Intermediate-Level Japanese EFL Students' Short-Term Writing Development: Examining Fluency and Complexity in Relation to Quality

Keiko HIROSE

Abstract

This paper examines whether and how intermediate-level Japanese EFL students change (or improve) their writing during a semester-long writing course. First, to examine short-term writing development, it compares pre- and post-course compositions in light of both fluency and complexity measures and writing quality ratings. Next, it investigates which measures contribute to overall writing quality ratings. Third, relationships among the 11 measures of writing fluency and complexity are examined. Multiple regression analysis reveals that 2 measures, total number of words and the Guiraud index (*G*), significantly explain total scores, whereas no syntactic complexity measures do so. Subsequently, exploratory factor analysis provides a three-factor solution indicating different loadings for fluency and lexical and syntactic complexity. The findings suggest that (a) fluency and *G* develop in parallel, whereas syntactic complexity changes little; (b) no syntactic complexity measures significantly contribute to the total scores; (c) lexical and syntactic complexity measures do not load on the same factor, indicating independent relations from each other; (d) *G* is partially related to fluency, whereas no other lexical measures are; and (e) syntactic complexity measures for different linguistic units (T-unit, clause, and sentence) show distinct patterns from one another. Lastly, several directions for future research are indicated.

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

Introduction

This study is oriented to longitudinal L2 writing studies that search for indices that capture L2 writing development (Bulté & Housen, 2014; Celaya & Navés, 2009). The previous studies have commonly examined linguistic changes in terms of complexity, accuracy, and fluency (CAF) (Wolfe-Quintero, Inagaki, & Kim, 1998). The present study examines intermediate-level Japanese EFL students' short-term writing development based on linguistic features of students' written products, particularly from the perspectives of fluency and linguistic complexity in relation to writing quality.

To date, many longitudinal studies examining L2 writing development have been conducted over a short term, such as several months (Ishikawa, 1995). A short course of one semester has been considered insufficient to capture substantial improvement in syntactic complexity, as Ortega (2003), who reviewed and synthesized the findings of such previous studies, found "a negligible to small-sized change" (p. 511). However, recent studies examining advanced-level ESL students' writing development in one semester have found noticeable improvements in syntactic complexity (Bulté & Housen, 2014; Connor-Linton & Polio, 2014). Although changes or improvements made by EFL students may be different from, and smaller than, those of ESL students (Ortega, 2003), Casanave (1994) found that positive changes in Japanese EFL students' writing emerged in the first of the three semesters she examined. Her empirical study, although of a small scale, reported upward changes (i.e., longer, more complex, and more accurate writing). Thus, a study of one-semester-long writing development can be regarded as worthy of being conducted. Lower proficiency-level EFL students' writing development is also in need of investigation.

In this study, writing development is approached with the assumption that an increase or a decrease in either a fluency or a complexity measure cannot be considered either a positive or a negative change in its own right. It has been shown that the relationship between fluency and complexity of a student

written product is not straightforward, and the two are sometimes even inversely related. A trade-off may exist between fluency and complexity, like the one between complexity and accuracy (Skehan, 2009). As reported by Larsen-Freeman (1978), for example, although the number of words per composition increased with student proficiency, it decreased over time within the highest proficiency student group. This tendency toward decreasing fluency might be related to the development of syntactic complexity. At the same time, such a seemingly non-linear reverse relationship has also been noticed among different aspects of syntactic complexity itself. A decrease in one aspect may lead to an increase in another. More specifically, some previous studies proposed that developmental patterns "move from coordination to subordination to the reduction of clauses to phrases" (Wolfe-Quintero et al., 1998, p. 73).

Thus, multiple measures should be employed to fully embrace student writing development in longitudinal research. Norris and Ortega (2009) have called for studies with factor analytic designs. Similarly, drawing on the findings of existing literature, Wolfe-Quintero et al. (1998) suggest 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). It is, therefore, important for a longitudinal study of L2 writing development to capture changes or development by adopting multiple measures to examine interrelationships among fluency and complexity measures. Disentangling fluency and complexity was attempted in this study by performing a factor analysis.

Not only the relations among linguistic measures, but also those between these measures and writing quality are in need of investigation. Past studies have found that more complex syntax does not necessarily directly lead to better quality writing (Crossley & McNamara, 2014; Yang, Lu, & Weigle, 2015). Fluency and complexity are both "multifaceted and multidimensional concepts" (Housen & Kuiken, 2009, p. 464). They need to be operationally defined. In the following sections, my conceptualizations of fluency and of syntactic and lexical complexity are specified, and measures employed in this study are explained. After writing quality is clarified, the research gap to be

filled by this study is presented.

Fluency

Writing fluency needs to be unraveled as a construct, because "there is no agreed-upon definition" (Abdel Latif, 2013, p. 99). As pointed out by Pallotti (2009), fluency is "a multidimensional construct, in which sub-dimensions can be recognized … Once it is established which of these sub-dimensions is at issue, it is in principle relatively transparent what is being measured" (pp. 591–592). The sub-dimensions meant by Pallotti are "breakdown (dys)fluency, indexed by pausing; repair (dys)fluency, indexed by measures such as reformulation, repetition, false starts, replacements; and speed, with measures such as syllables per minute" (Skehan, 2009, pp. 512–513). Although these sub-dimensions are primarily proposed for speaking fluency, fluency can be generally operationalized as comprising the two seemingly related aspects of dysfluency and speed. Lack of dysfluency features contributes to high speed. Studies that examined speaking fluency employed dysfluency measures (e.g., the percentage of pause times, the number of fillers per minute, or the number of reformulations per minute) and speed measures (e.g., the number of words per minute) (Sakuragi, 2011; Tavakoli & Skehan, 2005). The sub-dimensions of dysfluency and speed can also be applied to writing fluency. However, it is difficult to encapsulate both sub-dimensions without digging into the writing process. Past studies have traced the writing process by audio- or video-taped and think-aloud protocol data.

Writing has been found to be a highly cognitive process in which writers generate ideas related to a given topic; plan content and organization, both globally and locally; translate ideas into language; and review, in a recursive manner (see Flower & Hayes, 1981, for their influential L1 writing model). Because attentional resources are finite, writers have limited attention to pay to many aspects of writing simultaneously. Low-level and high-level processes "may compete for mental resources" (Bereiter & Scardamalia, 1987, p. 95). For example, attention to low-level concerns about mechanics may interfere with high-level planning of organization. Comparing L1 and L2 writing processes using think-aloud protocols in a within-subject design,

Whalen and Ménard (1995) revealed that for the L1 task, low-level linguistic-level processing accounted for half, whereas for the L2, it comprised as much as 78%. While putting the generated ideas into L2, students, especially those with lower L2 proficiency levels, are likely to encounter difficulty with "both linguistic knowledge (vocabulary, grammar, and orthography) and fluency or accessibility of linguistic knowledge (lexical retrieval and sentence building)" (Schoonen, van Gelderen, de Glopper, Hulstijn, Simis, Snellings, & Stevenson, 2003, p. 175). Thus, writing fluency is partially related to the facility with which they can do such lexical retrieval and syntactic processing.

