Exploring Japanese EFL Students’ Short-Term Writing Development

HIROSE, Keiko
Aichi Prefectural University

Abstract
This study examined how intermediate-level Japanese EFL students’ written English changed over 15 weeks. The study aimed to identify measures of students’ L2 short-term writing development in terms of fluency, syntactic, and lexical complexity. Two groups of undergraduates, who differed in prior writing background, received an English writing course of similar content. The study first compared pre- and post-course English compositions within each group, and then between-group comparisons of the pre- and post-course compositions were made, respectively. Nine linguistic measures were employed including raw fluency, and ratio measures of syntactic and lexical complexity. The results revealed that the less experienced students showed a significant increase in fluency, lexical complexity, and one syntactic complexity measure (number of T-units per sentence), whereas the more experienced students made no significant changes in any measure. Based on the findings, implications are discussed and several directions for future research are provided.

Keywords: EFL writing development, fluency, syntactic complexity, lexical complexity

Introduction

Background of the Study
This classroom-based study focuses on Japanese EFL university students’ writing development over a semester, and is oriented by two related strands of research. One is research concerned with L2 writing development. This strand has commonly examined development in terms of complexity, accuracy, and fluency (CAF) (e.g., Wolfe-Quintero, Inagaki, & Kim, 1998), and longitudinal investigations have measured L2 writing development by using numerous CAF measures (e.g., Celaya & Navés, 2009). The second strand of research related to the present study is classroom-based studies that examine intervention effects, more specifically, those conducted in Japanese EFL classrooms (e.g., Hirose & Sasaki, 2000). The L2 writing development strand of related research is explained first, followed by a review of the classroom-based research strand examining the effects of English writing instruction.

The present study examined L2 writing development in terms of the linguistic features of students’ writing, particularly the fluency, and syntactic and lexical complexity of the writing. In this study, linguistic development was approached with the assumption that an increase or a decrease in fluency and complexity cannot be considered either a positive or a negative
change in its own right. As explicitly pointed out by Pallotti (2009), CAF and development should be considered separate in the sense that “CAF refer to the properties of language performance as a product, while development is a process, with its sub-dimensions such as a route and rate” (p. 594). It has been shown that the relationship between fluency and complexity of student writing is not straightforward, and they are sometimes even inversely related. As reported by Larsen-Freeman (1978), for example, although the number of words per composition increased with student proficiency, it decreased within the highest proficiency student group. This tendency of decreasing fluency might be related to the development of complexity. Decreasing composition length with increasing proficiency may be accounted for by students’ inclination to use complex phrases such as nominalizations and modification, rather than coordination or subordination. Through the use of nominalizations, for example, clauses can be reduced to phrases, resulting in a fewer number of words. At the same time, such a seemingly non-linear reverse relationship has also been noticed among different aspects of syntactic complexity itself. For example, Wolfe-Quintero et al. (1998) summarized the developmental patterns some previous studies proposed as a “move from coordination to subordination to the reduction of clauses to phrases” (p. 73). Thus, as indicated by Norris and Ortega (2009), “a decrease in subordination at the highest levels of proficiency may be related to an increase in the overall complexity of the language performance” (p. 566). Therefore, relationships between fluency and complexity should be elucidated.

In the following sections, fluency, syntactic and lexical complexity are operationally specified, and then measurements are explained. Next, studies that have investigated L2 writing development using such measures are reviewed. After that, intervention studies are discussed in relation to the focus and content of the English writing instruction implemented in the Japanese EFL context. Finally, the research gap to be filled by the present study is specified.

Measurement of L2 Writing Development

**Fluency.** As pointed out by Housen and Kuiken (2009), fluency in L2 writing has been under-researched, especially compared to fluency in L2 speaking. Fluency is considered “a multidimensional construct, in which sub-dimensions can be recognized” (Pallotti, 2009, p. 591). Regarding writing fluency, such sub-dimensions consist of at least the following two. One is a quantitative sub-dimension of fluency in which the amount of written production, especially within an allocated time, is measured, whereas the other sub-dimension considers the writing process, i.e., how fluently writers can write with little pausing or hesitation. The latter sub-dimension is measured by counting the number of sentences written “without interruption, or without engaging in other activities such as reading back or rehearsing” (Raimes, 1985, p. 243) or by the analysis of pausing behavior while writing (e.g., Kobayashi & Rinnert, 2013). The latter aspect of fluency has been traced by video-taped and think-aloud data or by counting the number of words produced per minute. The studies that captured this process sub-dimension of fluency tend to have a smaller number of participants (e.g., eight in Raimes, 1985; one in Kobayashi & Rinnert, 2013), compared to those that addressed the product sub-dimension of fluency. This study focused on the quantitative aspects of writing.
fluency. Because the participating students wrote compositions in the classroom situation within the same time limit, the amount of production was considered a valid gauge of fluency.

Previous studies that investigated quantity of written production have employed measures such as the total number of words (e.g., Henry, 1996; Storch, 2005), along with the total number of clauses (e.g., Robb, Ross, & Shortreed, 1986) and T-units (“minimal terminable unit”) (e.g., Ishikawa, 1995; Ross, Shortreed, & Robb, 1988). T-unit (TU) is “one main clause plus whatever subordinate clauses are attached to that main clause” (Hunt, 1966, p. 737). In addition to the numbers of words (W), clauses (C), and TU, the number of words per T-unit (W/TU) has also been used as a fluency measure in some studies (e.g., Ishikawa, 2006; Yamanishi, 2011), along with words per clause (W/C) (e.g., Celaya & Navés, 2009). Although Wolfe-Quintero et al. (1998) listed these measures as the most appropriate measures of fluency development, W/TU and W/C have also been among the most frequently used syntactic complexity measures in previous studies (e.g., Ortega, 2003).

