MC	1121	4.85	≤.0001	0.60
MI ≤ FC	0.68	14.65	≤.0001	2.26	MI ≤ FC	1220	4.19	≤.0001	0.65
MI ≤ SC	0.65	13.20	≤.0001	2.16	MI ≤ SC	1181	3.82	≤.0001	0.63
MI ≤ FI			*ns*		MI ≤ FI	1367	2.36	.02	0.72
FI ≤ MC	0.61	6.89	≤.0001	2.03	FI ≤ MC			*ns*	
FI ≤ FC	0.60	6.46	≤.0001	2.00	FI ≤ FC			*ns*	
FI ≤ SC	0.57	6.04	≤.0001	1.90	FI ≤ SC			*ns*	

In contrast, the "Moderate incorrect" items were more slowly judged than the other categories, with medium effect sizes, as were the "Slow incorrect" items. Thus, judgment speed does to some extent reflect reformulation fluency: Incorrect items are likely to be judged more slowly than correct ones. A more conclusive relationship was found in the accuracy with which the incorrect categories were judged, since all incorrect reformulation categories were considerably less accurately judged than the correct ones, with large effect sizes. The fact that both measures distinguish the incorrect reformulations is noteworthy. On the one hand, it means that these participants either took longer to judge their incorrect reformulations or were less accurate in doing so, or both. This is a disappointing finding from their perspective, as it essentially means that they cannot hear these errors. On the other hand, this effect indicates that the categorization schema is substantiated by the

judgment data. This is pedagogically useful information, as it gives a clearer picture of the kinds of errors that need greater attention than an immediate recast in a face-to-face context can provide. Furthermore, the lack of clear RT-based identification of the FI and FC categories suggest that very few of the items correspond to "mistakes" in Corder's terms. In other words, the delayed CF has highlighted a category of errors that are not "systematic," in the sense of automatized, but are pedagogically meaningful and, in terms of providing a basis for greater complexity, useful errors that need further attention.

RELATIONSHIP BETWEEN MEASURES OF ACCURACY AND FLUENCY

ELC GROUP

The correlation between accuracy on the ERT and TGJT was found to be very high, at r = .767, p < .0001 (Table 9). Some of the variation could be attributable to the response bias shown by the group. As discussed above, some participants did not judge any of their items as incorrect. Assuming that they were not simply being uncooperative, the conclusion must be that every one of their items *sounded* correct to them. This would decrease the strength of the relationship between the accuracy measurements on the two tests, as would cases in which participants judged an item as ungrammatical when it was in fact grammatical.

Table 9: Correlation matrix for measures of fluency and accuracy on ERT and TGJT, ELC group

	1	2	3	4
1. TGJT Reaction Time (RT)	--			
2. TGJT Accuracy	-.027 *.629*	--		
3. ERT WPM	-.123 *.030*	.171 *.003*	--	
4. ERT Accuracy	-.092 *.103*	.767 *.000*	.133 *.018*	--

Note: significance shown in italics

IEP GROUP

For the IEP group, the association between accuracy measures on the ERT and TGJT was weaker than for the ELC group, but still strong at r = .593, p < .0001 (Table 10). Again, the conclusion is that at least some students recognized errors in their own production. Further support for the interrelation of the constructs accuracy and fluency is hinted at by the weak but significant correlation between RT and ERT accuracy (r = -.222): the more accurate their reformulations (by TL norms), the faster their judgments. To confirm this hypothesis, the bivariate correlation between RT and judgment (whether the participant thought an item was grammatical) was calculated. The result (r = -.313, p < .0001) is suggestive of the interplay between these variables and another, or possibly others, such as general reaction speed or confidence. Further research will be necessary to determine the precise nature of this interaction.

