【論文】

Examining Punctuation Errors as Validity Evidence for L2 Academic Writing Assessment

Tomoko OYAMA

Abstract

Focusing on accuracy in punctuation, the present study investigated whether and to what extent erroneous punctuation marks (or lack thereof) quantitatively differ among different levels of essays in a university-level English placement test data. Results show that appropriate punctuation use could be a valid indicator for differentiating writing levels for inexperienced writers, but not for intermediate and advanced level writers. Pedagogical implications are discussed.

Keywords: writing assessment, academic writing, integrated writing task, punctuation, error analysis, validation research

Introduction

Analyzing errors in essays can provide pedagogical implications, for L2 learners, writing instructors, and faculty teaching discipline-specific courses (Ferris, 2002; Ferris & Roberts, 2001). Especially some grammatical errors cannot be ignored, as they directly interfere with textual coherence and impede readers' comprehension (e.g., punctuation, sentence structure, tenses). Thus, a certain degree of grammatical accuracy is to be prioritized in written production (Hinkel, 2017). Yet, little work has been done to investigate whether it is a valid indicator in L2 writing assessment. To fill the void, the present study will address whether accuracy in punctuation would differentiate writing score levels in academic essays written based on integrated argumentative prompts.

Literature Review

Punctuation Errors in Academic Writing

Punctuation plays a crucial role in establishing well-organized information structure and cohesion within and across sentences in written discourse (Moore, 2016). Yet, it can be a perennial source of confusion for L2 writers, and studies have reported that punctuation errors abound in L2 writing

(Bayraktar et al., 1998; Chan, 2010; Hart, 2017). Ferris (1999) suggests ESL instructors treat grammar errors in L2 writing, based on survey responses from both students and subject-matter and ESL/EFL language instructors. Hinkel (2013) also explains that grammar instruction and feedback for writing can facilitate the development of their skills directly relevant to their production of written texts for professional and academic purposes (p. 16).

In her recent account on what English grammars should be prioritized in L2 teaching, Hinkel (2017) reviewed some previous works and identified the following three as those that should receive attention for teaching in classroom due to its direct interference with meaning: sentence structure, verb phrase, and punctuation. Indeed, these errors that can impede the smooth delivery of written communication can be considered anomalies, and thus could impact the processing of fluent reading (Keating & Jegerski, 2015). This leads readers to take more time in comprehending the information presented, which leads to Ferris’s following account (1999):

> “studies of university subject-matter instructors suggest that at least some English-speaking university faculty are less tolerant of “typical” ESL errors than of “typical” native speaker errors, and that professors feel that students’ linguistic errors are bothersome and affect their overall evaluation of student papers” (p. 8)

Some studies done on the production of punctuation marks in L2 English writing have taken a rigorous approach to coding learner errors. For instance, Hart (2017) investigated punctuation errors in students of L1 Chinese, which included the following (p. 183):

(1) a. The participants are allowed to converse if the instructions have already been given.
b. The participants are allowed to converse, if the instructions have already been given.

(2) a. Hseih and Fu (2003) increased this number to ten although they searched within a smaller area.
b. Hseih and Fu (2003) increased this number to ten, although they searched within a smaller area.

According to Hart’s account, (1b) and (2b) entail errors, as the comma is not to be used if there is an independent clause coming at the beginning of the sentence. The exceptions apply to this rule only when the first independent clause and the following dependent clause are combined with a connecting word that signals concession or contrast such as *unless* and *although*. However, some authors provide different explanations as to what constitutes an error. Dawkins (1995) consider these ‘erroneous’ ones to be valid because “All writers, evidently, want a sentence to say what they intend it to say” (p. 537), as seen below:

(3) a. Today John went to school. (p. 537)
    b. Today, John went to school. (p. 537)

(4) a. John asked for a date when he got the nerve. (p. 538)
    b. John asked for a date, when he got the nerve. (p. 538)

