A Mixed Methods Study on Formative Assessment of EFL Writing of College Students in Japan: Focusing on the Effectiveness of Schematized Teacher Feedback on Coherence

Kana Matsumura

Abstract

The acquisition of performance skills in English as a foreign language (EFL) has long been considered of great importance. To be a good English writer, EFL learners are expected to learn not only to produce grammatically correct sentences in L2 but to present coherent sentences when writing a paragraph. Coherence is considered a significant quality of effective writing (Bamberg, 1984; Lee, 2002; Richards, 1990; Wilkinson, 1990). At the same time, coherence is often considered difficult to learn and difficult to teach as well, especially in the English as a second language (ESL) classroom settings (Cerniglia et al., 1990; Lee, 2002). In EFL writing classes, instructors regularly provide writing feedback that is premised on the evaluation of student writing during classroom writing activities. Given these issues related to coherence, there should be a potential demand for effectively teaching this "essential element of good writing" (Cerniglia et al., 1990, p. 229), coherence, in the classroom.

The aim of this study was to explore how to teach and learn coherence effectively in an introductory English writing class in college through classroom-based formative language assessment. To address this issue, the present study attempted to generate and use diagnostic writing feedback with a schematized tree diagram developed by a web-based annotation tool. This study focuses on argumentative, stand-alone paragraph writing, which requires a certain framework and emphasizes logical coherence. Writing revision is a process of "reading" one's own draft, and a graphical display was used not only to aid in reading but also

facilitate a visual aid in understanding the structure of the entire sentence at a single glance. The research design was based on the hypothesis that annotated diagrams, generated through sentence tagging and linking procedures, are effective for learning coherence.

The study adopted a convergent mixed methods design using an intervention design (Creswell, 2015). The participants were 45 Japanese university undergraduates in two EFL academic writing classes taught by the author, with an intervention group (n = 23) given schematic teacher writing feedback and a control group (n = 22) given conventional text-based feedback. For the quantitative analyses, writing was scored using a modified analytic rating scale for paragraphs and analyzed with FACET for examining the appropriateness of rating-scale functioning or raters' consistency, while mixed-between-within multivariate analysis of variance (MANOVA) was used for exploring the effectiveness of schematized teacher feedback over three occasions of pre-, post-, and retention design. The qualitative analyses consisted of an anomaly analysis on the annotated diagram of the student writing sample and a thematic analysis of the student questionnaire responses. The annotated diagram is a form of assessment record which can provide stakeholders with information about students' writing ability. Therefore, both of the analyses were conducted to examine the effectiveness of feedback qualitatively. The assessment was justified in the framework of assessment use argument (AUA) (Bachman & Palmer, 2010).

As a result, in RQ 1, which tested the consistency and appropriateness of the rating scale, Claim 4 in AUA was confirmed so that the subsequent series of assessments was implemented successfully. In RQ2, the results of both quantitative and qualitative studies showed that the interpretation of the scores provided raters, annotators, and students with information about the ability to be assessed. Moreover, not only score interpretation but the annotated tree diagram were also found in another form of assessment record to be meaningful to both annotators

and students to convey the information in a very understandable way, thereby indicating the meaningfulness of the interpretation (Claim 3). Last but not least, for RQ 3, a mixed methods approach consisting of a series of quantitative and qualitative studies to address the effectiveness of schematized feedback provided findings to show positive intended consequences (Claim 1).

Keywords: formative assessment, diagnostic assessment, graphic display, coherence, annotation scheme, schematized feedback, convergent mixed methods with intervention design, assessment use argument

Chapter 1 Introduction

In this introduction, several key issues related to teaching and learning in an EFL writing class, the study goal, and the research design of the study are discussed. This study is a classroom-based assessment focusing on the teaching and learning of coherent English paragraph writing at a basic level in universities in Japan. The rationale for focusing on writing coherence is as follows. The ability to present the writer's opinion with relevant supporting detail to develop an argument systematically in written form is required in the academic domain (Council of Europe, 2001). This ability to construct ideas in well-organized structure following the format of the English language is a basic skill that is also needed in the general domain. Learning how to write paragraphs, the basic unit of coherent sentences in English, is an important skill for EFL learners to acquire at the introductory level in order to write reports and papers in the academic domain in the years ahead.