**Syntactic complexity.** Syntactic complexity is a multi-dimensional construct with interrelated sub-constructs, which have been identified and measured at the sentential, clausal, and phrasal levels (Bulté & Housen, 2012). To grasp multi-dimensional syntactic complexity, multiple measures were adopted in past studies, but rarely in a single study. In their review of the syntactic complexity research, Bulté and Housen (2012) found most of the studies employed only one or two measures from among a total of 27 syntactic complexity measures identified. Accordingly, Bulté and Housen claimed a multi-dimensional construct of complexity has not been sufficiently operationalized in the existing L2 research. Providing both theoretical and empirical justifications, Norris and Ortega (2009) argued for the measurement of dynamic aspects of syntactic complexity multi-dimensionally. More specifically, they recommended the following dimensions to be employed in the same study: general or overall complexity, complexity via subordination, sub-clausal complexity via phrasal elaboration, and complexity via coordination in cases where low-proficiency level data are included. Given impetus from their call for such research, this study drew on Norris and Ortega’s recommendation.

Thus, this study used five measures to grasp these multiple dimensions of syntactic complexity. First, to capture general complexity, W/TU and the number of S-nodes per T-unit (SN/TU) were employed. The former measure has been the most commonly employed in past studies, and in fact, Ortega (2003) found it was the only measure employed in L2 longitudinal writing studies she surveyed that examined L2 students’ linguistic development. The measure with S-node in the numerator was also added in the present study because it is considered to “show a greater sensitivity for measuring small differences in complexity at relatively low levels of proficiency” (Norris & Ortega, 2009, p. 566). In fact, SN/TU has been used in studies with lower-level Japanese EFL students (e.g., Ishikawa, 2006; Yamanishi, 2011). S-node is equivalent to a verb phrase (VP) (either finite or non-finite). Finite VPs are independent, adverbial, adjectival, and nominal, whereas non-finite VPs are infinitive, gerund, and participle. Second, the number of clauses per T-unit (C/TU) was adopted to measure complexity by subordination (i.e., finite clausal subordination). As in the case of W/TU, C/TU has also been used in previous L2 writing research (e.g., Storch, 2005; Yang, Lu, & Weigle, 2009).
Third, W/C was used to reflect complexity via sub-Clausal or phrasal elaboration, regarded as most suitable for upper proficiency levels (Norris & Ortega, 2009). As pointed out by Bulté and Housen (2012), however, W/C is controversial as a ‘pure’ phrasal complexity measure, and clause length depends on how a clause is defined. Previous studies have defined clauses differently, which makes comparison of the results problematic. In the present study, a clause was defined as containing a visible subject and a finite verb, including independent, dependent, or subordinate clauses. Although studies such as Bulté and Housen (2014) and Storch (2005) included non-finite verbs in counting a clause, a non-finite verb was not counted here as constituting a clause, in accordance with other previous studies (e.g., Yang et al., 2015). Lastly, TU/S, which shows the amount of coordination per sentence, thus reflecting clausal coordination, was also employed to capture different dimensions of complexity from measures with T-unit in the denominator (see Bardovi-Harlig, 1992, for justification of using the sentence). As argued by Bardovi-Harlig (1992), TU/S is considered discriminatory for beginning levels of L2 development (Norris & Ortega, 2009).

**Lexical complexity.** Lexical complexity also encompasses multiple aspects and there are numerous measures available to gauge lexical complexity. These are differentiated between “text-internal measures (so called because the text itself is sufficient for their calculation) and text-external measures (which require some sort of general reference material, usually based on word frequency)” (Skehan, 2009, p. 514). Although the scope of the measurement was limited, the present study adopted one text-internal measure. Text-internal measures include type-token ratio (i.e., the total number of word types divided by the total number of word tokens) and the Guiraud index (the number of word types divided by the square root of the number of word tokens; word types/$\sqrt{\text{word tokens}}$). Because the type-token ratio has been criticized for being affected by text length (e.g., Ellis & Barkhuizen, 2005), the Guiraud index ($G$) has been employed in many studies to eliminate the influence of composition length (e.g., Bulté & Housen, 2014). Rather than an index of sheer diversity, Bulté and Housen (2014) considered “$G$, which captures both diversity and productivity, to be a useful complement to the Diversity index, especially for the analysis of timed writing samples that are not controlled for length” (pp. 49–50) as in the written data collected in the present study. Therefore, this study employed $G$ as a lexical measurement. Although there are other aspects of lexical complexity, such as lexical density and lexical sophistication (Bulté & Housen, 2012), it was beyond the scope of the analysis to investigate lexical complexity in its entirety.

**L2 Writing Development Research**

Most longitudinal studies examining L2 writing development to date have been conducted over a short term, such as several months. However, Casanave’s (1994) longitudinal study examined four Japanese EFL students’ writings over a period of three semesters. Also Kobayashi and Rinnert’s (2013) longitudinal study investigated one Japanese student’s linguistic, process, and social dimensions of writing development over two and a half years.

A short course such as one semester has been considered insufficient to capture substantial improvement in syntactic complexity. For example, Ortega (2003), who reviewed and synthesized the findings of such short-term studies, found “a negligible to small-sized
change” (p. 511) in W/TU. Nevertheless, several recent studies examining ESL students’ writing development in a semester-length university writing program found noticeable improvements in syntactic complexity (e.g., Bulté & Housen, 2014). Although changes or improvements made by EFL students at the college/university level may be smaller than those of ESL students (Ortega, 2003), Casanave (1994) reported positive changes including longer, more complex, and more accurate text construction emerged in Japanese EFL students’ writing in the first semester. Thus, a study of one-semester-length writing development was regarded as worthy of being conducted.

As advocated by Bulté and Housen (2012), multiple complexity measures should be employed in the analyses to fully embrace student writing development. In a longitudinal study of L2 writing development, relationships need to be elucidated not only among the syntactic complexity measures, but also in relation to other measures of fluency. Drawing on the findings of existing literature, Wolfe-Quintero et al. (1998) suggested that “fluency and complexity measures may be related to the construct ‘development,’ but that accuracy measures may be related to a different construct ‘error’” (p. 118). Taking these concerns into consideration, multiple measures of fluency and complexity were adopted while accuracy was not examined in the present study.

