研究論文

A Case Study of Complexity, Accuracy, and Fluency in A Short-term Language Program

短期語学プログラムにおける複雑さ・正確さ・流暢さのケーススタディ

Barrie Matte
バリー・マット

コミュニケーションの妥当性や成功を測定するスピーキングに関わる研究には、複雑さ・正確さ・流暢さ（CAF）を観点に測定するものが多いが、このような評価観点に即したスピーキング指導を扱った研究はほとんどない。本稿では、大学生１名を対象に開発した一ヶ月間の事前プログラムと、そのプログラムの効果をCAF各側面で測定した結果を報告する。研究の結果、CAF各側面を強化する言語プログラムを準備することが可能であること、またCAFを意識した指導が学習者にとって有益であることが示唆された。さらに、明示的な訂正フィードバックの使用は、目的言語の構造や特徴の顕在化を促し学習者のメタ言語的知識に寄与することも示唆された。本稿では、指導法や研究法の意義も合わせて論じる。

キーワード
複雑さ、正確さ、流暢さ（CAF）、スピーキング評価、明示的な修正フィードバック、タスクの繰り返し

1. Introduction

This study was the result of a second language（L2）English student seeking advice for preparing to do a study-abroad program in Canada. My interest in corrective feedback lead me to thinking how I could incorporate it into a language program which would take place over the course of a single month. I decided that, because the body of second language acquisition（SLA）research on complexity, accuracy, and fluency（CAF）of L2 spoken language is greatly lacking in studies employing measurable intervention, this would be an opportunity to investigate how these dimensions of language production, with the use of explicit corrective feedback as a treatment, could be developed in a short-term study, hence the present study. To help the student develop the language needed to prepare for his study-abroad program, a speaking task was developed and corrective feedback was provided in each session throughout the study.

2. Literature Review

2.1 Complexity, Accuracy, and Fluency

The dimensions of complexity, accuracy, and fluency are essential when measuring aspects of interaction such as communicative adequacy and communicative success. Norris and Ortega (2009) commented that without measuring L2 CAF, measuring the effects of tasks, teaching, and other stimuli on the development of language competencies is quite difficult (p. 557).

Revesz et al. (2014) studied the extent to which CAF and linguistic proficiency contributed to *communicative success*, i.e., communication where “participants recognize and respond to the expectations of what to say and how to say it, contingent on what other participants do and what the context is” (p. 3). They found that performances that were more lexically diverse and syntactically complex, along with accuracy and fluency, received higher communicative adequacy ratings (p. 15), reaffirming the need to measure complexity as a component of spoken language production. In a similar study, de Jong et al. (2012) found that increases to both lexical and grammatical knowledge, which they termed *linguistic knowledge skills*, predicted greater success towards speaking performance for the advanced participants in their study (p. 27). Their findings suggest that the more learners increase in L2 ability, the more they benefit from lexical input. Therefore, it is necessary for lexical knowledge to be included in the CAF paradigm as lexical knowledge precedes grammatical knowledge.

Accuracy within the CAF paradigm has been measured in numerous ways, such as in Revesz et al. (2014), who used the ratio of errors per 100 words together with specific accuracy measures of grammatical forms (p. 8). Iwashita et al. (2008) found that global accuracy (accurate usage of grammatical forms) was a strong predictor of overall speaking performance (p. 42). Another important finding in their study was that when one aspect of language (e.g., lexis, fluency, or accuracy) was not as good as other aspects, the rating of overall speaking proficiency did not necessarily diminish, rather a combination of aspects was suggested to impact their assessment (p. 43). This relates to the similar findings of Skehan (2009), who proposed a Trade-off Hypothesis amongst the dimensions of CAF where the likelihood of all three paradigms increasing simultaneously is unlikely. Rather, he argued fluency is typically accompanied by either of the other two, not both (p. 512). In Ellis’ (2009) meta-analysis of studies on task type and their effect on CAF, task type was found to have a variety of effects on accuracy. The main findings that Ellis (2009) summarized were that L2 learners with lower proficiency increase in accuracy due to planning time more so than more advanced learners, and possible influences on accuracy depending on task type（pp. 496‒497）.