One of the key issues related to coherence in writing is its conceptual vagueness. This ambiguity is also represented by the vagueness of the descriptors on the rating scale. Knoch (2007a) noted that "vague descriptions (on the rating scale) might lie in the rather vague nature of coherence itself" (p. 109). Therefore, early in this study, several key terms related to coherence were defined so as to avoid the vagueness of the definition in the subsequent discussion. Two of these are presented here as examples. *Coherence* is defined as the "semantic property of discourses, based on the interpretation of other sentences" (van Dijk, 1977, p.93). *Cohesion* is a structural connection between sentences and their components that is explicitly illustrated with formal markers (van Dijk, 1977), which in this study is considered an important component of coherence.

The study focused on formative teacher feedback in classroom-based assessment in a college EFL academic writing class. To be specific, the study was conducted to investigate the effectiveness of schematized feedback with a tree

diagram generated by an annotation tool developed by the author, who is also the instructor as well. The study was conducted using a convergent mixed-methods with intervention design (Creswell, 2015) under the assessment use argument (AUA) (Bachman & Palmer, 2010) as a justification framework.

Chapter 2 Literature Review

The previous studies reviewed herein are diverse, with topics covering multiple subjects. The following are the topics reviewed therein: the current issues related to EFL writing in Japan, writing feedback, diagnostic language assessment of EFL writing, coherence in writing, text annotation tools and graphic display as structural representation, and relevant rating scales for diagnostic assessment. Among these, the most specific to this study is the review of text annotation tools and graphic display as structural representation.

There has been research on the use of reading comprehension in L1 as represented by graphic organizers. However, while the effectiveness of graphic display application in L2 language learning has been pointed out, it has not been sufficiently investigated in L2 reading research (Jiang & Grabe, 2007). On the other hand, several studies in L2 writing research have utilized graphic displays to detect and categorize coherence errors in EFL writing (e.g., Ahmadi & Parhizgar, 2017; Kawase, 2020; Skoufaki, 2009; Yamashita, 2019). Among the annotation schemes, Mann and Thompson's (1988) rhetorical structure theory (RST) is one of the most widely used. While RST's detailed discourse analysis is excellent for rigorous discourse analysis, it requires specialized knowledge to determine relation labels. Since this study assumes use for evaluation and analysis in actual classes and presentation to and understanding by students, a simpler annotation tool was desired. This is the rationale for using the Tool for Interactive Argument Annotation (TIARA), which was recently developed by a research group at the

Tokyo Institute of Technology (Putra et al., 2020). TIARA's usability, features, and analysis methods are presented in the Methodology section below.

The following are the three main research questions in the AUA framework. RQ 3 has four sub-research questions.

RQ 1: Do the rating scale and the text annotation used in the present study function properly?

RQ 2: How can students' overall writing performance and organization of their writing be characterized/interpreted through the rating scale and the annotation scheme?

RQ 3: To what extent and how does the type of teacher feedback (conventional versus graphic) affect (1) the overall writing performance and organization of their writing across occasions (initial draft, revised draft of the initial task, and a transfer task) in terms of scores, (2) revision time on task, (3) ideational and rhetorical coherence in the transfer task (through the analysis of information obtained by the annotation tool), and (4) students' rewriting behaviors and perceptions?

Chapter 3 Methodology

The following section outlines the methodology adopted in the study. The present study adopted a mixed-methods design consists of quantitative and qualitative studies. In quantitative study, assessment records are given in the form of scores, while in the qualitative study, tree-shaped annotated diagrams and students' responses to the questionnaire are assessment records to be analyzed.

The participants were 45 Japanese university undergraduates in two EFL academic writing classes taught by the author, with an intervention group (n = 23) receiving schematic teacher writing feedback and a control group (n = 22) receiving conventional text-based feedback. They wrote argumentative paragraphs,

whose topics were taken from Eiken Grade Pre-1 and Grade 2. The study was conducted over three writing occasions with two tasks differing in difficulty plus another task to examine the retention skill. Two tasks of different levels of difficulty were given in a counterbalanced manner. The test followed a process writing format, with an initial draft followed by different types of feedback, a second draft, which is a revision task of the initial draft, and a transfer task. The raters (n = 6) with Ph.D. or M.D. degrees have had experience with teaching writing to EFL students at university. The annotators (n = 2), who are also the raters, have had experience with using the annotation used in the study for three years.