Pause analysis sheds light on what hinders writers from writing fluently, that is, without pauses. Pauses can occur in every stage of the process: planning, translating, and reviewing. A retrospective think-aloud protocol analysis of pausing while writing showed that lower-proficiency Japanese EFL students often paused to do lexical searching or syntactic processing while attempting to translate their generated ideas into English, resulting in smaller amounts of production than their higher-level counterparts: 132.8 vs. 168.0 words; 4.35 vs. 6.14 words per minute (WPM), respectively (Hirose, 2005).[1] In that comparative study, the lower group paused more and produced fewer words, whereas the higher group paused less to produce longer compositions with higher quality. These contrastive results were partly derived from different degrees of automaticity or fluency of linguistic processing between the two groups, suggesting that the higher group's writing with fewer pauses derived from more fluent retrieval of words and structures. Therefore, the number of pause times during writing was found to be in inverse relation to the length of produced text. In other words, fewer dysfluent features led to higher writing speed.

Is there any yardstick available by which to assess writing speed? In terms of the total words produced in 30 minutes, the 200–300 word range is considered an acceptable length for ESL compositions (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981), and the Independent Writing Task in TOEFL iBT, which "requires writing an essay that states, explains, and supports the writer's opinion on a given issue," presents a benchmark of a minimum of 300 words in 30 minutes (Educational Testing Service, 2012,

p. 39). The TOEIC writing test has an opinion essay type, which has the same minimum word length (Educational Testing Service, 2007). Dividing the word range 200–300 by 30 minutes results in 6.67–10 WPM. Against these international benchmarks, most Japanese EFL students' writing fluency studied previously had not yet reached an acceptable level. Mizumoto (2008), for example, had Japanese university students with lower English proficiency write an argumentative composition under a 60-minute condition. Although the students were asked to write more than 150 words, they produced a mean total of 147.7 words (slightly less than 2.5 WPM). Although the average estimate of 2.5 WPM should be treated with caution, given an argumentative topic in timed writing conditions, these Japanese EFL students produced far less than the recommended 6.67–10 WPM, indicating a severe lack of writing fluency.

In this study, the speed sub-dimension of writing fluency was examined by dividing the total amount of text produced by the writing time, and the dysfluency aspects were not directly captured for the following reasons. First, the amount of production was considered a valid gauge of fluency within the same time limit. In addition to the total number of words, lengths of clauses and T-units (minimal terminable unit) were employed. Writing fluency is not obviously a word-level construct, and adopting production units longer than a word is regarded as important (note that Abdel Latif, 2013, recommended the length of translating episodes as a measure of writing fluency). Second, although incorporating dysfluency features into the analysis by counting the number of production units written "without interruption, or without engaging in other activities such as reading back or rehearsing" (Raimes, 1985, p. 243) would have been more desirable, a relatively large sample (N = 138) made it difficult to collect think-aloud protocols in the non-computer classroom situation (recall Raimes collected protocols from eight students). Third, the participants wrote compositions with pencil and paper. The recent advancement of technology has made it possible to examine the writing process by the use of an online measure (Schoonen, Snellings, Stevenson, & van Gelderen, 2009) and keystroke logging methods on a computer (Leijten & Van Waes, 2013). It should be noted that Japanese university students,

whose L1 has an agglutinating morphology, are not expected to be adept enough to keystroke their English writing on a computer.[2] Students' lack of familiarity with keystroking in English would have caused another obstacle to fluent writing, which made it questionable to use an online method in this study.

Inclusion of writing quality in the present analysis made it possible to examine writing fluency manifested in the amount of production in relation to writing quality. The relation between writing fluency and quality is complex because higher writing speed may lead to larger quantity, but not necessarily to higher writing quality (Bereiter & Scardamalia, 1987).

Syntactic Complexity

Like fluency, syntactic complexity is a multi-dimensional construct with interrelated sub-constructs. Its sub-constructs have been identified and measured at the sentential, clausal, and phrasal levels (Bulté & Housen, 2012). To grasp such 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 noted 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) also 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: (a) general or overall complexity, (b) complexity via subordination, (c) sub-clausal complexity via phrasal elaboration, and (d) complexity via coordination in cases where low-proficiency level data are included. Given impetus from their call for such research, this study drew on their recommendations to adopt multiple measures to examine their interrelationships.

Accordingly, this study used five measures to grasp these multidimensions of syntactic complexity. First, to capture general complexity, the number

of words per T-unit (W/TU) and the number of S-nodes per T-unit (SN/TU) were employed. W/TU has been the most commonly employed in past studies, and in fact, it was the only measure employed in the L2 longitudinal writing studies Ortega (2003) surveyed that examined L2 students' linguistic development. SN/TU was also added in the present analysis 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 (Ishikawa, 2006; Yamanishi, 2011). SN 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 (Storch, 2005; Yang, Lu, & Weigle, 2015). Third, the number of words per clause (W/C) was used to reflect complexity via sub-clausal or phrasal elaboration. As argued 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 in this study as constituting a clause, in accordance with other previous studies (Ishikawa, 1995; Yang et al., 2015). Lastly, the number of T-units per sentence (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, TU/S is considered discerning for beginning levels of L2 development (Norris & Ortega, 2009).

Phrasal-level syntactic complexity was not embraced in this study for the

following reasons. First, phrasal complexity is considered a sign of later development. Previous research shows that with increasing proficiency, students are inclined to use complex phrases such as nominalizations and modification, instead of coordination or subordination. Thus, a pure phrasal complexity measure such as the length of noun phrases may be regarded as most suitable to analyze advanced-level students' writing (Norris & Ortega, 2009), as targeted by Bulté and Housen (2014). With intermediate-level EFL students' writing, this study adopted measures considered to be more sensitive for lower proficiency-level students such as coordination and subordination, as well as one clausal level measure (W/C). Another reason for not including phrasal-level analysis was that phrasal-level coding is difficult to perform manually. Furthermore, although many studies employed computer programs such as 'Coh-Metrix' and 'L2 Syntactic Complexity Analyzer,' comparisons between coding by means of such computer programs and by human hands found the agreement for phrasal-level analyses is lower than that for clause and T-unit level analyses (Lu, 2010; Polio & Yoon, 2018).

Lexical Complexity

Just like syntactic complexity, lexical complexity also encompasses multiple sub-dimensions. There are numerous measures available to gauge various sub-dimensions of lexical complexity (Bulté & Housen, 2012). This study attempted to capture three aspects of lexical complexity by employing measures of diversity, density, and sophistication.