Table 10: Correlation matrix for measures of fluency and accuracy on ERT and TGJT, IEP group

	1	2	3	4
1. TGJT Reaction Time (RT)	--			
2. TGJT Accuracy	-.125 *.085*	--		
3. ERT WPM	-.072 *.318*	.116 *.108*	--	
4. ERT Accuracy	-.222 *.002*	.593 *.000*	.170 *.018*	--

The more surprising discovery in the correlation data (Table 9), however, was the very weak relationship between RT and measures of spoken fluency. More fluent production did correlate with faster RT ($r = -.123$) but so weakly as to suggest that fluency and RT are only tangentially related. In fact, only the 79 "Fast correct" items showed a significant correlation between ERT fluency and TGJT reaction time ($r = .459$, $p < .0001$). These items, representing 15% of the total, could qualify as "mistakes," having been both reformulated and judged fluently and accurately.

DISCUSSION

The evidence presented in this investigation has been used to address three questions. In answer to the first, *Can learners identify the grammatical status of their own and peers' spoken language?* it was found that both groups on average could accurately discern grammaticality approximately 80% of the time, but that both groups performed much more poorly with ungrammatical items than grammatical ones: They could generally not hear their own errors. The fact that the ELC group was faster is possibly explained by the proficiency difference between the two groups: the ELC group was more proficient and had greater exposure to the TL, and despite greater complexity of production, their average reaction time suggests that they are able to rely more on their implicit knowledge than the IEP group. We would expect this to be the case, but there are other possible explanations for the average RT difference. A likely candidate is general listening proficiency: the ELC group consisted largely of Arabic and Spanish L1 learners whose educational experiences and learning styles generally prioritize auditory learning (e.g. Reid, 1987, p. 96, Table 3). However, in this TGJT, the participants are listening to themselves, and both the form and content of the utterances are known beforehand, drawn as they are from CF given on the students' own production, which should have offset at least some of the effects of differential listening proficiency.

Moreover, it was found that both groups were biased towards judging their production as grammatical, confirming that many of their errors sound correct to them. Only 26% of the ungrammatical items were correctly judged to be ungrammatical, which is understandable given the fact that all ERT items represented language on which the participants had already received CF and so presumably thought they had reformulated correctly. It was hypothesized that items that were originally "mistakes," easily corrected on the ERT to become "Fast correct" items in this categorization, would result faster RTs in judgment. Some support for this comes from the accuracy with which these items were judged and the moderate correlation between reformulation fluency and judgment speed. A "slip" in the original small group discussions would likely be fluently and accurately reformulated on the ERT and judged rapidly and correctly, and the TGJT data support that analysis. However, two larger category of errors, "Moderate incorrect" and "Slow incorrect" items, were identifiable on the basis of both RT and accuracy: Participants' judgments on these items were both more inaccurate and slower. A provisional conclusion, therefore, is that the majority of learners find it challenging to recognize ungrammaticality that is neither a "slip" nor a systematic "error," but may be instead what Edge (1989, pp. 9–11) calls an "attempt."

In addressing the second question, *What is the relationship between accuracy in production and accuracy in judgment of grammaticality?*, correlations between ERT production and TGJT recognition for both groups on the accuracy measure were strong. Overall, the participants correctly judged the grammaticality of 79% (ELC) and 77% (IEP) of the items, indicating that learners at these two proficiency levels are likely to be inaccurate in their judgments approximately 20% of the time. It is precisely this inaccuracy that is of pedagogical interest, since these are the forms that most likely represent "attempts" in need of further remediation. Furthermore, ERT items categorized as incorrect, whether fast or slow, were generally also incorrectly judged. Again, this argues for more systematic CF than face-to-face options can provide. Wouters and John (2020) point out the majority

of participants in their study reported that they only realized they were mispronouncing the target phonemes when they did the perception task, for which the analog in this study would be the TGJT.