Such a study of L2 writing development through multiple measures is also worthwhile in light of the following strand of L2 writing intervention studies. Many intervention studies to date tend to use a limited number of measures to investigate the instructional effects. As pointed out by Connor-Linton and Polio (2014), intervention studies that investigate change over time “might use only one or a few measures” (p. 7), and thus they urge the need to employ more measures.

Classroom-Based Studies Examining the Effects of English Writing Instruction

The second strand of research related to the present study comprises classroom-based intervention studies that investigate whether (and if so, how) students improve their English writing after receiving writing instruction for a semester or a longer period. Such studies have also been conducted in Japanese EFL university classrooms. In Japan, most students enter university with six years of highly controlled, formal English education without having taken an English writing course per se before. It is common for many Japanese students not to have written English beyond the sentence level prior to university. Indeed, they have rarely expressed themselves in English. Furthermore, at university, English writing is not required for most students. English writing courses are often integrated with other language skills like speaking and listening, or are provided as a means to reinforce grammatical structures or vocabulary (Ishikawa, 1995).

Thus, Japanese EFL students not only need English writing instruction but also English writing experience. More specifically, Sasaki and Hirose (1996) found that Japanese EFL novice writers lacked both the knowledge of English expository writing and paragraph-level writing experience. Thus, knowledge along with writing experience can be considered essential components of writing instruction for inexperienced Japanese university students. Classroom-based studies have been conducted on the effects of writing instruction on student writing. Given instruction in English paragraph organization such as time order and cause and
effect for a semester, students’ English writing did not improve solely as a result of analyzing and studying model paragraphs, which suggests students need writing instruction that incorporates writing experience (Hirose & Sasaki, 2000). The same study investigated whether writing experience operationalized as weekly out-of-class journal writing combined with paragraph instruction positively influenced students’ subsequent writing. Results revealed that the students gained in terms of the total number of words and improved mechanics, with no significant improvement of the overall quality.

Although its effects on writing quality have not been consistently positive, improving fluency has been shown to be an established effect of writing instruction incorporating journal writing (e.g., Hirose, 2005). For example, investigating the effectiveness of one-year English instruction using several practices including journal writing, Ross et al. (1988) reported those students who practiced journal writing developed fluency especially in narrative writing. In sum, the findings of these intervention studies suggest that facilitating writing practice works at the early stages of university-level EFL writing courses, providing additional “evidence for a fundamental role for fluency-aimed intervention in the teaching of FL writing, particularly at incipient levels of proficiency” (Ortega, 2009, p. 238).

Taking the above findings into consideration, alternative writing practices to fluency-aimed journal writing that might lead to quality improvement were created and implemented. More specifically, expository paragraph writing practice, together with paragraph development instruction, was assigned during every class in a semester-length writing course for novice Japanese students, and the effects on student writings were investigated (Hirose, 2012b). In this paragraph-focused instruction, peer feedback was implemented, in which students wrote a composition following a newly learned paragraph organization at home and then read and gave feedback on each other’s paragraphs in class. Impetus for integrating peer feedback, a new activity for most Japanese students, originated from major concerns of motivating them to write English beyond a paragraph level while learning about paragraph development every week. The classroom-based studies of this type of instruction (Hirose, 2012a; 2012b) found that Japanese students had positive attitudes towards peer feedback and enhanced motivation towards English writing. The instruction was also found to have exerted positive influences over fluency (in terms of number of words), but not necessarily on overall writing quality, resonating with the findings of previous intervention research.

Previous studies that suggest fluency improvement is not compatible with improvement of writing quality have not examined specifically how students’ written English changed in measures other than total number of words and overall evaluation scores. Thus, the question remains as to how students’ written English changes in terms of aspects other than fluency, after they receive paragraph instruction and experience with writing in a semester course. To scrutinize change over time, how their written English changes in terms of syntactic and lexical complexity needs to be fully elucidated. Furthermore, previous studies (e.g., Hirose, 2012b) also found that the instructional effects differed depending on student writing background, such as L2 writing proficiency level and prior writing experience, thus calling for research taking students’ writing background into consideration.
The Present Study

Thus, the present study attempted to identify measures to elucidate Japanese EFL students’ writing development over a semester. For this purpose, the study compared the pre- and post-course compositions written by two groups of Japanese undergraduates who received semester-length English writing instruction that focused on paragraph development and paragraph-level expository writing practice facilitated by peer feedback. The two groups of students differed in their prior English writing experiences, but not substantially in their English language proficiency and English writing proficiency levels.

The research questions of this study were:
1. Do Japanese EFL students change or improve their compositions after receiving a semester-length course incorporating English paragraph instruction and regular writing experience?
2. How does the change or improvement differ according to students’ prior writing background (instruction and experience)?

Method

Participants

The participants of the study were two groups of Japanese university students taking English writing courses taught by the author (N = 51; 8 males and 43 females). Neither group of students was taking any other writing class. The two groups differed in their prior English writing backgrounds (i.e., experience and instruction). Group A (n = 22; 5 males and 17 females) were first-year students who had little prior English writing instruction or experience, whereas Group B (n = 29; 3 males and 26 females) had received one year of English writing instruction.

The same writing background questionnaire survey was administered to A and B groups to examine how their prior instructional and personal writing experiences, as well as their perceptions of English writing, differed at the commencement of the courses. Neither Group A nor Group B had experienced peer feedback previously. The questionnaire results confirmed that Group A had received less writing instruction and experience than Group B. The differences were tested for significance using chi-square analysis. The difference between the two groups was significant for instruction of two items: (a) paragraph development and (b) how to write research papers. Regarding their prior English writing experience, Group B reported having more experience in writing (a) diaries/journals, (b) summaries or paraphrases of materials read, (c) reports or research papers, and (d) letters/e-mails, chats than Group A.