The final dimension of CAF is fluency, which at the macro level can be defined as general L2 proficiency. Fluency can also be measured at the micro level through breakdown fluency（Baker-Smemoe et al., 2014; Iwashita et al., 2008; Revesz et al., 2014; Skehan, 2009;）, repair fluency（Revesz et al., 2014; Skehan, 2009）, speed（Baker-Smemoe et al., 2014; Iwashita et al., 2008; Revesz et al., 2014; Skehan, 2009）, and length of run（Baker-Smemoe et al., 2014; Skehan, 2009）. Breakdown fluency, which is a measurement of silence and pausing behavior typically measured through the number of silent and filled pauses per 100 words, was found to be the strongest predictor of communicative adequacy in several studies（Iwashita et al., 2008; Revesz et al., 2014）, which suggests that fewer filled pauses indicate higher proficiency to an interlocutor or rater. Baker-Smemoe et al.（2014）reported similar findings, further suggesting that using spoken fluency measures to estimate proficiency might be most effective for learners with more advanced L2 proficiency, as lesser gains amongst lower proficiency L2 learners were found in their study. Revesz et al.（2014）found similar results, where fewer false-starts and self-repairs typically indicated higher proficiency amongst participants. Skehan（2009）proposed the deepening of fluency's definition to include pause boundaries as a measure to assess mean length of run. Native speakers regard pause boundaries as a place for online planning, while L2 learners tend to either extend or shorten boundaries as a result of processing unforeseen lexical choices（p. 514）. Mean length of run（Towell et al., 1996）is calculated as the mean number of syllables produced in utterances between pauses of .28 seconds and above. In fluency tasks, advanced learners of French and English demonstrated that increases in fluency were a result of the period of residence abroad and that the mean length of run was largely attributable to the proceduralization of different kinds of knowledge, including procedural knowledge of syntax and of lexical phrases. Commenting on morpho-syntax development of L2 learners, Polat and Kim（2013）proposed that naturalistic（untutored）learning of an L2 leads to overuse of high-frequency features of a language, such as vocabulary and simple grammatical features, which can lead to misleading fluency measures（pp. 203‒204）. Corrective feedback during output has also been suggested to increase fluency measures in speaking performance, as L2 learners modify their output according to the input they receive（Mackey, 2007; Muranoi, 2007; Stafford et al., 2012）. With regards to these findings, it is thus important to measure all dimensions of CAF, as complexity, accuracy, and fluency are dependent on influences from one another in order to be more accurate.

2.2 Explicit Corrective Feedback

Corrective feedback（CF）has been shown to have a great influence on the development of spoken language, as it can increase the salience of target features of an L2. Muranoi's（2007）examination of production models indicated that any opportunity for output is important to L2 development, however cognitive awareness and noticing must occur as well, which is where explicit CF is necessary（pp. 76–77）. Among the many types of feedback that are provided in interlocution, implicit types（clarification requests, recasts, and repetition）and explicit types（elicitation, provision of metalinguistic clues, and explicit error correction）are found in many studies of language acquisition. Lyster and Saito's（2010）meta-analysis of 15 quantitative CF studies indicated that recasts, prompts, and explicit correction all had significant effects; however, prompts were slightly more effective that recasts（p. 283）. One implication from these findings is that how CF is provided contributes more to noticing than any other factor. It is also necessary to note the importance of salience, as there is no benefit to learners if they do not notice the feedback they are provided. This was suggested in Stafford et al.（2012）, in which receptive gains, but not productive gains, were found to occur.

2.3 Task Repetition

Repetition, rehearsal, and time on task have been proposed as possible methods of improving L2 fluency. Ellis（2009）and de Jong and Perfetti（2011）emphasized the importance of using repetition and rehearsal for increasing results in fluency measures, and similar suggestions came from Du's（2013）study of immersion programs, where time on task is key to gains, especially within the first month of a program（pp. 140–141）. In addition to taking into account how much time is needed to foster development of an L2, the number of times a task needs to be repeated and the appropriate feedback necessary must be considered as demonstrated in Aljaafreh and Lantolf's（1994）study in which corrective feedback was used in private tutoring sessions for a writing task that was repeated weekly during a 2-month language program. While Aljaafreh and Lantolf focused on the use of corrective feedback and negotiation to improve a writing task, Lynch（2007）used a *student-initiated* task in which students were to take responsibility to improve their performance on a speaking task by transcribing their performance verbatim before self- and peer-correcting it, and then finally having a teacher check it（p. 310）. Lynch found that compared to students receiving feedback solely from the teacher, participating in the transcription process seemed to assist in the uptake of the corrective feedback, leading to fewer errors over time（pp. 315–316）. The benefit of repeating a single task appears to narrow the focus of the language, while increasing the amount of time for improving the target language containing errors.