The final research question, *What is the relationship between fluency in production and reaction time in judgment?* remains difficult to answer with any degree of confidence. The investigation did not reveal a strong, linear relationship between fluent production and the speed of judgments. This supports findings such as those of Coppieters (1987) and Birdsong (1989, p. 61), who notes that "metalinguistic performance reflects idiosyncratic skill parameters, which vary across task and across individuals." Instead, the tentative conclusion is that reaction time in the judgment of an item depends more on whether a participant judges it to be grammatical or not than on factors of fluency in production. If the item "sounds right," the RT will be much faster, by an average of 1.5 seconds. The analysis behind this finding is based on comparatively little data, since the majority of TGJT items happened to be (TL) grammatical, and in addition the response bias found in both groups resulted in a great imbalance between judgments as grammatical and as ungrammatical. Additional research, in which a better balance of grammatical and ungrammatical items could be planned (permitting an increased number of judgments as ungrammatical) would enable a more robust investigation of the origin of L2 metalinguistic intuitions. In terms of CF, the implications of this investigation contribute to our overall understanding of the various choices before teachers. One clear finding is that learners' metalinguistic judgments of their own errors are no more accurate or faster than their judgments of peer errors, provided they have also received CF on these, which supports the conclusion that the practice of using peer error as a source of CF input is not harmful, as some have suggested (Truscott, 1999).

Another reliable finding is that there is no simple way to distinguish between "errors" and "mistakes" based on judgment speed. It is possible that the original errors from the small-group conversations are in fact mostly systematic errors and not slips, since there is a degree of discernment employed by teachers at the moment of data collection (Hunter, 2012). However, the response bias revealed that the less proficient IEP group was better able to recognize their own inaccuracies, which might lead to the conclusion that they make more "slips," a phenomenon documented by Poulisse:

> The large number of L1-based slips in beginners' L2 speech can be explained [by the fact that] L1 procedures are largely automatized, while L2 procedures are not yet…As a result, sometimes the L1 procedures will accidentally take the place of the required L2 procedures. (Poulisse, 2000, p. 145)

Poulisse found that proficiency-related differences emerge at the lexical level (mostly substitutions) followed by the phonological and morphological level (mainly verb forms). Hence the IEP group may have produced more "L1-based slips" (such as *another country people* in Item 1–38 above) and were more able to recognize them, but not necessarily correct them. The ELC group produced fewer and recognized fewer, and thus the remaining ungrammaticality can be assumed to be more systematic. Therefore, the hypothesis that all learners should be able to recognize only correctible ungrammaticality (Corder's "mistakes") must be refined to incorporate L2 proficiency. In any case, it is difficult to support Corder's conclusion that "mistakes are of no significance."

Given the finding of a response bias, it is reasonable to conclude that many errors do "sound right" to learners, which in itself is a strong argument for CF of this kind: in the absence of systematic CF, these errors are likely to continue to "sound right" since the learner's own (and peers') production of them is likely to be more frequent, or at least more salient, than the target forms in the input (see also Butler Platt & MacWhinney, 1982, p. 412 for a description of a similar mechanism in child L1 acquisition). This might seem counter-intuitive, but the fact is that a learner's total *grammatical* linguistic repertoire in the target language is a minute subset of that language. Her *ungrammatical* repertoire is not a subset of the TL but is similarly minuscule by comparison. Therefore, *any* aspect of the learner's repertoire is statistically far more likely to occur in her own production, not to mention

thoughts or private/inner speech (De Guerrero, 1994; 2018), than in the TL input. This fact underscores the problems posed by Pawley and Syder (1983), that is, the unidiomatic production of language learners and the subsequent impact it has on fluency.

As concerns judgments of one's own production, where this has been considered at all in the literature, it has generally been assumed that learners will judge their own production as grammatical, as discussed above; however, this is far too simplistic a picture since it ignores the issue of "attempts," and it assumes that accurate production and accurate recognition are motivated by the same cognitive mechanisms, which the current study, like Gass' (1983, p. 280), does not support. Gass' conclusion is that learner intuitions of grammaticality become more analytical as proficiency increases:

> Sentences "felt" wrong to the students without their having an accurate idea of why they were wrong. It is suggested here that part of what is involved in becoming more proficient in a second language is the progression from more gestalt-like to analytical analyses. We might further speculate that indeed the analyzed aspect is a necessary precondition for fluency in an L2, more so than for an L1. (Gass, 1983, p. 285)