Lexical complexity measures are generally 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). Text-internal measures include type-token ratio (TTR) (word types/word tokens). Because the TTR is criticized for being affected by text length, the Guiraud index (*G*) (word types/√word tokens) was used to compensate for the influence of the length of texts. To measure lexical diversity this study employed *G* instead of the Diversity index (Malvern, Richards, Chipere, & Durán, 2004) following Bulté and Housen's (2014) observation. Rather than

an index of sheer diversity, Bulté and Housen considered "*G*, which captures both diversity and productivity, …, especially for the analysis of timed writing samples that are not controlled for length" (pp. 49–50) as in the data collected in this study.

In addition to *G*, another text-internal measure different from TTR-based indices was used. As a measure of lexical density, the number of content words divided by the total number of words (ConW/W) was calculated. The composition is considered dense if it has many content words in relation to the total number of words (Laufer & Nation, 1995).

Besides the two text-internal measures (*G* and ConW/W), one text-external measure was used for analysis. Among available measures, P_Lex was employed as an index of lexical sophistication (Meara & Bell, 2001). P_Lex was chosen over the Lexical Frequency Profile (Laufer & Nation, 1995), another well-known measure of lexical sophistication, because the former "works best with texts that are not longer than 300 words" (Meara & Miralpeix, 2017, p. 45), whereas the latter recommends texts of "at least 300 words long" (Nation, 2008, p. 84) for its application and same-length texts for comparison purpose. Almost all of the present data were short texts of less than 300 words and their lengths differed (see Table 6). In fact, there was only one text longer than 300 words. By running a P_Lex program, the lambda values (λ), which "typically range from 0 to about 4.5, with higher figures corresponding to a higher proportion of infrequent words" (Meara & Bell, 2001, p. 11) were obtained. The higher the lambda values, the lower the frequency of the words that were used.

Objective Measures Employed for the Present Analysis

Table 1 lists the 11 measures employed in this study. Simple counts of words (W), clauses (C), and T-units (TU) were employed to measure fluency because they were considered valid for tapping writing fluency in timed writing. These measures have been used to measure fluency in past studies: W (Storch, 2005), C (Robb, Ross, & Shortreed, 1986), and TU (Ishikawa, 1995). TU is "one main clause plus whatever subordinate clauses are attached to that main clause" (Hunt, 1966, p. 737). Regarding syntactic complexity,

five indices were adopted to capture its multi-dimensionality (Bulté & Housen, 2012; Norris & Ortega, 2009). Lexical complexity was measured in three different ways, as previously explained.

Among these measures, W/TU and W/C have been controversial as to whether they measure fluency or syntactic complexity (Sakuragi, 2011). For example, W/TU has been used as a fluency measure in some studies (Ishikawa, 2006; Yamanishi, 2011), as has W/C (Celaya & Navés, 2009). Wolfe-Quintero et al. (1998) listed both of them as the most appropriate measures of fluency, whereas others such as Ortega (2003) have considered them among the most frequently used syntactic complexity measures. The previous studies using these measures have treated them as either type of measure. Thus, this issue is also addressed in the present analysis.

Table 1
A List of Linguistic Measures Used in the Present Study

1. *Fluency*
   W, TU, C
2. *Syntactic complexity*
   overall complexity: W/TU, SN/TU
   complexity by subordination (finite clausal subordination): C/TU
   complexity by coordination (clausal coordination): TU/S
   complexity via sub-clausal or phrasal elaboration: W/C
3. *Lexical complexity*
   diversity: *G*
   density: ConW/W
   sophistication: P_Lex

Writing Quality

In order to measure writing quality, an analytic scale consisting of the five criteria of content, organization, language use, vocabulary, and mechanics was used (see **Appendix A**). The scale was an adapted version of Jacobs et

al.'s (1981) ESL Composition Profile (see the ***Composition rating*** section for details) that had been empirically validated (see Yamanishi, 2004). In this study, the total of the five subscores was operationalized to reflect overall writing quality.

The Present Study

This study attempted to identify linguistic measures to elucidate intermediate-level Japanese EFL students' writing development over a semester. For this purpose, the study compared the pre- and post-course compositions written by Japanese undergraduates who received one semester of English writing instruction. The comparison encompassed composition ratings as well as 11 linguistic measures. Such L2 writing research from a multi-dimensional perspective is worth conducting, particularly in light of the necessity of L2 writing intervention studies. Many intervention studies to date tend to use a limited number of measures to examine the instructional effects, although there are exceptions (Ishikawa, 1995), and the effects have not been conclusive. As urged by Connor-Linton and Polio (2014), intervention studies should employ multiple measures to investigate change over time. Furthermore, writing quality was not found to be consistently positive (Hirose & Sasaki, 2000), so it is important to include a quality component in intervention studies.

The present study addresses the following three research questions (RQs):

1. Do intermediate-level Japanese EFL students change or improve in their writing after a semester-long English writing course?
2. How is writing quality related to fluency and syntactic/lexical complexity?
3. How are fluency and syntactic/lexical complexity related to each other?

Method

Participants

The participants of the study were Japanese EFL university students (N = 69; 15 males and 54 females). Their English proficiency levels were mostly CEFR B1–B2, ranging from low- to high-intermediate. They were students in four intact English writing courses taught by the researcher (Class 1: n = 14; Class 2: n = 15; Class 3: n = 22; Class 4: n = 18). They all wrote English compositions in class at the outset and the end of the semester-long courses (hereafter pre-course and post-course composition, respectively). Their English writing levels did not differ significantly at the beginning of the courses. That is, the results of a non-parametric test showed no significant difference among them in the pre-course compositions. The compositions from the four classes of students were thus combined in the present analysis because they were considered comparable as writing samples drawn from the population of intermediate-level Japanese EFL university students with relatively little writing experience.

Content of Instruction

All the participants received English writing instruction that dealt with paragraph organization, such as *comparison/contrast* and *cause/effect* structures, and facilitated writing experience. Knowledge of English writing and beyond paragraph-level writing experience have been identified as among explanatory factors for Japanese EFL students' writing ability, and weak writers were found to lack both (Sasaki & Hirose, 1996). Thus, combining knowledge instruction with writing experience can be considered essential for them. The assumption is that gaining knowledge about English writing plus writing experience would lead to writing development.

In this instruction, writing experience was realized through a composition assignment of at least one paragraph and in-class peer feedback writing every class. Exactly the same class procedure was used in all the four courses. The classes met separately once a week for 90 minutes each over the course

of a 15-week semester. The first half of the class time was 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 first half was spent reading feedback and engaging in spoken feedback. The other half of the class was spent on English paragraph instruction. Students learned about English paragraphs by reading and analyzing sample paragraphs, and then outside of class they wrote compositions on their chosen topics related to a specific paragraph organization covered in a previous class.