The instrument for scoring students' writings, the analytic rating scale of the English as Second Language (ESL) Composition Profile (Jacobs et al., 1981) was modified for rating paragraph writings. For annotating students' writings, TIARA was employed.

For quantitative analysis, the many-facet Rasch measurement (MFRM; Linacre, 1989; Linacre & Wright, 1993; McNamara, 1996) was employed to examine the severity and consistency of rating scores as well as the propriety of the rating scales. A mixed-between-within multivariate analysis of variance (MANOVA) was conducted with a between-groups independent variable (two different groups based on feedback type: conventional vs. graphic) and a within-groups independent variable (three writing occasions with a repeated design). This analysis allowed examination of the effects of feedback type on score improvement across writing occasions for four writing skill criteria. In addition, descriptive statistics were used to examine the mean, standard deviation, and minimum and maximum time required for the analysis comparing on task revision time.

The qualitative analysis consisted of an anomaly analysis of the frequency, location, and types of anomalies with the annotated diagram of the student writing sample. A thematic analysis (Braun & Clarke, 2006; Creswell & Cresswell Báez

2021; Flick, 2014; Takagi, 2021) of the student questionnaire responses regarding rewriting behaviors and perceptions of teacher feedback was conducted. The qualitative research results from the anomaly analysis and tree-diagram shape analysis were merged with the scores obtained from the quantitative study, and the results were analyzed, examined, and discussed from the perspective of a mixed methods approach.

Chapter 4 Results

The following is a brief summary of the main findings of the quantitative study. An MFRM analysis showed the consistency of ratings (consistency was confirmed as two ratings in pairs, not two raters) in terms of severity. In addition, the overall reliability of the rating scale was found to be consistent enough to be used in the subsequent analyses, although there was some limitation in terms of threshold distance for some writing skill criteria. The results of a mixed-between-within MANOVA showed that there was no significant statistical difference between the two groups but that in the transfer task, a significant increase from the first draft was observed in the intervention group in some of the criteria. Finally, descriptive statistics showed that the average revision time on task was shorter for the intervention group that received graphical feedback.

Before conducting a qualitative study, the following four types of inter-annotator reliability showed sufficiently high agreement in the annotators' judgments: tagging/labeling the relation of the text unit, linking the source-target sentence, and locating and classifying coherence anomaly. The main findings of the qualitative study are as follows. The intervention group had a lower frequency of coherence anomalies in the whole writing sample than the control group. Along this line, the number of writings free of coherence anomalies was greater in the intervention group than in the control group. Furthermore, regarding the shapes of

the tree diagrams generated by the annotation tool, it was found that the percentage of the three types of poorly formed tree diagrams (horizontally wide, unbalanced, and vertically long) was higher in the control group (50%) than in the intervention group (30%). This suggests that the intervention group produced writing that could be represented by a well-balanced shape diagram, resulting in higher organization scores than the control group. Finally, the students' responses to the questionnaire regarding revisions in their writing and teacher feedback revealed the following. Students in the intervention group reported more revisions and more detail. In addition, students in the intervention group provided more objective feedback, such as understanding how their writing was viewed by others. These results can be considered positive consequences of the schematized feedback.

Chapter 5 Discussion

This section discusses the three research questions by reviewing both quantitative and qualitative results to integrate them in order to draw interpretations about the efficacy of schematized teacher feedback on EFL argumentative paragraph writing. These research questions are discussed based on the framework of the AUA (Bachman & Palmer, 2010). Each research question and sub-research question is intended to address each of the claims in AUA. RQ 1 addresses the consistency or reliability of the rating scale and annotation scheme (Claim 4). RQ 2 responds to the meaningfulness of interpretation of students' performance on the rating scale and in the annotation scheme (Claim 3), and RQ 3 addresses the effectiveness of schematized teacher feedback (Claim 1) followed by four sub-questions. RQ 3(1) responds to the effectiveness of performance on writing skill criteria across occasions (Claim 1), RQ3(2) the efficiency of revision time on task (Claim 1), RQ3(3) the effectiveness of performance on ideational and rhetorical coherence in transfer task (Claim 1), and RQ3(4) addresses the extent to

which and how the schematized feedback was beneficial as identified by the responses related to their perceptions (Claim 1).