Wray (2000; 2008; 2018), however, argues that "gestalt-like" analyses derive from implicit knowledge of frequency of occurrence in input, on which L1 speakers can rely but L2 cannot, and this knowledge can be seen as the backdrop against which the unusual (i.e. ungrammatical or borderline) will stand out. In contrast, "analytical analyses" derive from explicit knowledge of "rules," which L2 speakers may rely on to a far greater extent than L1 but which take longer to process. With the addendum that the learner's own output, whether TL grammatical or not, becomes input for gestalt/implicit knowledge, Wray's position is one whose explanatory power vis-à-vis error production in SLA is greater than any other so far elaborated. If anything, then, what is involved in becoming more proficient in a second language may be the opposite of what Gass suggests: the progression from analytical analyses to more gestalt-like analyses, or at least a balance between the two. In the terminology of this study, this would mean first drawing students' attention to any incorrect items, particularly "Fast incorrect" ones, and second encouraging them to target the "Moderate" and "Slow" items with a view to developing greater accuracy and fluency in both production and recognition.

PEDAGOGICAL APPLICATION

The foregoing discussion implies a rationale not only for continued investigation of learners' ability to judge grammaticality as a basis for understanding language development but for the inclusion of grammaticality judgment as part of a CF methodology. At the very least, the awareness that ungrammatical items sound correct (and vice versa) could draw learners' attention (Schmidt, 1990; Schmidt, 2012; Wouters and John, 2020) to how closely their language conforms to TL norms. As Schmidt (2012, p. 11) writes: "There remains more than sufficient reason to hypothesize that individual differences in the degree to which learners pay attention to and notice grammatical features of the input may partly account for their relative success in [grammatical accuracy in] language learning." The following, therefore, describes an attempt to integrate grammaticality judgments into a delayed CF methodology, allowing for individual agency and choice but using individual learner response data to guide that choice. In brief, the platform[3] saves a learner's recorded reformulations of her (or peers') errors, whether they are correct or not, as part of the delayed CF cycle. These can then be presented to the learner in a TGJT. The learner's response data are used to prioritize individual items for practice according to four criteria, weighted as shown in parentheses:

[3] The platform, which is still under beta testing, can be seen at www.comsem.net.

a. Number of attempts previously made (10%)
b. Number of days since last practiced (20%)
c. Average reaction time for previous correct attempts (20%)
d. Percentage of previous attempts that were correct (50%)

Normalized scores (z scores) are generated for these four variables (using the inverse of b and c, as higher figures are less desirable) and multiplied by the criterion weight to generate a figure out of 100, relative to every other item. Items with lower numbers appear in green, indicating that generally the learner can correctly identify the grammaticality of the item, has done so recently, and so on. Conversely, items that need attention are coded red (shown in Figure 1 as darker shading and shorter bars). The learner selects items to practice, and the items are presented in random order in either text or audio form according to the learner's choice. As soon as the item appears (or after the audio stops playing), a timer is displayed, counting up in tenths of a second, to remind the learner of the time pressure (Figure 2). If no judgment has been made after ten seconds, or if the learner clicks "Skip", the attempt is recorded as incorrect.