Data

The English compositions 69 participants wrote before and after they took the courses were the major data sources for this study. All participants wrote on an argumentative topic, taking one of the two given positions and supporting it in 30 minutes in class. This type of task was chosen for several reasons. First, many studies that examine Japanese EFL students' writing have used such argumentative tasks (Kamimura, 2006; Sasaki & Hirose, 1996). Second, this type of writing is what participating students are expected to achieve and they themselves identify as such (recall TOEIC and TOEFL writing tasks). Most importantly, all the participants of this study received the writing instruction described previously, and they 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 B** for writing prompts).[3] Both topics were considered equally familiar to these students, most of whom were working part-time and had an interest in studying abroad. The students were not informed about the topics beforehand and were not allowed to use dictionaries.

Data Analysis

Because compositions were all hand-written in class, the 138 compositions were typed in word documents for further analysis.

Composition rating

All 138 compositions were scored by two English-speaking instructors with MAs in TESOL, according to an adapted version of Jacobs et al.'s (1981) ESL Composition Profile (Yamanishi, 2004). Unlike the original version, the adapted version has equal weighting (10 points each) for the five criteria of content, organization, language use, vocabulary, and mechanics (see **Appendix A** for descriptors). Although the rating descriptors are the same as the original, the ratings are given on a 1–10 scale (Poor 1–2; Fair 3–5; Good 6–8; Very good 9–10). In order to avoid a possible order effect, the raters scored the compositions in opposite orders. They were not told the same writers produced two compositions, not to mention the order in which they were written. The sum of the two raters' scores was used for the present analyses, with a possible range of 10–100. When the two raters' total scores differed by more than 5 points, the researcher resorted to a third rater who has a similar background to that of the two raters and used the two closest scores among the three according to Jacobs et al.'s recommended procedure. In this way, five analytical scores and a total score were obtained for each composition.

Quantitative linguistic analysis

This study partially relied on computer-based analyses. The words (tokens), types, and *G* were calculated using AntConc (Version 3.3.1; Anthony, 2012). Similarly, the lambda values (λ) were computed using P_Lex (Version 3.00; Meara & Miralpeix, 2017). Except for lexical analyses, the other measures such as C, TU, SN, and S were identified manually by two Japanese researchers/teachers of English. When there were discrepancies between the two, the researcher coded the data and resolved the disagreements through discussion until 100% agreement was reached. In counting, sentence fragments were not regarded as T-units. Thus, T-units could occur across periods (Ishikawa, 2006). Manual identification was used for syntactic analyses because human coding is considered desirable for analyzing L2 texts, especially those produced by lower-level students. As pointed out by Bulté and Housen (2014), computer-based analyses "may still be too rigid to

accurately and fully identify, segment, and parse the L2 learner productions" (p. 48).

Statistical analysis

For RQ 1, the pre- and post-course compositions were compared in terms of six ratings (five analytical and one overall) and 11 linguistic measure scores by using paired *t*-tests. Because multiple tests were employed, a Bonferroni adjustment was made (Tabachnick & Fidell, 2013). The alpha level was set to 0.0029 (i.e., 0.05/17) because overall 17 *t*-tests were conducted; thus, only those tests that resulted in values at or below the alpha level were accepted as significant.

For subsequent analyses addressing RQs 2 and 3, the pre- and post-course composition scores were combined (N= 138).[4] For RQ 2, a stepwise multiple regression analysis was conducted to investigate the (best) prediction of overall writing quality. For RQ3, to explore how each of the 11 linguistic measures is related to the others, Pearson correlation coefficients were calculated, and factor analysis was performed to investigate the interrelationships among them using SPSS version 22. Because no hypothesis was formed concerning the clusterings of the 11 measures, exploratory factor analysis was considered the most appropriate method for the present study.[5]

Results and Discussion

RQ1: Do intermediate-level Japanese EFL students change or improve in their writing after a semester-long English writing course?

For RQ1, the pre- and post-course composition scores were compared. As shown in Table 2, the interrater reliability for the total pre- and post-course composition scores was acceptably high (0.84 and 0.84, respectively). On the other hand, the reliability estimates for some subscores, especially mechanics, were relatively low.

Table 2
Interrater Reliability Estimates for Composition Scores

	Pre-Course Composition	Post-Course Composition
Total Score	0.84	0.84
Content	0.81	0.77
Organization	0.60	0.75
Vocabulary	0.64	0.76
Language Use	0.62	0.59
Mechanics	0.45	0.55

Note. Interrater reliability estimates are based on the coefficient alpha formula.

Table 3 shows the means and *SD*s of pre- and post-course composition total scores and subscores and the results of repeated *t*-tests. The results of the *t*-tests showed there were significant differences between the two in all rated measures, and the effect sizes were relatively large for the differences. The largest effect size was found for the total score ($r = .59$). After a 4-month writing instruction period, statistically significant changes were found for the total and all the subscores, although the difference in mechanics should be treated with caution because the reliability estimates were relatively low (see Table 2).

Table 4 shows the means and *SD*s of the 11 linguistic measure scores of pre- and post-course compositions. The results of repeated *t*-tests showed significant differences in four out of 11 measures. It should be noted that the effect sizes for these objective measures were mostly smaller than those of the subjective (evaluation) ratings. The scores on all the count measures (W, C, and TU) increased significantly. On average, students wrote 28 more words, 3.8 more clauses, and 3 more T-units in their post-course composition. *G* also increased significantly, whereas the lexical density ratio and lexical sophistication measure, which resorted to an external word frequency profile, did not. Thus, lexical improvement was made, but only partially.

Furthermore, no significant increases in syntactic measures were observed. The results of *t*-tests showed the differences were small enough not to be significant in all the ratio measures of syntactic units, TU, C, and S. As shown in Table 4, depending on the syntactic unit, the results yielded different tendencies. More specifically, all the scores on the measures with the TU in the denominator decreased, whereas other syntactic measures with either S or C in the denominator did not. Thus, despite a greater number of TUs, they were slightly shorter in length in the post-course compositions.

Table 3
Pre-Course vs. Post-Course Composition Scores

	Pre-Course Composition		Post-Course Composition			Effect Size
Measure	*M*	*SD*	*M*	*SD*	*t*	*r*
Total Score (100)	66.93	7.12	73.54	8.71	−6.01*	.59
Content (20)	13.52	2.02	15.16	2.08	−5.73*	.57
Organization (20)	13.39	1.93	14.59	2.23	−4.21*	.45
Vocabulary (20)	13.33	1.36	14.42	1.85	−4.42*	.47
Language Use (20)	13.12	1.39	14.22	1.81	−4.38*	.47
Mechanics (20)	13.77	1.46	14.81	1.68	−4.02*	.44

df = 68. *$p < .0029$.