On the other hand, the English proficiency levels and writing abilities of the two groups were not significantly different. Their English proficiency levels were mostly intermediate (ranging from low- to high-intermediate) and considered higher than average for Japanese university students in the sense that their TOEIC scores fell between 500 and 800. Furthermore, their pre-course English compositions did not reveal any significant difference in terms of total scores (content, organization, language use, vocabulary, and mechanics) evaluated by three external raters (see the Data section below).
Class Procedure and Amount of Writing

Both A and B groups received paragraph-focused instruction using introductory course books. The same class procedure was used with all the students. The classes met separately once a week for 90 minutes each, over the course of a 15-week semester. For 12 weeks, prior to each class, students were assigned to write a composition with a minimum length of one paragraph; they then wrote feedback on their partners’ compositions in English. The first 45 minutes of the class time were devoted to peer feedback activities based on the writing assignments. In this part, students spent approximately 20 minutes paired with partners, reading each other’s compositions and writing feedback, and the remainder of the time was spent reading peer feedback and engaging in spoken feedback. The other half of the class was spent on English paragraph instruction using the course book. Students learned about English paragraphs by reading and analyzing sample paragraphs in English, and then were expected to write compositions related to paragraph organizations covered in a previous class.

Table 1 shows how much writing each group did during the course. Group A wrote an average of 119.3 word-long compositions, whereas Group B wrote an average of 155.6 word-long compositions over the semester. The results of a t-test showed that Group B produced statistically more words per composition than Group A [t (49) = 4.82, p = .000, r = .57]. For composition assignments, students of both groups chose to write mostly about themselves and about topics relating to themselves, thus utilizing opportunities for self-expression.

<table>
<thead>
<tr>
<th></th>
<th>Group A (n = 22)</th>
<th>Group B (n = 29)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of compositions produced</td>
<td>11.2 ± 0.69</td>
<td>10.5 ± 1.09</td>
</tr>
<tr>
<td>Total number of words per composition</td>
<td>119.3 ± 16.8</td>
<td>155.6 ± 35.8</td>
</tr>
</tbody>
</table>

Data

The pre- and post-course English compositions were the major data sources for the present study. At the beginning and end of the semester, students wrote an argumentative task, taking one of two given positions, and supporting it in 30 minutes. This type of task was chosen for several reasons. First, many studies dealing with intervention effects have used such tasks (e.g., Hirose & Sasaki, 2000; Kamimura, 2006). Second, this type of writing is what students are expected to achieve. For example, the Independent Writing Task in TOEFL iBT “requires writing an essay that states, explains, and supports the writer’s opinion on a given issue” (Educational Testing Service, 2006, p. 42).

Importantly, all the participants of the present study had received expository writing instruction that covered such elements as the topic sentence, the body, and organizational patterns (time order, comparison and contrast, cause and effect, etc.) and words/phrases typically used with them at the time they wrote post-course compositions. Although the instruction did not specifically include argumentative essay writing, students were expected to
exercise their learned knowledge about and experience of English expository writing in an argumentative task.

Two topics “university students and part-time jobs” and “English learning and studying abroad” were used (see Appendix A), but in opposite orders for the two groups. Two topics were used in the present study to avoid possible influences of a practice effect of previous writing on the same topic. Both topics were considered equally familiar to these students, most of whom were working part-time and had interests in studying abroad. The students were not informed about the topics beforehand and were not allowed to use dictionaries.

For the evaluation, three English-speaking instructors with MA’s in TESOL or education scored both groups’ pre- and post-course compositions without knowing which were either pre- or post-course compositions, according to an adapted version of Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey’s (1981) ESL English Composition Profile (see Yamanishi, 2004). Ratings were assigned equally (10 points each) for the five criteria of content, organization, language use, vocabulary, and mechanics. Each student’s composition score was the sum of the three raters’ scores (the full total score = 150). Table 2 shows the two groups’ pre- and post-course composition total scores. The results of t-tests found no significant differences between the two groups either in pre-course or post-course total scores [pre-course $t(49) = 0.13, p = .90, r = .02$; post-course $t(49) = 1.29, p = .20, r = .18$]. However, Group A’s post-course total scores were significantly higher than their pre-course total scores [$t(21) = 3.25, p = .004, r = .33$], whereas there was no significant difference between the pre- and post-course compositions of Group B [$t (28) = 1.42, p = .167, r = .07$]. Both groups improved their post-course composition scores, but it was Group A (the less experienced group) that made significant improvement.

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Group A ($n = 22$)</th>
<th>Group B ($n = 29$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>Pre-Course</td>
<td>96.95</td>
<td>9.47</td>
</tr>
<tr>
<td>Post-Course</td>
<td>103.23</td>
<td>8.35</td>
</tr>
</tbody>
</table>

Note. the full total score = 150. Reliability estimates were based on Cronbach $\alpha$.

**Data Analysis**

Because compositions were all hand-written, the 102 compositions (two each by all the students) were typed in word documents for further analysis. For quantitative analysis, the pre- and post-course compositions were compared using nine measures to examine fluency, and syntactic and lexical complexity. To measure fluency, simple counts of words (W), clauses (C), and T-units (TU) were employed. These measures were considered valid not only for assessing but also for comparing writing fluency (refer to the Fluency section above). In counting, sentence fragments were not regarded as T-units. T-units could occur across periods (Ishikawa, 2006). For example, “Because there are many foreign people who can speak English and many English schools in Japan” (see a student’s sample pre-course composition in Appendix B) was not regarded as a T-unit. Regarding syntactic complexity, the
following five measures were adopted in the present analysis to capture its multidimensionality (Bulté & Housen, 2012; Norris & Ortega, 2009): W/TU, C/TU, W/C, SN/TU, and TU/S. As a measurement of lexical complexity, one measure, G, was chosen.

Except for the numbers of W (tokens) and G calculated by AntConc (2012) version 3.3.1, (created and provided on the web by Laurence Anthony), the other raw measures (C, TU, SN, and S) were analyzed manually by two researchers/teachers of English. When there were discrepancies between the two raters, the author coded the data and resolved the disagreements through discussion until 100% agreement was reached.