Figure 1: Interface for selection of items for individualized TGJT

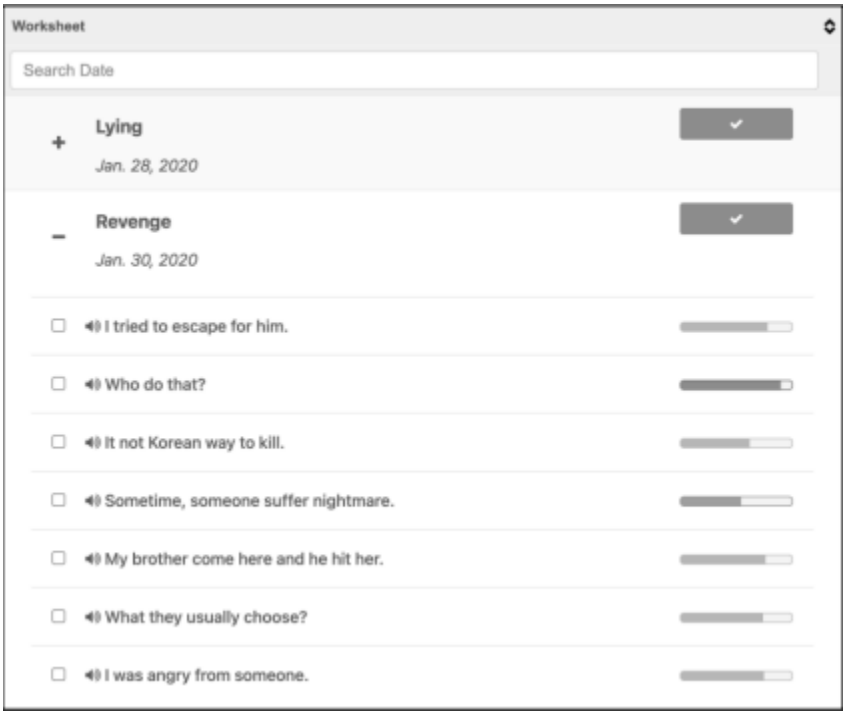
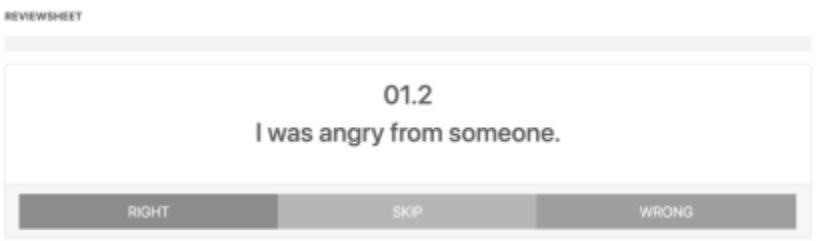


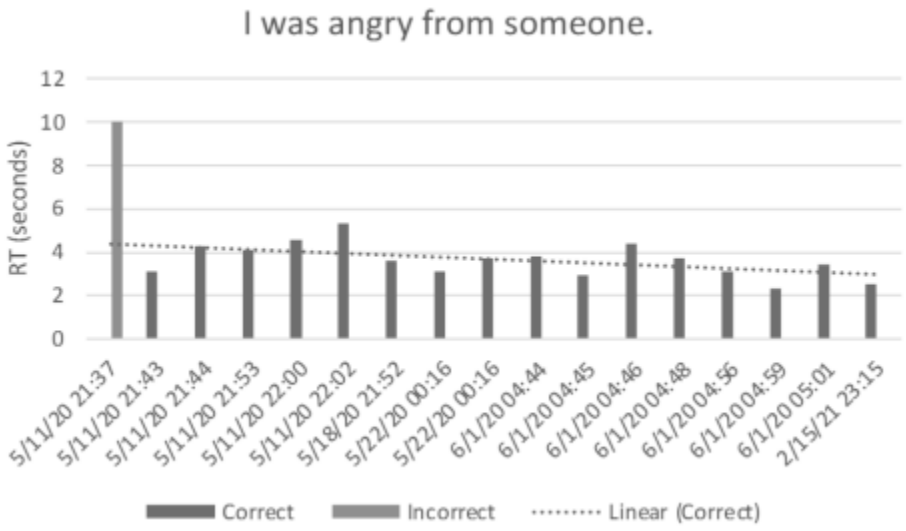
Figure 2: An item in the timed grammaticality judgment test



To date, insufficient data have been collected to permit robust analysis but as an example, data from one learner (Figure 3) indicate several trends. First, she tends to practice items multiple times both in the same session and over multiple sessions; second, a common pattern is that she judges some items incorrectly at first and then gets consistently faster over time; third, some items continue to be inaccurately judged, especially

after an interval of a few weeks, suggesting that they continue to sound correct (or incorrect) to her. As more data are collected, it may become possible to explore whether this type of practice can play a facilitating role in the acquisition of new language forms.