Increased numbers of W, C, and TU produced within the given time may be a positive indicator of increased fluency. The significantly greater *G* may lend support to the increased vocabulary scores. Despite there being no significant changes in any syntactic measures, human rating scores of language use and mechanics improved significantly in post-course compositions. Therefore, there seems to be no trade-off relation between fluency and syntactic complexity. Fluency increased and syntactic complexity did not decrease, with increased writing quality.

Table 4
Pre-Course vs. Post-Course Compositions: 11 Measures

	Pre-Course Composition		Post-Course Composition			Effect Size
Measure	*M*	*SD*	*M*	*SD*	*t*	*r*
W	138.00	46.07	165.97	39.62	−5.32*	.54
C	18.62	5.91	22.43	5.92	−4.78*	.50
TU	11.00	3.51	14.06	3.51	−5.92*	.58
TU/S	1.09	0.16	1.12	0.13	−1.18	.14
C/TU	1.71	0.27	1.62	0.30	2.25	.26
W/TU	12.72	2.19	12.03	2.13	2.26	.26
SN/TU	2.52	0.54	2.36	0.46	1.92	.23
W/C	7.48	1.08	7.49	0.99	−0.11	.01
G	6.14	0.85	6.54	0.73	−3.75*	.41
ConW/W	0.50	0.04	0.49	0.03	1.22	.15
P_Lex	1.10	0.32	1.18	0.31	−1.66	.20

df = 68. *p < .0029.

RQ 2: How is writing quality related to fluency and syntactic/lexical complexity?

For multiple regression analysis, the total score was used as the dependent variable to reflect overall writing quality. As Table 5 shows, all subscores were highly correlated with the total score and among each other ($r > .7$; significant at $p < .01$). The interrater reliability for the total score was .86.

Table 5
Correlation Matrix for Composition Rating Scores

	1	2	3	4	5	6
1. Total Score	1.00					
2. Content	.90**	1.00				
3. Organization	.88**	.84**	1.00			
4. Vocabulary	.92**	.81**	.77**	1.00		
5. Language Use	.89**	.77**	.72**	.84**	1.00	
6. Mechanics	.87**	.74**	.73**	.80**	.77**	1.00

$N = 138$. $^{**}p < .01$ (two-tailed).

Descriptive statistics for the 12 measure scores

Table 6 presents descriptive statistics for the 12 scores (one overall quality and 11 linguistic measures) used for the analysis. The absolute values of skewness and kurtosis for these measures did not exceed two, except the kurtosis of C. Although the distribution of C was a little peaked, this was not considered a major problem for further analysis because “underestimates of variance associated with positive kurtosis (distribution with short, thick-tails) disappear with samples of 100 or more cases” (Tabachnick & Fidell, 2013, p. 80).

Table 6
Descriptive Statistics for the 12 Measures

Measure	*M*	*SD*	Range	Skewness	Kurtosis
Total Score	70.23	8.60	55 – 93	0.53	−0.34
W	152.41	44.92	56 – 323	0.46	0.85
C	20.53	6.20	9 – 47	0.89	2.11
TU	12.53	3.82	5 – 26	0.51	0.46
TU/S	1.11	0.15	0.70 – 1.67	0.70	1.85

Measure	*M*	*SD*	Range	Skewness	Kurtosis
C/TU	1.67	0.29	1.10 – 2.54	0.60	0.20
W/TU	12.37	2.18	8.88 – 18.57	0.69	−0.08
SN/TU	2.44	0.51	1.50 – 4.14	0.81	0.49
W/C	7.48	1.04	5.26 – 10.73	0.54	0.38
G	6.53	0.82	4.45 – 8.66	0.09	−0.17
ConW/W	0.49	0.03	0.41 – 0.57	0.003	−0.17
P_Lex	1.14	0.31	0.50 – 2.00	0.40	−0.25

$N = 138$.

Multiple regression analysis

A stepwise multiple linear regression analysis was performed to identify which linguistic measure or combination of measures best predicts the overall writing quality. For this analysis, the total score was regressed against all the 11 measures used for the present analysis (independent variables), yielding $F(2, 135) = 26.67$, $p < .001$. The adjusted coefficient of determination (henceforth, R^2) was 0.273. This indicates that 27.3% of the variance of the total score was explained by the model. The model reached statistical significance, although it did not account for a large percentage of the variance. According to Cohen (1988), this is a large effect size.

As Table 7 shows, the *t* statistics for the beta values of W and *G* were significant ($p < .05$), while the other nine measures did not significantly contribute to explaining the total score. The largest beta coefficient was .402 for W. This means W made the strongest contribution to explaining the total score. The beta value for *G* was lower (.184), indicating it made less of a contribution than W. As Step 1 yields, W alone explained 51.1% of the total score. W and G also had a significantly positive correlation ($r = .59$, see Table 8). Figure 1 is a path diagram illustrating an explanatory model of intermediate-level EFL writing based on the results of the present analysis. The results indicate that using a large number of words and many different words are indicators of higher overall writing quality, whereas none of the

syntactic measures are. It is noteworthy that the measures that significantly contributed to the writing quality (W and *G*) were among those that improved significantly over the semester.

Table 7
Stepwise Multiple Regression Analysis for the Total Composition Score

	Unstandardized coefficients		Standardized coefficients			
	B	Std. error	Beta	*t*	*p*	R^2 (adjusted R^2)
Step 1						.261 (.256)
Constant	55.33	2.24		24.70	.000	
W	.098	.014	.511	6.93	.000	
Step 2						.283 (.273)
Constant	46.23	4.99		9.27	.000	
W	.077	.017	.402	4.46	.000	
G	1.94	.95	.184	2.04	.044	

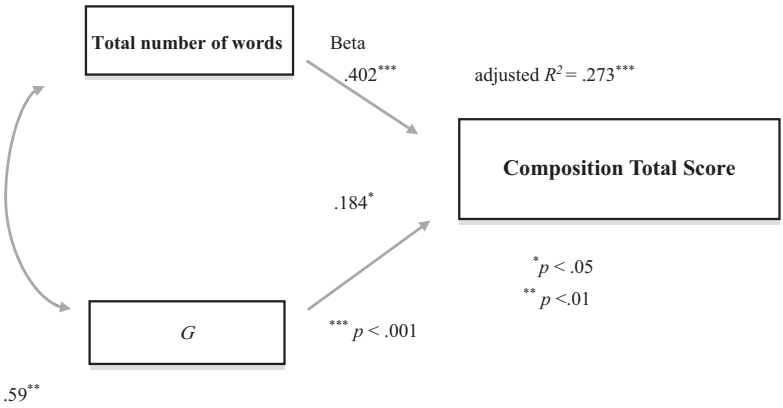

Figure 1
An Explanatory Model of EFL Composition Total Score

RQ3: How are fluency and syntactic/lexical complexity related to each other?