For Research Question (RQ) 1, the pre- and post-course compositions were compared within each group by using paired t-tests. Because multiple tests were employed, a Bonferroni adjustment was made (Tabachnick & Fidell, 2013). Nine comparisons were made for each group to answer RQ1; thus, the alpha level of 0.05 was divided for the study by the number of comparisons (i.e., 0.05/9), and only those tests that resulted in below the 0.0056 level were accepted as significant. For RQ2, the two groups’ pre- and post-course compositions were compared, respectively, by using independent t-tests. Nine comparisons were made between Groups A and B in RQ2. A Bonferroni adjustment was made because multiple tests were conducted; thus, again the alpha level of 0.05 was divided by the number of comparisons (i.e., 0.05/9), and the significance level was set at 0.0056.

**Results and Discussion**

**RQ1: Pre-Course vs. Post-Course Composition Within-Group Comparisons**

Table 3 shows the means and SDs of the nine measures of pre- and post-course compositions written by Group A (the less experienced group). The results of repeated-t tests showed there were significant differences between the two compositions in five out of nine measures. Results also showed large effect sizes for the differences in these five measures (r >.6). Group A gained significantly in the total numbers of W, C, TU, and G. For instance, the post-course compositions were longer than the pre-course compositions by 46.4 words. There was also a significant difference in one of the syntactic complexity measures, TU/S, which was employed to measure clausal coordination. As shown in Table 3, Group A produced an average of nearly 2 TU/S in post-course compositions, as opposed to 1 TU/S in pre-course compositions. Although this coordination index increased, other syntactic measures all showed decrease. This finding indicates Group A improved only beginning-level syntactic complexity because, as stated above, coordination provides evidence of complexity at the beginning levels of development. As reported in the Data section, Group A’s post-course composition scores improved significantly, which matched their significant gains in fluency and lexical measures and one syntactic measure. The findings suggest that writing fluency and lexical complexity improve as regular writing experience of compositions and peer feedback accumulates, even for a 15-week semester (see Appendix B for a Group A student’s pre- and post-course compositions).

On the other hand, in the case of Group B (the more experienced group), the results of the paired t-test showed no significant differences on any measures (see Table 4). Ranging from medium to negligible, the effect sizes were mostly small with the exception that the total numbers of W and TU indicated medium effect sizes (r >.3). The students numerically gained
in terms of W, C, TU, and G in the post-course compositions, but not at the designated level of significance. Thus, the present findings imply that in the case of the more experienced group, writing fluency and lexical complexity may develop more slowly or level off even despite more writing experience. Group B showed no significant improvement in their post-course composition scores. These non-significant results are worthy of attention considering their significantly greater amount of English writing during the semester (recall Table 1).

Table 3

Pre- and Post-Course Compositions for Group A

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre-Course Composition</th>
<th>Post-Course Composition</th>
<th>t-value</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>W</td>
<td>108.55</td>
<td>31.66</td>
<td>154.91</td>
<td>26.64</td>
</tr>
<tr>
<td>C</td>
<td>15.14</td>
<td>4.30</td>
<td>22.91</td>
<td>4.35</td>
</tr>
<tr>
<td>TU</td>
<td>8.68</td>
<td>2.32</td>
<td>13.82</td>
<td>3.11</td>
</tr>
<tr>
<td>TU/S</td>
<td>1.06</td>
<td>0.18</td>
<td>1.84</td>
<td>0.52</td>
</tr>
<tr>
<td>C/TU</td>
<td>1.76</td>
<td>0.26</td>
<td>1.70</td>
<td>0.32</td>
</tr>
<tr>
<td>W/TU</td>
<td>12.67</td>
<td>2.65</td>
<td>11.53</td>
<td>2.12</td>
</tr>
<tr>
<td>SN/TU</td>
<td>2.67</td>
<td>0.69</td>
<td>2.26</td>
<td>0.48</td>
</tr>
<tr>
<td>W/C</td>
<td>7.23</td>
<td>1.26</td>
<td>6.83</td>
<td>0.89</td>
</tr>
<tr>
<td>G</td>
<td>5.54</td>
<td>0.77</td>
<td>6.38</td>
<td>0.61</td>
</tr>
</tbody>
</table>

$df = 21, \ p < .0056$

Table 4

Pre- and Post-Course Compositions for Group B

<table>
<thead>
<tr>
<th>Measure</th>
<th>Pre-Course Composition</th>
<th>Post-Course Composition</th>
<th>t-value</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>W</td>
<td>171.38</td>
<td>40.63</td>
<td>187.28</td>
<td>45.52</td>
</tr>
<tr>
<td>C</td>
<td>22.48</td>
<td>5.31</td>
<td>24.24</td>
<td>7.20</td>
</tr>
<tr>
<td>TU</td>
<td>13.55</td>
<td>3.09</td>
<td>15.45</td>
<td>4.01</td>
</tr>
<tr>
<td>TU/S</td>
<td>1.13</td>
<td>0.17</td>
<td>1.12</td>
<td>0.11</td>
</tr>
<tr>
<td>C/TU</td>
<td>1.68</td>
<td>0.29</td>
<td>1.58</td>
<td>0.28</td>
</tr>
<tr>
<td>W/TU</td>
<td>12.71</td>
<td>1.92</td>
<td>12.37</td>
<td>2.34</td>
</tr>
<tr>
<td>SN/TU</td>
<td>2.31</td>
<td>0.36</td>
<td>2.44</td>
<td>0.47</td>
</tr>
<tr>
<td>W/C</td>
<td>7.66</td>
<td>0.97</td>
<td>7.87</td>
<td>1.06</td>
</tr>
<tr>
<td>G</td>
<td>6.73</td>
<td>0.68</td>
<td>6.91</td>
<td>0.79</td>
</tr>
</tbody>
</table>

$df = 28$

In summary, Group A significantly increased fluency and lexical complexity measures in their post-course compositions, whereas Group B made no substantial gains in any measure used in the present analysis. Although there was no common indicator of writing development between the two groups, both groups gained raw fluency measures and neither group...
advanced significantly in syntactic complexity measures involving general complexity (W/TU, SN/TU), clausal subordination (C/TU), and sub-clausal or phrasal complexity (W/C).