Figure 3: One learner's TGJT response data for one item



CONCLUSION

In this exploratory investigation, the relationship between accuracy, complexity, and fluency was investigated from the perspective of learners' recognition of grammaticality in their own production; it is one of very few studies of learner intuitions that uses the learner's own production as data, and the only such study to date that uses audio recordings of the learners themselves. The present investigation thus represents a response to Corder's challenge to develop a more sophisticated method of studying and analyzing learner errors. Given the difference in proficiency between the groups in this study, the complexity of the language—and errors—that each group produces were also different, and a CF methodology that did not account for these differences would be seriously flawed. Put another way, interlanguage development is a moving target, and CF must accommodate not only differences between groups but also within them. It follows therefore that the CF methodology should have approximately similar effects irrespective of learner proficiency, as was found here in the accuracy of error recognition by learners in both groups.

While this study has presented evidence in support of a systematic CF approach that integrates recognition fluency by individual learners, it is not without limitations. First, the small number of participants meant a paucity of "Fast incorrect" items, a category of particular interest in CF research. Second, participants took the TGJT only once, which meant that any development in automaticity on the individual ERT/TGJT items, such as might be measured by the coefficient of variability (Hui & Godfroid, 2021), was not possible. As more data are collected through the platform described above, it should be feasible to investigate proceduralization of individual items. Third, the pressure to judge quickly, reinforced by the displaying of RT on the screen after each judgment, may have affected TGJT performance by causing stress or by encouraging some participants to sacrifice accuracy for speed. Finally, "production fluency" in this investigation was operationalized in terms of fluency on a timed reformulation test. While this approach has intuitive appeal (and face validity from the point of view of holding learners accountable in the CF process), it cannot be said to equate to spontaneous production of correct forms in communication, nor does it give any indication of whether learners can generalize from specific items to larger systems. These are both areas that merit further study.

Whether this line of investigation ultimately turns out to be fruitful will depend on refinements in the methodology such as those proposed. However, the creation of a database of learner errors together with accompanying audio recordings of learner reformulations will almost certainly prove valuable for a range of SLA investigations. The pedagogical utility of such a database is that it will permit ongoing refinements of the CF methodology by permitting students to compare their current and former production. Further research comparing similar populations at different proficiency levels is needed to determine whether faster and more accurate judgments are a function of increases in general proficiency or a direct result of the CF. If the former is the case, this type of TGJT, which is inexpensive and easy to administer, might be expanded in SLA research and extended to assessment purposes. If the latter is true, it provides a strong theoretical and empirical basis for the systematic, delayed corrective feedback approach presented here.

REFERENCES

1. Bardovi-Harlig, K. (2002). A new starting point? Investigating formulaic use and input in future expression. Studies in Second Language Acquisition, 24, 189–198, doi: 10.1017/S0272263102002036.
2. Birdsong, D. (1989). Metalinguistic performance and interlinguistic competence (Springer series in language and communication; 25). Springer-Verlag.
3. Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. Language Learning, 33, 1–17, doi: 10.1111/j.1467-1770.1983.tb00983.x.
4. Butler Platt, C., & MacWhinney, B. (1982). Error assimilation as a mechanism in language learning. Journal of Child Language, 10, 401–414, doi: 10.1017/S0305000900007844.
5. Chambers, F. (1997). What do we mean by fluency? System, 25(4), 535–544, doi: 10.1016/S0346-251X(97)00046-8.
6. Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. Language Learning, 33(3), 343–377, doi: 10.1111/j.1467–1770.1983.tb00546.x.
7. Cohen, A., & Robbins, M. (1976). Toward assessing interlanguage performance: The relationship between selected errors, learners' characteristics, and learners' explanations. Language Learning, 26(1), 45–66, doi: 10.1111/j.1467-1770.1976.tb00259.x.
8. Coppieters, R. (1987). Competence differences between native and near-native speakers. Language, 63(3), 544–573, doi: 10.2307/415005.
9. Corder, S. P. (1967). The significance of learners' errors. IRAL, 5(4), 161–170, doi: 10.1515/iral.1967.5.1-4.161.
10. Custers, R., & Aarts, H. (2007). In search of the nonconscious sources of goal pursuit: Accessibility and positive affective valence of the goal state. Journal of Experimental Social Psychology, 43(2), 312–318, doi: 10.1016/j.jesp.2006.02.005.
11. De Guerrero, M. C. M. (2018). Going covert: Inner and private speech in language learning. Language Teaching, 51(1), 1–35, doi: 10.1017/S0261444817000295.
12. De Guerrero, M. C. M. (1994). Form and function of inner speech in adult second language learning. In J. P. Lantolf & G. Appel (Eds.), Vygotskian approaches to second language research (pp. 83-116). Norwood, NJ: Ablex Publishing Company.
13. DeKeyser, R. (2007). Practice in a second language: Perspectives from applied linguistics and cognitive psychology (The Cambridge applied linguistics series). Cambridge University Press.
14. Edge, J. (1989). Mistakes and correction. Longman.
15. Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. TESOL Quarterly, 42(3), 375–396, doi: 10.1002/j.1545-7249.2008.tb00137.x.
16. Ellis, R. (1991). Grammaticality judgments and second language acquisition. Studies in Second Language Acquisition, 13, 161–186, doi: 10.1017/S0272263100009931.
17. Ellis, R., & Barkhuizen, G. P. (2005). Analysing learner language (Oxford applied linguistics). Oxford University Press.