Correlations among the 11 linguistic measures

Pearson product-moment correlations were computed among the 11 items under study. As reported in Table 8, a majority of the raw frequency measures were highly correlated with each other: W and C (r = .89); W and TU (r = .84); C and TU (r = .83). It is noteworthy that the three measures with T-unit in the denominator had significantly high correlations with each other: W/TU and SN/TU (r = .83); C/TU and W/TU (r = .68); C/TU and SN/TU (r = .67). It is equally notable that W/C showed a different tendency from these three measures, correlating negatively with C/TU. Another interesting finding is that the raw frequency measures and the corresponding ratio measures all showed a significant negative correlation: TU and SN/TU (r = -.41); TU and C/TU (r = -.34); TU and W/TU (r = -.32); C and W/C (r = -.20). Moreover, *G* had positive significant correlations with W (r = .59), TU (r = .45), and C (r = .45). Lastly, lexical complexity measure scores all showed low correlations with each other: *G* and ConW/W (r = .07); *G* and P_Lex (r = .01); ConW/W and P_Lex (r = .08).

Table 8
Pearson Correlation Coefficients for the 11 Measures

	W	C	TU	TU/S	C/TU	W/TU	SN/TU	W/C	*G*	ConW/W	P_Lex
W	1.00	.89**	.84**	.22**	.03	.22**	.02	.26**	.59**	−.10	−.01
C		1.00	.83**	.25**	.21*	.04	−.06	−.20*	.45**	−.18*	−.13
TU			1.00	.25**	−.34**	−.32**	−.41**	.03	.45**	.05	−.01
TU/S				1.00	−.01	−.06	−.14	−.06	.19*	−.09	−.05
C/TU					1.00	.68**	.67**	−.37**	−.22**	−.38**	−.20*
W/TU						1.00	.83**	.42**	.03	−.25**	−.04
SN/TU							1.00	.23**	−.09	−.17*	−.04
W/C								1.00	.32**	.16	.22*
G									1.00	.07	.01
ConW/W										1.00	.08
P_Lex											1.00

* p < .05. ** p < .01 (two-tailed).

Exploratory factor analysis

Prior to performing exploratory factor analysis, the suitability of the data for the analysis was checked. The assumption of normality was checked by examining whether each variable was normally distributed. As reported in Table 6, skewness and kurtosis values for the 11 measures indicated relatively normal distributions. The correlation matrix (Table 8) revealed the presence of many coefficients of .3 and above. Furthermore, Bartlett's test of sphericity reached statistical significance, showing the data was spherical ($\chi^2 = 1968.77$, $df = 55$, $p = .000$). The Kaiser-Meyer-Olkin value was .613, exceeding the recommended value (.60) for a good factor analysis (Tabachnick & Fidell, 2013). Thus, the data were interpreted as appropriate for exploratory factor analysis (Kaiser, 1974).

Table 9
Varimax Rotated Factor Matrix for the Three Factor Solution

	Factor 1	Factor 2	Factor 3
W	**.97**	.21	.11
C	**.94**	.10	−.33
TU	**.93**	−.33	.03
G	**.58**	−.01	.28
W/TU	.00	**.99**	.15
SN/TU	−.15	**.84**	.01
C/TU	−.07	**.79**	**−.61**
W/C	.10	.28	**.96**
TU/S	.25	−.05	−.05
ConW/W	−.07	−.30	.26
P_Lex	−.04	−.07	.25
Eigenvalues	3.13	2.63	1.64
% of variance explained	28.44	23.92	14.91
Cumulative % of variance explained	28.44	52.36	67.27

Note. Factor loadings above .45 are shown in boldface.

The maximum likelihood method with Varimax rotation was adopted to analyze the data. Both orthogonal and oblique rotations were conducted. As recommended by Pallant (2013), oblique (Promax) factor solutions were implemented to check whether the extracted factors were correlated. Because oblique solutions indicated the extracted factors were not correlated, the Varimax method was employed for the final analysis. The Eigenvalue was set at 1.00. The chi-square goodness of fit index (Chi^2 value) obtained by the Maximum Likelihood estimation procedure indicates adequate fit: χ^2= 30.6, *df* = 25, *p* = .20. As Table 9 shows, three factors were obtained from the analysis. The first factor accounts for 28.44%, the second for 23.92%, and the third for 14.91%, in all accounting for 67.27% of the total variance. Thus, each factor explained a considerable proportion of the unique variance observed.

The first factor received very high loadings from the three raw measures W, C, and TU, which were presumed to represent writing fluency. Furthermore, *G* also loaded on this factor, thus indicating the first factor partly encompasses lexical complexity. The first factor includes W and *G*, which explained the overall writing quality significantly (recall the results reported in the *RQ2* section). Although this needs further verification, the present findings imply fluency and lexical complexity may be partially related to each other. These findings may also provide a link to those of a previous study (Hirose, 2018) that suggest fluency and *G* may develop in parallel in the early stages of EFL writing.

The second factor had the three measures with T-unit in the denominator, which could be seen as indices of syntactic complexity. They highly loaded together on this factor; W/TU loaded most highly, followed by SN/TU and C/TU. The raw measure TU showed a low, and negative loading on this factor, indicating an inverse relationship with the three ratio measures that had TU in the denominator. In other words, the smaller the number of T-units, the more words, clauses, and S-nodes per T-unit.

The third factor was associated with two measures, W/C and C/TU. The highest loading was provided by W/C, which was heavily loaded only on this factor. C/TU, on the other hand, showed a low, but negative loading, indicating

an inverse relationship with W/C: in other words, the less the subordination, the longer the clauses. This inverse relationship between C/TU and W/C may lend support to the developmental prediction that advanced students draw on complexity at the phrasal level rather than through subordination (Ortega, 2003). To further support this prediction, phrasal complexity measures such as the mean number of words per noun phrase need to be employed in future studies, as urged by Bulté and Housen (2012). As shown in Table 9, W/TU produced a low positive loading on the third factor, while loading highly positively on the second factor. It is noteworthy that C/TU, which loaded highly positively on the second factor, showed a negative loading on the third factor. Although this complex loading pattern is difficult to decipher, the third factor could be interpreted to reflect syntactic complexity, like the second factor, but of a different underlying construct.

As shown in Table 9, TU/S, the amount of coordination per sentence, showed low factor loadings on all the three factors (below .30), in particular, very low negative loadings on the two syntactic factors (Factors 2 and 3). Additionally, TU/S showed weak and mostly negative correlations with the other syntactic complexity measures, not to mention the raw fluency measures (Table 8). These findings suggest that TU/S may be distinct from the other syntactic measures.