As described in the Results section above, except for revision time on task, the quantitative analysis identified few significant differences in scores. However, in some cases, there were differences in the transfer task or among the criteria. Interestingly, while differences between the two groups in the organization score was expected as the different type of feedback effect, there was actually a larger difference in the language and vocabulary scores. While it is difficult to pinpoint the cause of this phenomenon, one possible explanation is that the annotation diagrams used as feedback were separated by text units, making it easier to focus on a single sentence and to find errors at the sentence level. It is also possible that more attention was paid to each individual word because of the decrease of the cognitive load because of schematized feedback, which resulted in better scores in these criteria for the intervention group. This may be a common factor with the following decrease in revision time on task.

Revision time on task showed the effectiveness schematized feedback. That is, the intervention group rewrote more efficiently than the control group. The visual argument hypothesis that presupposes "graphical representations are effective because, due to their visuospatial properties, their processing requires fewer cognitive transformations than does text processing and does not exceed the limitations of working memory" (Vekiri, 2002, p. 281) and the experimental result regarding the effectiveness of graphical representation on time on task (Winn et al., 1991) led to one of the RQs in the present study related to time on task.

While the quantitative results showed a few significant differences between the two groups, the qualitative findings revealed some more positive effect in the intervention group given schematized feedback in terms of the perception of the revision process, the perception of the feedback, and the frequency of the coherence anomaly and the shape of the annotation diagram. It should be noted

that some of the students in the intervention group showed changes in their behavior during the writing planning stage, such as thinking using a tree diagram.

On the other hand, in terms of quantitative analysis, it is not clear whether the lack of a significant difference between the control group receiving conventional text-based feedback and the intervention group receiving schematized feedback is due to an actual lack of significant differences in feedback effects or to other issues, such as the raters, the rating scale used in this study, or the interaction of all of these factors. In addition, a small sample size decreases the statistical power, so it may be difficult to statistically confirm significant effects in classroom assessments.

Finally, regarding the teacher's effort in creating the feedback, since this study used two tasks with different levels of difficulty, the feedback was created twice. However, the students' comments indicated that the graphical feedback left a strong impression, and therefore it is considered effective even when implemented once per semester.

Chapter 6 Conclusion

The present mixed methods study addressed issues regarding the teaching and learning of ideational and rhetorical coherence in EFL paragraph writing in terms of formative classroom-based assessment of university students at a basic level in Japan. The study was conducted to investigate the efficacy of schematized feedback with a tree diagram generated by an annotation tool. The results showed that there was no significant statistical difference between the two groups, except for some results concerning the transfer task. On the other hand, the results of the qualitative study showed some positive effect with the intervention group provided with the schematized feedback, such as the reduction in time on task during revisions and the well-balanced shape of annotated diagrams in the transfer task,

which indicates high-quality organization of the passage. Furthermore, the results of a questionnaire showed that the intervention group students tended to be more sensitive and attentive to writing coherence than the control group. Some students in the intervention group were observed to show behavioral changes in the planning process.

Despite some limitations associated with empirical research conducted in a classroom setting, such as small sample size and difficulties in controlling the educational environment among control groups from an ethical point of view, the possibility of certain qualitative effects due to the fact that the author is an instructor who knows the students well, this study provides several pedagogical implications. One of them is the effectiveness of graphic displays for EFL writing activities in teaching and learning. In addition, since the annotation process itself, such as text segmentation, tagging, and connecting text units, may provide an opportunity for L2 writers to improve their understanding of coherence, it would be worthwhile to conduct future research on the further use of annotation tools by instructors and on students' own attempts to engage in annotation work.