3. The Present Study

3.1 Purpose

The purpose of the present study was to prepare a second-year undergraduate university student for an 11-month study abroad program in Canada, taking place in the academic year after the study was completed. To prepare the student, it was decided that the focus of the sessions should be on improving his spoken language through the dimensions of CAF and a repeated speaking task, to which he would receive explicit corrective feedback targeting errors and misuse of vocabulary, grammar, and oral production (intonation, etc.). Measures of CAF and explicit corrective feedback have been operationalized as follows:

Complexity: The use of a variety of grammatical features and vocabulary

Accuracy: Few errors in grammar and vocabulary, and few self-repairs

Fluency: The length of mean run (how much is spoken between pause boundaries)

Explicit corrective feedback:

1) Targeted explicit feedback on specific errors, such as in misuse of vocabulary or erroneous grammar, followed by a discussion of the appropriate vocabulary or grammar (i.e., raising metalinguistic awareness of the language targets)

2) Targeted explicit feedback on spoken language from the recordings, such as pause boundaries and intonation, followed by a discussion of ways to make the student's speaking sound more natural sounding (i.e., raising metalinguistic awareness of his language production)

3.2 Hypotheses

As this study acknowledges the importance of CAF in L2 spoken language production, in addition to the effects of CF and task repetition, the following hypotheses were made:

1. Provision of explicit corrective feedback on grammatical and lexical features will result in increases across all dimensions of CAF, which will be noticeable to both myself and the student, as well as to raters listening to the recordings.

2. Task repetition will create greater awareness of errors and misuses of grammatical and lexical features within the participant, which will be noticeable in the recordings to both myself and the student, as well as to raters listening to the recordings.

4. Methodology

4.1 Participant

The participant in this study was a 19-year-old male student at a mid-sized university in Western Japan named Rin（pseudonym）. Rin was a second-year international studies major who studied English in class six times each week, but who also practiced using English independently（daily）using various websites, graded readers（extensive reading）, timed reading practice, and through interacting with foreign students and English teachers on campus. He had studied English throughout elementary school, junior high school, and senior high school, and had also studied English for three years at a cram school（*juku* in Japanese）. However, he had never studied English abroad prior to this program. He had taken both the TOEFL and TOEIC tests within the previous year, and had a TOEFL score of 413. Rin indicated a dislike of studying English prior to entering university, but found his non-elective courses to be quite enjoyable and decided to pursue the language more seriously. From the observations I made during this study, Rin appeared to have a very high aptitude for learning English, which together with his determination to improve through independent studying contributed to the ease in which he made proficiency gains over the short period of the study.

4.2 Procedures

Over the course of a month, I met Rin seven times on campus（one hour per session）. With Rin's study abroad program commencing in the following year, the following questions were chosen for him to answer and record in each session（except session 4）for the purpose of developing his L2 CAF when discussing the study abroad program:

1）"Tell me about your study abroad program next year" and

2）"Why do you want to study abroad?"

During the study, Rin was asked to listen and reflect on the recordings, first with a focus on grammatical errors（sequencing errors made in sessions 1 and 2）, then on changes to his intonation and lexical content（fluency issues which were addressed in sessions 3, 4, and 5）, and finally, issues with pause boundaries（addressed in sessions 6 and 7）. The aim of this treatment was to increase his metacognitive awareness of his own language production.

Each recording was recorded in an empty classroom at the university at a 192kbps bit rate in .mp3 format, using a Sony ICD-UX544F recorder. The recordings were checked for volume and clarity, and no problems were found.

4.3 Data Collection

The first session included a description of the purpose and objectives of the study and the signing of an informed consent form by the student, followed by the completion of a background questionnaire. There was no treatment done in this session, however a practice task was recorded of Rin answering the questions about his study abroad plans.

Sessions 2 and 3 began with Rin listening to recordings made at the end of each previous session, which was then followed by a discussion focusing on grammatical errors made in the recordings. Explicit CF was then given to Rin, and he was asked to answer the questions once more at the end of the session, which was recorded.

Sessions 4 to 6 also began with Rin listening to the previous sessions' recordings, however in these sessions he was asked to comment on any changes he noticed in the recordings. I asked questions such as “What are the main differences you hear in the intonation you used to answer the first question?” and “In the recording from session 3 compared to the recording from session 4, do you notice any differences in the way you discussed what you want to do in Canada?” The target in these three sessions was to increase the lexical and grammatical content of Rin's answers (in accordance with the proposed operationalization of complexity being a result of usage of a variety of vocabulary and grammatical features in his answers). From session 4, Rin had changed his plan for studying abroad (he extended the length of his stay in Canada), resulting in an adjustment to his answers to accommodate this change of plans.