18. Eskildsen, S. W. (2009). Constructing another language: Usage-based linguistics in second language acquisition. Applied Linguistics, 30(3), 335–357, doi: 10.1093/applin/amn037.
19. Gass, S. M. (1983). The development of L2 intuitions. TESOL Quarterly, 17(2), 273–291, doi: 10.2307/3586654.
20. Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: a focus on access to fluency. The Canadian Modern Language Review, 61(3), 325–353, doi: 10.3138/cmlr.61.3.325.
21. Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. TESOL Quarterly, 22(3), 473–492, doi: 10.2307/3587290.
22. Han, Y. (1996). L2 learners' explicit knowledge of verb complement structures and its relationships to L2 implicit knowledge. PhD Thesis, Temple University.
23. Han, Y. (2000). Grammaticality judgment tests: How reliable and valid are they? Applied Language Learning, 11(1), 177–204.
24. Hui, B. (2020). Processing variability in intentional and incidental word learning. Studies in Second Language Acquisition, 42(2), 327–357, doi: 10.1017/S0272263119000603.
25. Hui, B., & Godfroid, A. (2021). Testing the role of processing speed and automaticity in second language listening. Applied Psycholinguistics, 42(5), 1089–1115, doi: 10.1017/S0142716420000193.
26. Hulstijn, J., & Schmidt, R. (1994). Consciousness in second language learning. AILA Review, 5– 112.
27. Hunter, J. (2012). 'Small Talk': Developing fluency, accuracy, and complexity in speaking. ELT Journal, 66(1), 30–41, doi: 10.1093/elt/ccq093.
28. Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. System, 32, 145–164, doi: 10.1016/j.pragma.2007.02.011.
29. Krashen, S. D. (1981). Second language acquisition and second language learning. Pergamon Press.
30. Krashen, S. D. (1982). Principles and practice in second language acquisition. Pergamon Press.
31. Krashen, S. D. (1985). The Input Hypothesis: Issues and Implications. Longman ELT.
32. Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. Journal of applied linguistics, 1. doi: 10.1558/japl.1.1.49.55872.
33. Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. Language teaching research, 15(1), 11-33, doi: 10.1177/1362168810383328.
34. Lintunen, P., Mutta, M., & Peltonen, P. (2019). Fluency in L2 Learning and Use. Multilingual Matters.
35. Lyster, R. (2004). Research on form-focused instruction in immersion classrooms: Implications for theory and practice. French Language Studies, 14, 321-341. doi: 10.3138/cmlr.55.4.457
36. Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. Studies in Second Language Acquisition, 19, 37-66. doi: 10.1017/S0272263197001034.
37. MacWhinney, B. (2018). A unified model of first and second language learning. Sources of variation in first language acquisition: Languages, contexts, and learners. Amsterdam: John Benjamins, 287–312, doi: 10.2167/illt047.0.
38. McLaughlin, B. (1987). Theories of second language learning. Edward Arnold.
39. McLaughlin, B. (1990). "Conscious" versus "unconscious" learning. TESOL Quarterly, 24(4), 617–634, doi: 10.1093/applin/11.2.113.
40. Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. Language and communication, 19(1), 191-226.
41. Payne, T. E. (2010). Complementation: Noun complements vs. post-nominal modifiers. In Understanding English Grammar: A Linguistic Introduction (pp. 205–227). Cambridge University Press.
42. Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2019). A methodological synthesis and meta-analysis of judgment tasks in second language research. Second Language Research, 0267658319828413, doi: 10.1177/0267658319828413.
43. Poulisse, N. (2000). Slips of the tongue in first and second language production. Studia Linguistica, 54(2), 136–149, doi: 10.1111/1467-9582.00055.