RQ2: Between-Group Comparisons of Pre- and Post-Course Compositions

The means and SDs and the results of independent t-tests of pre-course compositions written by Groups A and B are shown in Table 5. The results reveal there were significant differences in all fluency measures (W, C, and TU) and the lexical complexity measure (G), suggesting fluency and lexical complexity differentiated between the two groups’ writing at the pre-course stage. All four significant results (W, C, TU, and G) generate effect sizes which are large (r > .6) (see Table 5). Group B produced significantly more words, clauses, and T-units, with more lexical complexity, than Group A. At the start of the course, Group B wrote over 62 words on average more than Group A. On the other hand, there were no significant differences in any syntactic measures that compared the length of linguistic units, such as T-units and clauses, or complexity by subordination and coordination. Moreover, significant differences in the raw fluency and lexical measures did not correspond to any variation in overall writing quality as evaluated by the external raters. As reported in the Data section, there was no significant difference in writing quality between the two groups’ pre-course compositions (recall Table 2).

Table 6 shows means and SDs along with the results of independent t-tests on the post-course compositions. Group B still wrote significantly more than Group A, but not as markedly as in the pre-course compositions. The results of t-tests showed the differences between the two were small enough not to be significant in the other measures that were significantly different in the pre-course compositions: the numbers of syntactic units, C and TU, and G. Gaining significantly in terms of the numbers of C, TU, and G (refer to the RQ1 section), Group A caught up with Group B in these measures. For example, Group A’s number of clauses in post-course compositions was not significantly different from that of Group B.

Table 5
Pre-Course Compositions: Group A vs. Group B

<table>
<thead>
<tr>
<th>Measure</th>
<th>Group A (n = 22)</th>
<th>Group B (n = 29)</th>
<th>t-value</th>
<th>Effect size r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>W</td>
<td>108.55</td>
<td>31.66</td>
<td>171.38</td>
<td>40.63</td>
</tr>
<tr>
<td>C</td>
<td>15.14</td>
<td>4.30</td>
<td>22.48</td>
<td>5.31</td>
</tr>
<tr>
<td>TU</td>
<td>8.68</td>
<td>2.32</td>
<td>13.55</td>
<td>3.09</td>
</tr>
<tr>
<td>TU/S</td>
<td>1.06</td>
<td>0.18</td>
<td>1.13</td>
<td>0.17</td>
</tr>
<tr>
<td>C/TU</td>
<td>1.76</td>
<td>0.26</td>
<td>1.68</td>
<td>0.29</td>
</tr>
<tr>
<td>W/TU</td>
<td>12.67</td>
<td>2.65</td>
<td>12.71</td>
<td>1.92</td>
</tr>
<tr>
<td>SN/TU</td>
<td>2.67</td>
<td>0.69</td>
<td>2.31</td>
<td>0.36</td>
</tr>
<tr>
<td>W/C</td>
<td>7.23</td>
<td>1.26</td>
<td>7.66</td>
<td>0.97</td>
</tr>
<tr>
<td>G</td>
<td>5.54</td>
<td>0.77</td>
<td>6.73</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Significant differences between the two groups’ post-course compositions were found in three measures: W, TU/S, and W/C. Except TU/S, the effect sizes are medium (r>.3) for the differences in W and W/C. Group B still outperformed Group A in terms of overall composition length. On average, Group B wrote 32 more words than Group A (recall Group B wrote 62 more words in the pre-course compositions). It is noteworthy that there was a significant difference in the two syntactic complexity measures with other than TU in the denominator (TU/S and W/C), whereas there were no significant differences in C/TU, W/TU, and SN/TU (reflecting clausal subordination and general syntactic complexity). In TU/S (reflecting clausal coordination), Group A outperformed Group B (M_A = 1.84 vs. M_B = 1.12), meaning the less experienced group used significantly more clausal coordination. This finding should be interpreted with caution because the results of Levene’s test showed the variances of TU/S for the two groups were not the same, thus violating the assumption of equal variance, and TU/S is considered to have a predictive power mainly at the beginning level as mentioned previously.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Group A (n = 22) M</th>
<th>SD</th>
<th>Group B (n = 29) M</th>
<th>SD</th>
<th>t-value</th>
<th>Effect size r</th>
</tr>
</thead>
<tbody>
<tr>
<td>W</td>
<td>154.91</td>
<td>26.64</td>
<td>187.28</td>
<td>45.52</td>
<td>2.97*</td>
<td>.39</td>
</tr>
<tr>
<td>C</td>
<td>22.91</td>
<td>4.35</td>
<td>24.24</td>
<td>7.20</td>
<td>0.77</td>
<td>.11</td>
</tr>
<tr>
<td>TU</td>
<td>13.82</td>
<td>3.11</td>
<td>15.45</td>
<td>4.01</td>
<td>1.58</td>
<td>.22</td>
</tr>
<tr>
<td>TU/S</td>
<td>1.84</td>
<td>0.52</td>
<td>1.12</td>
<td>0.11</td>
<td>6.37*</td>
<td>.80</td>
</tr>
<tr>
<td>C/TU</td>
<td>1.70</td>
<td>0.32</td>
<td>1.58</td>
<td>0.28</td>
<td>1.42</td>
<td>.20</td>
</tr>
<tr>
<td>W/TU</td>
<td>11.53</td>
<td>2.12</td>
<td>12.37</td>
<td>2.34</td>
<td>1.32</td>
<td>.18</td>
</tr>
<tr>
<td>SN/TU</td>
<td>2.26</td>
<td>0.48</td>
<td>2.44</td>
<td>0.47</td>
<td>1.36</td>
<td>.19</td>
</tr>
<tr>
<td>W/C</td>
<td>6.83</td>
<td>0.89</td>
<td>7.87</td>
<td>1.06</td>
<td>3.69*</td>
<td>.46</td>
</tr>
<tr>
<td>G</td>
<td>6.38</td>
<td>0.61</td>
<td>6.91</td>
<td>0.79</td>
<td>2.60</td>
<td>.35</td>
</tr>
</tbody>
</table>