In session 4, Rin was asked to write out portions of his answers, and then we revised vocabulary and grammatical errors together (his usage of ordinals), using elicitation and explicit focus on grammatical tenses and word selection. In having Rin rehearse from a set answer, his speaking began to take on a “scripted” sound, which lacked natural-sounding rhythm. Due to time constraints, no recording was made in this session.

During session 5, each of the previous recordings were played, and Rin was asked to comment on changes that occurred that he could notice. We then discussed how his answers became longer and used more complex vocabulary and grammar. To address the issue of Rin's answers sounding unnatural, he was asked to make a bullet-point list of key words from the previous recording, which was then used to help him focus on key words rather than a script. After his first rehearsal, Rin was asked to remove any unnecessary words which resulted in the list being reduced from 57 to 46 words. After a third rehearsal, the list was further reduced to 34. During the final portion of the session, minor lexical and grammatical issues were addressed (such as a failure to distinguish between *be going to* and *will*, and Rin's tendency to drop the final o in *Toronto*), then a recording was made.

Session 6 involved listening to the recording from session 5, followed by a discussion. Rin and I both agreed that while his language had become more complex and accurate, pauses in his speech did not seem natural, and were causing apparent disfluency. A second point discussed was that the recording from session 5 still sounded somewhat scripted. In order to address the first issue, I played the recording and paused it at a point where a natural pause should occur. Next, the recording was played in its original form and Rin was asked to comment on the difference, to which he said that it sounded more natural with a longer pause. He was then asked to rehearse his answer again, with an explicit emphasis on increasing his length of pausing, and the positioning of his pausing. A few suggestions were also given about his intonation of the final syllables within each pause boundary before feedback on his progress was provided. At this point, there was a noticeable difference in how natural Rin's speech sounded, which addressed the second issue. He was then asked to rehearse once more before a recording was made.

At the beginning of the final session, the second, fourth, and fifth recordings (from sessions 3, 5, and 6, respectively) were listened to, then Rin was asked to comment on what had changed with regards to his rate of speech, length of reply, and use of natural pause boundaries. Both of us agreed on most of the changes that had occurred, however a final suggestion was offered with regards to syntax: in order to link two of his sentences, increase his speaking rate, and reduce the number of total sentences, it was suggested that he use the relative pronoun *which*. Finally, after several rehearsals, Rin was asked to make a final recording.

4.4 Measurement

In order to measure Rin's speaking performance, I developed a scale (Appendix A) to assess the recordings made of Rin answering the 2 questions, using the three dimensions of CAF. Complexity was operationalized as a measurement of lexical and grammatical features used in the recordings. The inclusion of lexis and grammar within the paradigm of complexity is similar to the suggestions of Skehan (2009), and rated in a similar fashion to the findings from several studies that complexity must also contain diversity (e.g., Norris & Ortega, 2009; Revesz et al., 2014).

Accuracy was operationalized as a measurement of global grammatical and lexical errors and misuses, as well as the amount of self-repairs that occurred. I hypothesized that in addition to fewer grammatical and lexical errors, fewer self-repairs would predict higher accuracy in a similar fashion to the process of proceduralization (de Bot, 1996; Revesz et al., 2014).

Providing an operationalization for fluency was most challenging, as it has been defined in numerous ways in SLA literature. In order to simplify the rating process for the raters, fluency was operationalized simply as the length of mean run, or the amount of language spoken within natural pause boundaries, and to what degree that is attributable to the rater's sense of "fluent" spoken language（Towell et al., 1996; Skehan, 2009）.

4.5 Analyses

To assess Rin's speaking performance, recordings 2, 5, and 7 were given in a different order to each of the raters, all of whom are English language instructors. Recording 1 was a practice, meaning it did not contain enough information to make it comparable to the information the participant provided in the following recordings. In order to differentiate the recordings from one another, recordings 3 and 6 were not given to the raters, as they were quite similar to the recordings proceeding them（recordings 4 and 7）. Four of the raters were L1 English speakers, and the fifth rater was a Japanese L2 speaker of English with very high proficiency. A rating rubric（Appendix B）was given to each rater, who assigned a score of 1–3（3 being the highest score）to each dimension of CAF, as well as assigning an order to each of the three recordings they listened to.