It is also worth mentioning that W/TU and W/C did not load heavily on the same factor. W/TU and W/C did not load on the first factor but on the second and the third factor, respectively; this suggests that these measures tap syntactic complexity rather than fluency. Although Wolfe-Quintero et al. (1998) pointed to W/TU and W/C as “the best measures of fluency” (p. 29), the present findings did not support their claim. They are both more likely to measure syntactic complexity (Norris & Ortega, 2009), but W/TU and W/C behaved somewhat differently in the present analysis, implying they capture different dimensions of syntactic complexity. The different factor loadings found for W/TU and W/C seem to be compatible with Navés’s (2007) findings of an exploratory factor analysis cited in Celaya and Navés (2009), in which W/C showed a different loading from other syntactic complexity measures such as W/TU. On the other hand, in Oh’s (2006)

study cited in Norris and Ortega (2009), W/TU and W/C loaded on the same factor. Thus, the relationship between W/TU and W/C should be further examined. Additionally, the relationships between measures with T-unit in the denominator, for example, the relationship between W/TU and C/TU, should also be further examined (Polio, 2001).

Regarding lexical complexity, the factor analysis results revealed its three sub-components, lexical diversity, density, and sophistication, loaded differently, not sharing the same factor. As reported previously, these three measures were not found to be substantially correlated with each other either (Table 8). Regarding lexical density, Laufer and Nation (1995) questioned its validity as a pure lexical measure because lexical density is affected by syntactic structure. They argued that fewer function words may result from "more subordinate clauses, participial phrases and ellipsis, all of which are not lexical but structural characteristics of a composition" (p. 309). Sub-dimensions of lexical complexity and interrelationships between them are in need of further investigation.

Conclusion and Limitations

The present study attempted to capture the short-term writing development of intermediate-level Japanese EFL students. Unlike advanced ESL students (Bulté & Housen, 2014), they did not progress syntactically in any measure employed in the study. On the other hand, fluency measures and one lexical complexity measure were significantly enhanced in tandem in a semester. Furthermore, W and *G* significantly accounted for overall writing quality. The results suggest that fluency and lexical complexity can be conceived of as keys to higher quality writing for lower-level EFL students. Greater fluency does not necessarily mean higher quality. However, the present study found students also improved their overall writing quality; in other words, increased words did not sacrifice writing quality. Polio (2001) noted that: "Fluency may have no relation to quality or, possibly, a negative one. If, however, L2 writers can write more quickly, particularly if quality does

not suffer, as a result of an intervention, then we can say that development has taken place" (p. 106). The present results definitely show a case in point, or even present a stronger case, because writing quality significantly increased too, rather than being sacrificed. Although the relationship between writing fluency and quality should be further investigated, the present results suggest that for this population of intermediate-level EFL students, fluency, in parallel with lexical complexity, seems to be a positive indicator of good argumentative writing.

Future studies should further pursue multi-dimensional aspects of writing fluency by incorporating the dysfluency aspect into the analysis. Previous studies that examined L2 speaking fluency by using both dysfluency and speed measures suggest they are independent from each other, implying they do not share the same underlying construct. Sakuragi (2011), for example, examined fluency of L2 students of Japanese by means of dysfluency and speed measures. The exploratory factor analysis revealed these two measure scores did not load on the same factor, which concurred with the results of other studies (Tavakoli & Skehan, 2005). Furthermore, writing fluency should be examined by extending the measurements to capture the real time writing process, as recent studies (Leijten & Van Waes, 2013; Van Waes & Leijten, 2015) have done.

Distinct from fluency, syntactic complexity remained unchanged in a short period of four months. 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. Further studies that employ more syntactic measures would be necessary to be conclusive, and such studies should employ clausal and phrasal indices (Crossley & McNamara, 2014; Kyle & Crossley, 2018). Furthermore, the present non-significant results in terms of syntactic complexity may raise such pedagogical questions as what type of writing instruction and experience students need in order to develop syntactic complexity.

The present findings differ from those of Bulté and Housen (2014) in

several ways. First, Bulté and Housen 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. Conversely, they found no significant changes in any lexical complexity measures, including *G*. Furthermore, in spite of a limited focus on complexity, they found those measures that improved significantly were not consistent with those that significantly explained the overall writing quality. These contrastive findings between the two studies should be interpreted with caution, because their participants were advanced-level ESL students as opposed to the intermediate-level EFL students in the present study. Considering such differences in research design and participant proficiency levels, simple comparison of these results would not be desirable. Nonetheless, their multiple regression analysis found *G* was one of the four measures that significantly contributed to the overall writing quality, among which *G* had the largest beta coefficient. This finding may be congruent with that of the present study in that *G* was a significant contributory variable to the overall writing quality. More importantly, 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 provide support to Skehan's (2009) suggestion, based on comparisons between native and non-native speakers' spoken data, that "complexity may be more unidimensional in that lexical complexity and structural complexity go hand in hand" (p. 528) for native speakers, but for non-native speakers they do not. The factor analysis of the present study also revealed that syntactic complexity and lexical complexity scores did not load on the same factor. The different loadings may have derived from different ways in which syntactic and lexical complexity were conceived of and operationalized in the present study. Syntactic complexity was examined in terms of compositionality, whereas lexical complexity was analyzed in terms of diversity. Relationships between lexical and syntactic complexity require further investigation, in which the range of lexical as well as syntactic measures needs to be extended.

This study also found there was no direct correspondence between most of the subjective ratings and the objective measure scores, resonating with the results of previous studies (Crossley & McNamara, 2014; Yang, Lu, & Weigle, 2015). For example, in this study, language use improved significantly in human ratings, whereas no significant differences were found in any syntactic measures employed. Although other syntactic measures might be necessary to capture syntactic change, this gap might be related to the present analysis that did not fully encompass accuracy. Disentangling the relationship among multiple CAF measures remains for further research to pursue.

Notes

1 The total number of words was divided by the writing time. The total time excluded time spent for pre-writing but included pausing time while writing.

2 Keystroking in Japanese is different from that in English. Thus, in order to compose on computers in English, those Japanese students who may have become accustomed to the Japanese syllabary character input method have to learn the English alphabetic character input method.

3 Although the order of the two topics was not counterbalanced within each class, the order was alternated between the classes. These two topics were considered comparable and not significantly influential on the quality and quantity of student compositions for the following reason. A previous study used the same two topics for Japanese students with similar English proficiency as the participants of the present study (Hirose, 2012). Comparing the compositions on the two topics, the study found no significant difference in terms of writing quantity (= total number of words) or quality (= total scores).