44. Poehner, M. E., & Wang, Z. (2021). Dynamic assessment and second language development. Language Teaching, 54(4), 472-490, doi: 10.1017/S0261444820000555.
45. Pulvermuller, F. (2002). The neuroscience of language: On brain circuits of words and serial order. Cambridge University Press.
46. Reid, J. M. (1987). The learning style preferences of ESL students. TESOL Quarterly, 21(1), 87– 111, doi: 10.1515/iral.2004.42.4.335.
47. Reinders, H. W. (2005). The effects of different task types on L2 learners' intake and acquisition of two grammatical structures. PhD Thesis, The University of Auckland.
48. Schmidt, R. (1990). The role of consciousness in second language learning. Applied Linguistics, 11(2), 129–158, doi: 10.1093/elt/ccr036.
49. Schmidt, R. (2012). Attention, awareness, and individual differences in language learning. In Perspectives on individual characteristics and foreign language education (pp. 27–50). De Gruyter Mouton.
50. Schütze, C. T. (1996). The empirical base of linguistics: Grammaticality judgments and linguistic methodology. University of Chicago Press.
51. Schwartz, B. D. (1986). The epistemological status of second language acquisition. Second language research, 2(2), 120–159, doi: 10.1177/026765838600200202.
52. Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. Applied Psycholinguistics, 14(3), 369–385, doi: 10.1017/S0142716400010845.
53. Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. Language learning, 49(1), 93–120, doi: 10.1111/1467-9922.00071.
54. Spinner, P., & Gass, S. M. (2019). Using judgments in second language acquisition research. Routledge New York.
55. Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. Applied Psycholinguistics, 38(5), 1229– 1261, doi: 10.1111/lang.12241.
56. Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. Language Learning, 65(4), 860–895, doi: 10.1111/lang.12138.
57. 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(4), 747–790, doi: 10.1111/lang.12138.
58. Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. Applied Linguistics, 17(1), 84–119, doi: 10.1093/applin/17.1.84.
59. Truscott, J. (1999). What's wrong with oral grammar correction. The Canadian Modern Language Review, 55(4), 437-456, doi: 10.1191/026765898674803209.
60. Vygotsky, L. S. (1978). Mind in society: the development of higher psychological processes. Harvard University Press.
61. White, L. (1977). Error analysis and error correction in adult learners of English as a second language. Working Papers in Bilingualism, 13, 42–58, doi: 10.1080/09658410208667060.
62. Willis, D. (2003). Rules, patterns, and words: grammar and lexis in English Language Teaching. Cambridge University Press.
63. Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. Applied Linguistics, 21(4), 463–489, doi: 10.1111/j.1467–1770.1978.tb00313.x.
64. Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. Phraseology in foreign language learning and teaching, 123. doi:10.1093/applin/21.4.463.
65. Wray, A. (2018). Concluding question: Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.) Understanding Formulaic Language (pp. 248–269). New York, NY.: Routledge. doi: 10.4324/9781315206615-1.