5. Results

The results from the raters indicated that all five raters agreed on the same order of recordings. The raw data results from the five raters were then collected and input into Microsoft Excel, where they were calculated for their mean scores by adding together the score given by each rater and dividing the result by the number of raters. Table 1 shows the initial mean and standard deviations for CAF scores as reported by the five raters, which indicate that the raters were quite similar in their scores for each recording. Since the recordings were given in a randomized order to each rater, the salience of the CAF features in Rin's spoken language appear to be visible.

Table 1　*Rater mean CAF scores across recordings*

Recording	Complexity M（SD）	Accuracy M（SD）	Fluency M（SD）
2	1.4（0.55）	1.8（0.55）	1.6（0.55）
4	2.2（0.45）	2.2（0.45）	2（0）
6	2.6（0.55）	2.6（0.55）	3（0）

In order to further analyze the results from the raters, their individual responses were added together and a mean score was used to determine rater severity（see Table 2）. Higher mean scores indicate less severity while lower scores indicate greater severity. Raters 1 and 3 were the most severe, with a shared overall mean of 1.89, while raters 2 and 4 were slightly more lenient, and rater 5 was the most lenient of all raters. Close examination of the ratings suggested that the strength of the rater severity could have been influenced by the background of the raters, as raters 1 and 3 were both American-born, rater 2 was a landed immigrant in Australia from Vietnam, rater 4 was Canadian, and rater 5 was Japanese. These differences in the strictness of L2 spoken performance perception could stem from country of origin, gender, or even socioeconomic status, however that needs to be explored in a future study as such considerations were not taken into account during the course of this study.

Table 2　*Rater Severity Based on Overall Mean Scoring*

Rater	Overall Mean
1	1.89
2	2.33
3	1.89
4	2.11
5	2.56

The findings support hypothesis 1, that the provision of explicit corrective feedback on grammatical and lexical features would result in increases across all dimensions of CAF. The results from the five raters suggest that each dimension of CAF noticeably improved as a result of the treatment employed in this study, while fluency and complexity increased the most of the three. Skehan's（2009）Trade-off Hypothesis proposed that all three dimensions of CAF would not increase simultaneously. However, increases across all three dimensions seem to have occurred in the present study. While fluency was an issue during the treatment, the end results suggest that it had noticeably improved once explicit CF was provided for Rin's spoken pause boundaries.

With regards to hypothesis 2, it is difficult to make any substantial claims towards the effects of task repetition, however Rin commented that the combination of CF and task repetition（specifically, listening to past recordings, discussing them, and then practicing the speaking task again）helped him greatly, and the decrease in lexical and grammatical errors over the course of the study tend to agree with him.

6. Discussion

The main purpose of this study was to examine whether provision of CF on the three dimensions of CAF, together with task repetition, could result in noticeable gains within a short period of time. Throughout the sessions with Rin, it was apparent that changes in lexical and grammatical features of his language were occurring and asking him to reflect on this increased his metalinguistic awareness about such changes. The provision of explicit CF in each session forced him to become more aware of the language he was producing. This is similar to Mackey's (2007) suggestion that noticing features of language during output draws attention to one's linguistic capabilities and to new forms of the target language which may have just entered a learner's lexicon, or have yet to be proceduralized into their metalanguage. Through listening to recordings from previous sessions, awareness of Rin's own errors and misuses of the target structures increased, and after a few sessions he was able to reduce the number of errors, as well as the number of self-repairs that occurred in his speech. Aljaafreh and Lantolf (1994) suggested that guidance through feedback is good at the start of learning, but the learner must integrate some of what was provided in the classroom and be able to use it independently. In the present study, Rin was able to demonstrate that uptake of the CF occurred, as is evident in the mean scores the raters gave to each recording.

The effect of task repetition found in this study is proposed to have contributed to not only noticing but reducing the amount of cognitive load necessary for L2 spoken production. These findings are similar to those proposed by Ellis (2009) and de Jong and Perfetti (2011), who discussed the increases in fluency that come from repetition and rehearsal prior to output. Perhaps cognitive load reduction is the key to CAF increases, together with the increased salience provided from appropriate types of corrective feedback.