4 Just as Bulté and Housen (2014) and Crossley and McNamara (2014) used combined data from the same participants, all the rated scores of the pre- and post-course composition were combined for the present analysis.

5 Applying structural equation modeling would be more desirable to diagram the relationships among the variables and construct a theory-based model. Nevertheless, the present study did not aim to construct nor test a theory-based model of EFL students' writing yet.

Acknowledgments

The research reported in this article was supported by JSPS KAKENHI Grant Number JP23520685.

References

Abdel Latif, M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics*, *34*, 99–105. doi:10.1093/applin/amp073

Anthony, L. (2012). AntConc (Version 3.3.1) [Computer Software]. Waseda University. Retrieved from http://www.laurenceanthony.net/software/antconc/

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly*, *26*, 390–395.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 writing complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21–46). Philadelphia/Amsterdam: John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. doi:10.1016/j.jslw.2014.09.005

Casanave, C. P. (1994). Language development in students' journals. *Journal of Second Language Writing*, *3*, 179–201.

Celaya, M. L., & Navés, T. (2009). Age-related differences and associated factors in foreign language writing. Implications for L2 writing theory and school curricula. In R. Manchón (Ed.). *Writing in foreign language contexts: Learning, teaching, and research* (pp. 130–155). Bristol, UK: Multilingual Matters.

Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9. doi:10.1016/j.jslw.2014.09.002

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79. doi:10.1016/j.jslw.2014.09.006

Educational Testing Service (2007). *TOEIC® speaking and writing tests workshop manual*.

Educational Testing Service (2012). *Propell® handbook for the TOEFL iBT® test workshop*.

Flower, L. S., & Hayes, J. R. (1981). A cognitive process of writing. *College Composition and Communication*, *32*, 365–387.

Hirose, K. (2005). *Product and process in the L1 and L2 writing of Japanese students of English*. Hiroshima: Keisuisha.

Hirose, K. (2012). Written feedback and oral interaction: How bimodal peer feedback affects Japanese EFL students. *The Journal of Asia TEFL*, *9*, 3, 1–26.

Hirose, K. (2018). Exploring Japanese EFL students' short-term writing development. *JACET Journal*, 62, 69–88.

Hirose, K., & Sasaki, M. (2000). Effects of teaching metaknowledge and journal writing on Japanese university students' EFL writing, *JALT Journal*, *22*, 94–113.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, *30*, 461–473. doi:10.1093/applin/amp048

Hunt, K. (1966). Recent measures in syntactic development. *Elementary English*, *43*, 732–739.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, *4*, 51–69.

Ishikawa, T. (2006). The effects of task complexity and language proficiency on task-based language performance. *The Journal of Asia TEFL*, *3*, 193–225.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36.

Kamimura, T. (2006). Effects of peer feedback on EFL student writers at different levels of English proficiency: A Japanese context. *TESL Canada Journal*, *23*, 12–39. doi:10.18806/tesl.v23i2.53

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal*, *102*, 333–349. doi:10.1111/modl.12468

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly*, *12*, 439–448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, 30, 358–392. doi:10.1177/0741088313491692

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. doi:10.1075/ijcl.15.4.02lu

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. New York/ Basingstoke, UK: Palgrave Macmillan.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–19.

Meara, P., & Miralpeix, I. (2017). *Tools for researching vocabulary*. Bristol, UK: Multilingual Matters.

Mizumoto, A. (2008). Jiyûeisakubun niokeru goi no tôkeisihyô to hyôteisha no sôgôtekihyôka no kankei [Relationship between lexical indexes and holistic scoring by raters in an English essay]. *Gakushûsha côpasu no kaiseki ni motozuku kyakkanteki sakubun hyôka sihyô no kentô* [Investigation of objective composition evaluation indices based upon the analysis of learner corpus] (The Institute of Statistical Mathematics, Japan), 15–28.

Nation, P. (2008). *Teaching vocabulary: Strategies and techniques*. Boston, MA: Heinle, Cengage Learning.

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

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*, 492–518.

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (5th ed.). Maidenhead, UK: Open University Press.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*, 590–601. doi:10.1093/applin/amp045

Polio, C. (2001). Research methodology in second language writing research: The case of text-based studies. In T. Silva & P. K. Matsuda (Eds.). *On second language writing* (pp. 91–115). Mahwah, NJ: Lawrence Erlbaum.

Polio, C., & Yoon, H. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*, 165–188. doi:10.1111/ijal.12200

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly*, *19*, 229–259.

Robb, T., Ross, S., & Shortreed, I. (1986). Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, *20*, 83–95.

Sakuragi, T. (2011). “Fukuzatsusa seikakusa ryûchôsa” sihyô no kôseigainen

datôsei no kenshô: Nihongo gakushûsha no hatsuwabunseki no baai [The construct validity of the measures of complexity, accuracy, and fluency: Analyzing the speaking performance of learners of Japanese]. *JALT Journal*, *33*, 157–173.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, *46*, 137–174.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*, 165–202.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. Manchón (Ed.). *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Bristol, UK: Multilingual Matters.

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

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, *14*, 153–173. doi:10.1016/j.jslw.2005.05.002

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson Education.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). Philadelphia/Amsterdam: John Benjamins.

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition*, *38*, 79–95. doi:10.1016/j.compcom.2015.09.012

Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning*, *45*, 381–418.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu, HI: University of Hawai'i Press.

Yamanishi, H. (2004). Kôkôsei no jiyûeisakubun wa donoyôni hyôka sareteirunoka—Bunsekiteki hyôkashakudo to sôgôteki hyôkashakudo no hikaku o tôsiteno 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.

Yamanishi, H. (2011). Purosesu apurôchi niyoru paragurafu raitingu sidô to tandai ichinensei no raitingu no hattatsu [A process approach to paragraph-writing instruction and college freshmen's writing development]. *The Bulletin of the Writing Research Group, JACET Kansai Chapter*, 9, 1–13.

Yang, W., Lu, X., & Weigle, S. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, *28*, 53–67. doi:10.1016/j.jslw.2015.02.002

Appendix A: Evaluation Criteria

1	2	3	4	5	6	7	8	9	10
Poor		Fair			Good			Very good	

CONTENT

VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic

GOOD: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail

FAIR: limited knowledge of subject • little substance • inadequate development of topic

POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate

ORGANIZATION

VERY GOOD: fluent expression • ideas clearly stated/supported • succinct • well-organized • logical sequencing • cohesive

GOOD: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing

FAIR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development

POOR: does not communicate • no organization • OR not enough to evaluate

VOCABULARY

VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register

GOOD: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*

FAIR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*

POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate

LANGUAGE USE

VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions

GOOD: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*

FAIR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*

POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate

MECHANICS

VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing

GOOD: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*

FAIR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*

POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate

Appendix B: 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.