\(df = 49\), *p < .0056, TU/S \(df = 22.5\)

In contrast to TU/S, Group B outperformed Group A in W/C (measuring sub-clausal or phrasal complexity). Although Group B’s W/C did not differ significantly between pre- and post-course compositions, their W/C ratio in post-course compositions significantly surpassed that of Group A. With these intermediate-proficiency EFL students, then, clausal length might be more effective to discriminate close proficiency levels than T-unit-related measures. As pointed out by Ortega (2003), slightly more than one word difference per clause (\(M_A = 6.83\) vs. \(M_B = 7.87\)) can be considered a between-proficiency level difference. Notably, Group A’s number of clauses substantially increased, whereas their W/C decreased in post-course compositions.

**Summary and Implications**

Based on the results of the study, implications can be drawn for indices of EFL students’
short-term writing development. First, the findings suggest that in the case of inexperienced intermediate-level EFL students, fluency and lexical complexity improve immediately in a short period such as a semester course. In the pre-course compositions, the more experienced group significantly outperformed the less experienced group in terms of all the raw fluency measures and the lexical complexity measure. In other words, the difference in the two groups’ prior English writing experience/instruction was reflected in their writing fluency and lexical complexity. In the post-course compositions, the experienced group still outperformed the less experienced group quantitatively, yet the differences between the two groups decreased. After receiving writing instruction and accumulating writing experience for a semester, it appears the students with less prior writing experience made substantial improvements, and fluency and lexical complexity were indicators of such short-term progress. The findings suggest fluency and lexical complexity may develop in parallel in the early stages of EFL writing, and the writing instruction is effective with this novice EFL writer group.

On the other hand, the non-significant findings in all measures in the case of the more experienced group’s pre- and post-course compositions suggest that, instead of increasing continuously, fluency and lexical complexity seem to level off. Although further studies that employ more measures would be necessary to be conclusive, results indicate the possibility that development of fluency and lexical complexity is not linear. Alternatively, the present non-significant results may also suggest a limitation of the writing instruction for this more experienced EFL writer group, who may require a different kind of writing instruction to develop further. Students would likely profit from increased English writing experience as a means of self-expression in the longer term. At the same time, students could benefit from writing instruction incorporating other components in addition to the knowledge about and experience of English expository writing. Genre-based tasks, for example, may be another component necessary for writing development, as found effective with novice Japanese EFL writers in Yasuda (2011). While improving writing, the students raised awareness of genre including purpose, writer role, and audience. In such argumentative writing tasks as used in this study, writers need to clearly articulate their points of view. Genre-specific writing instruction would eventually be necessary for those students to develop their understanding of the importance of purpose, context, and audience in their writing.

The present findings also indicate that fluency precedes complexity, especially syntactic complexity, in EFL writing development. They suggest that improvement in fluency emerges immediately once students start accumulating English expository writing experience, whereas it may take longer for those who have previously become accustomed to writing. Further research should consider how students develop in fluency over a longer period of time. In the present study, the B group’s post-course composition length \( (M = 187.28, SD = 45.52, \text{Range } 85 – 323) \) implied most students had not reached an acceptable level of writing fluency yet. In this type of argumentative writing, a length of 200–300 words has been considered acceptable for ESL compositions written in 30 minutes (Jacobs et al., 1981), and similarly a minimum of 300 words is presented for the Independent Writing Task in TOEFL iBT (Educational Testing Service, 2006). The more experienced group of the present study, as well as the less experienced group, would need to increase writing fluency to achieve the
length required by these standards. Or it could be argued these benchmarks may be too high or unrealistic for these EFL students. If so, it should be determined what a realistic and acceptable level of writing fluency, or quantity, may be for them.

Distinct from fluency and lexical complexity, syntactic complexity remained unchanged in the short term. The results did not show a marked change in this regard for either group of students. Improvement in syntactic complexity seems to take more time than a semester, supporting Ortega’s (2003) conclusion that “roughly a year of college-level instruction” (p. 492) or more than a year is necessary to observe substantial changes in syntactic complexity of EFL writing. The non-significant results in terms of syntactic complexity may also raise questions such as what students would need in order to develop complexity, whether they can develop syntactic complexity naturally if they accumulate more writing practice, or whether they require intervention with a different instructional focus to improve syntactic complexity.

The present findings differ from those of Bulté and Housen (2014), who found significant improvements in syntactic complexity measures at all levels (sentential, clausal, and phrasal) including W/TU and W/C, two measures used in the present study too. In contrast, they found no significant changes in lexical complexity measures including G, whereas the current study found G significant. These contrastive findings between these two studies should be interpreted with caution. Bulté and Housen (2014) employed more measures of syntactic and lexical complexity, and their participants were ESL students including advanced-level students as opposed to the intermediate-level Japanese EFL students in the present study. Considering such participant differences as proficiency-level and English writing experiences, a simple comparison would be misleading. However, it should be noted that taken together these two studies suggest that “lexical complexity and syntactic complexity do not develop in parallel” (Bulté & Housen, 2014, p. 53), implying lexical complexity and syntactic complexity are independent. This implication may give support to Skehan’s (2009) suggestion, based on comparisons between native and non-native speakers’ spoken data, that for non-native speakers, lexical complexity and syntactic complexity “do not seem to be integrated so well” (p. 528), unlike the way they are for native speakers. The relationship between lexical complexity and syntactic complexity is in need of further investigation, in which the range of lexical measures needs to be extended to cover lexical density and lexical sophistication.