7. Conclusion

7.1 Implications

Through the provision of explicit corrective feedback and task repetition on target L2 structures, it may be possible to make gains across all dimensions of CAF, which suggests that more CAF studies need to be conducted using similar treatment methods for L2 spoken performance measures. This study found that such gains in CAF were possible in a short period of time, which further suggests that either the external factors or internal factors (individual differences, such as working memory and aptitude) contribute greatly to uptake of language and ability to produce modified output.

Skehan's (2009) Trade-off Hypothesis has value in many CAF studies, as few have examined all three dimensions of CAF simultaneously, and even fewer have provided some form of treatment. In the present study, the nature of the treatment was likely a factor in CAF gains, most likely due to the increases to metalinguistic awareness and opportunities to repeat the same task numerous times.

7.2 Limitations and Future Studies

This study would have benefited from several additions, including a larger number of participants, pre- and post-test proficiency measures, more careful selection of raters, follow-up interviews with raters about their responses, and using more comprehensive analytical software such as Facets in order to analyze the results. While this study was a good demonstration of incorporating CAF into a language program, it needs to be better developed for use in a classroom in order to benefit multiple learners.

Future studies involving CAF should similarly include some type of treatment, as few past studies have done so. Moving forward, more detailed analysis within each dimension of CAF is necessary, while ensuring that these implementations are employed in both assessment and in the treatment is essential to pushing the field ahead.

References

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal, 78*(4), 465–483. doi: 10.2307/328585

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728. doi: 10.1111/flan.12110

de Bot, K. (1996). The psycholinguistics of the Output Hypothesis. *Language Learning*, *46*(3), 529–555. doi: 10.1111/j.1467-1770.1996.tb01246.x

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568. doi: 10.1111/j.1467-9922.2010.00620.x

de Jong, N. H., & Steinel, M. P., Florijn, A. F., Schoonen, R., Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. doi: 10.1017/S0272263111000489

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, *97*(1), 131–143. doi: 10.1111/j.1540-4781.2013.01434.x

Ellis, R.（2009）. The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*（4）, 474–509. doi: 10.1093/applin/amp042

Iwashita, N., Brown, A., McNamara, T., & O’Hagan, S.（2008）. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*（1）, 24–49. doi: 10.1093/applin/amm017

Lynch, T.（2007）. Learning from the transcripts of an oral communication task. *ELT Journal, 61*（4）, 311–320. doi: 10.1093/elt/ccm050

Lyster, R., & Saito, K.（2010）. Oral feedback in classroom SLA. *Studies in Second Language Acquisition*, *32*（2）, 265–302. doi: 10.1017/S0272263109990520

Muranoi, H.（2007）. Output practice in the L2 classroom. In R. M. DeKeyser（Ed.）, *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*（pp. 51–84）. Cambridge, UK: Cambridge University Press.

Norris, J. M., & Ortega, L.（2009）. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics*, *30*（4）, 555–578. doi: 10.1093/applin/amp044

Polat, B., & Kim, Y.（2013）. Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics*, *35*（2）, 184–207. doi: 10.1093/applin/amt013

Revesz, A., Ekiert, M., & Torgensen, E. N.（2014）. The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics, 37*（6）, 1–22. doi: 10.1093/applin/amu069

Skehan, P.（2009）. Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*（4）, 510–532. doi: 10.1093/applin/amp047

Stafford, C. A., Wood Bowden, H., & Sanz, C.（2012）. Optimizing language instruction: Matters of explicitness, practice, and cue learning. *Language Learning*, *62*（3）, 741–768. doi: 10.1111/j.1467–9922.2011.00648.x

Towell, R., Hawkins, R., and 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

Appendix A

CAF Rating for Recordings

Score	Complexity	Accuracy	Fluency
3	• There is a variety in the use of grammatical features and vocabulary throughout the recording	• Few grammatical errors • Few vocabulary errors • Few self-repairs	• The length of mean run (how much is spoken between pause boundaries) is attributable to the fluency of the speaker
2	• There is a small amount of variety in the use of grammatical features and vocabulary throughout the recording	• Some grammatical errors • Some vocabulary errors • Small number of self-repairs	• The length of mean run (how much is spoken between pause boundaries) is somewhat attributable to the fluency of the speaker
1	• Use of grammatical features and vocabulary throughout the recording is limited to one or two tenses, and a small range of vocabulary	• Many grammatical errors • Many vocabulary errors • Many self-repairs	• The length of mean run (how much is spoken between pause boundaries) is not attributable to the fluency of the speaker

Appendix B

CAF Rating Rubric

Recording	Complexity	Accuracy	Fluency	Attempt Number
A				
B				
C				