The significant finding about W/C being a discriminator between the two groups’ post-course compositions may indicate that change (or improvement) in syntactic complexity occurs first at the sub-clausal or phrasal level. Examining low-proficiency Japanese EFL students’ narrative writing, Ishikawa’s (1995) findings supported the converging claim “that T-unit analysis does not work well for low-proficiency data” (p. 63), thus calling for higher-proficiency data for the T-unit analysis to be effective. The fact that no significant changes were found in the measures with the T-unit in the denominator in the present study seems to add to the claim that T-unit analysis does not work well with EFL intermediate-proficiency level data either.

The study points to several directions for future research. First, it is necessary to search further for indicators of student writing development over a longer period than one semester. Longer-term L2 writing development is in need of further investigation not solely from diverse perspectives but also with more in-depth analyses. In such a longitudinal study, how and when
syntactic complexity improves should be examined not independently, but in relation to accuracy and fluency. Focus on individual students' English writing development should also be made through in-depth studies. Concerning the relation between complexity and accuracy, several recent studies investigating ESL students' short-term writing development (e.g., Bulté & Housen, 2014) provide support for the claim that accuracy follows complexity. Disentangling the relationship among multiple CAF measures remains a goal for future studies. Moreover, how the measures of CAF are related to each other should be further investigated with regard to writing quality.

Notes

1 The text-external measures resorting to word frequency lists were not used in this study. Although limited, text-internal measures were considered sufficient to capture lexical improvement over a semester-long course.

2 The students of Group B were from two intact classes and participated in a previous study (Hirose, 2012b). In addition to sharing prior English writing backgrounds, they were comparable in many ways. Between the two subgroups, there were no significant differences either in the pre-course English composition scores/total words or post-course English composition scores/total words. Thus, the two subgroups’ data were combined for further analysis in the present study.

3 Some may question why two groups with different writing background received similar writing instruction, or more specifically, why Group B (those with prior instruction) deserved the paragraph-focused instruction. It is not unusual in Japanese EFL classrooms for their prior instruction not to solely target writing. Unlike the present writing instruction with a focus on writing, Group B had received integrative content that combined reading, discussion, and writing. Furthermore, it was Group B who held substantially more positive views of the instruction. In the post-course questionnaire, more students in Group B thought that the instruction had positive effects on their English abilities.

4 In the following class, the students received both their compositions and partners' peer feedback sheets with written teacher feedback. Teacher feedback included both holistic evaluative comments such as “very good,” or “good comments” but also such specific comments as “this needs examples,” or “I agree.” It was meant to complement peer feedback, compensating for its shortcomings, and grammar and spelling mistakes missed by peers were highlighted or corrected, and confusing parts were identified with question marks.

5 Group A wrote about “English learning and studying abroad” for pre-course compositions, and then “university students and part-time jobs” for post-course compositions, whereas Group B wrote in the opposite order. The two topics were not counterbalanced within the same group. Because both pre- and post-course writing sessions were administered during class hours, it was not feasible to ask half of each class to work on one topic while the other half worked on the other topic. These two topics were considered comparable and not significantly influential on the quality and quantity of student compositions for the following reason. The same two topics were used for two groups of Japanese students with similar English proficiency and writing background as Group B of the present study (Hirose, 2012a). One group wrote on one topic, whereas the other group wrote on the other topic. The
previous study found no significant difference between the English compositions these two groups wrote, in terms of writing quality (=total scores) or quantity (=total number of words). Composition scores were used only to check whether there were significant differences between pre- and post-course compositions of the two groups, respectively, and whether there were significant differences in either pre- or post-course compositions between the two groups.

Acknowledgements
The research reported in this paper was supported by JSPS KAKENHI Grant Number JP23520685. The author is very grateful to the three anonymous reviewers for their helpful comments and suggestions on an earlier version of this paper.

References


Yamanishi, H. (2004). Kokôsei no jiyûeisakubun wa donoyôni hyôka sareteirunoka — bunsekiteki hyôkashakudo to sôgôteki hyôkashakudo no hikaku wo tósiten no kentô —
[How are high school students’ free compositions evaluated by teachers and teacher candidates?: A comparative analysis between analytic and holistic rating scales]. *JALT Journal, 26*, 189–205.


Appendix A: Pre- and Post-Course Composition Prompts

In the readers’ column in an English newspaper, there has been a heated discussion about the issue of “university students and part-time jobs.” Some people think that students should not have part-time jobs, whereas others believe they should work part-time. Now the editor of the newspaper is calling for the readers’ opinions. Suppose you are writing for the readers’ opinion column. Take one of the positions described above, and write your opinion.

In the readers’ column in an English newspaper, there has been a heated discussion about the issue of “English learning and studying abroad.” Some think that people have to study abroad to improve their English, whereas others believe people can improve their English in Japan and don’t need to study abroad. Now the editor of the newspaper is calling for the readers’ opinions. Suppose you are writing for the readers’ opinion column. Take one of the positions described above, and write your opinion.

Appendix B: A Student’s Sample Compositions

Pre-Course Composition (Study Abroad):
I think we don’t have to study abroad in order to improve English abilities only. Because there are many foreign people who can speak English and many English schools in Japan. “Ekimae Ryugaku” is cheaper than studying abroad. But if you want to get experience, studying abroad may be better than “Ekimae Ryugaku.” The ways has a good point. SO I don’t know which is a good way.
If I study abroad, I want to go England. Because England has kings English and tradition.

Post-Course Composition (Part-Time Jobs):
I think university students should have part-time jobs. I also have part-time job. The reason why I have part-time is for my studying. My parents allowed me to go to university if I pay something I need, for example, my cell phone’s fee, a ticket for going to school, my cloth and so on. As I have to need money for these things, I agree with having part-time jobs. However, if I work hard, my grade may be low. I think people who have part-time jobs should also study. When our studying is regarded as good grade before having part-time jobs and after having them we can go on getting good grade, people who think the university students should have part-time jobs may admit to have them.