Based on Dawkins’ account, while (3a) is the “ordinary” sentence to provide, (3b) provides a more specific meaning when there is a context, for instance, where John was being hospitalized for one year before the production of this sentence (p. 537). Similarly, a comma between the first independent clause and the following dependent clause starting with *when* provides an emphasis to place meaningful stress on the dependent clause, as in (3b). Mann (2003) and others (e.g., Moore, 2016) holds a similar approach, taking punctuation as part of information management and its function. She also cautions against memorizing overly simplified rules for language learners, as it might damage the discoursal flow in writing. In studies looking at the level of accuracy in L2 writing in higher education, Ishikawa (1995), Polio (1997) and other scholars concur on this standpoint, coding errors based on the expectation that learner errors are taken rather leniently considering that learners (as L1 speakers) do indeed have the intentions to be understood, which is in line with Grice’ s Maxims (Grice, 1975).

Yet, at the same time, it is crucial to consider punctuation errors that do disrupt flow in discourse. Chan (2010), for instance, analyzed lexico-grammatical errors in essay of 200-300 words based on free-writing tasks written by 387 ESL students at secondary and higher education levels. 204 tokens of punctuation errors were collected in total, but the researcher found comma splice and fragment errors to be particularly problematic, and included into the punctuation error category. Below are some examples of those (p. 307).

(5) a. *I saw her face, I will know that she was very angry, so I will go to my room, and.
    b. *I have a very happy childhood. Because, my friend, my parents are very good.

In (5a), which is an example of a comma splice, two independent clauses are separated with a comma, and not correctly punctuated with a period to separate those clauses or combined with the use of a connecting word to create one stand-alone sentence. In (5b), a period is disrupting the cause-effect relation, which is to be established between the first independent and the following dependent clauses *without* the use of the period before *Because*. This is an instance of a sentence fragment. Although Dawkins (1995) claims that fragments are used intentionally to display certain intended meanings by a writer, in classroom teaching contexts, it would not be advisable to allow this as a standard approach, as suggested by scholars in recent years (e.g., Hinkel, 2017; Kolln et al., 2016; O’Conner, 2010). The next section discusses why errors merit investigation in L2 teaching and testing.

Analysis of Errors: Academic Writing

Error analysis has received some criticism for its limitations especially in relation to production tasks because of its perceived failure to capture the entirety of how L2 learners acquire and use language (Ellis, 1996); it can be said that the absence of evidence does not equal the evidence of absence. However, although some scholars strongly oppose treating grammatical errors in L2 writing (Truscott, 1996), Ferris (1999) and many other researchers in the field of language teaching and SLA advocate for taking errors into consideration from both students’ and instructors’ (both subject-matter and ESL/EFL) perspectives. Ferris (1999) emphasizes that surveys do show learners’ preference for receiving feedback on their errors from language teachers, as also discussed in recent works (Hinkel, 2013; Larsen-Freeman et al., 2016).

In her discussion of what grammars to be prioritized in L2 teaching, Hinkel (2017) provides her review of past literature that discuss ‘severe’ grammar errors that impede meaning and thus comprehension. Hinkel (2017) identifies the following three as those errors that should receive attention: sentence structure, verb phrase, and punctuation.

Psycholinguistic studies which use online measures (e.g., eye-tracking, self-paced reading tasks, ERP) have shown that anomalies, whether they are syntactic or semantic, can affect one’s processing of reading (Keating & Jegerski, 2015). It can also be said for erroneous punctuation errors; this type of error can affect the parsing by a parser during processing by disrupting the flow of cohesion and coherence, and thus would eventually lead to breakdown in reading. This leads to the following claim made by Ferris (1999):

> “studies of university subject-matter instructors suggest that at least some English-speaking university faculty are less tolerant of “typical” ESL errors than of “typical” native speaker errors, and that professors feel that students’ linguistic errors are bothersome and affect their overall evaluation of student papers” (p. 8)

Ginther and Grant (1997) conducted an error analysis of 180 essays from the Educational Testing Service’ s of *Test of Written English* (on a holistic scale from 1 to 6 with two argumentative essay topic prompts), coding errors based on parts of speech as well as errors in word form, choice, and omissions. The results of their error coding in relation to the test scores show that error frequency tends to reflect some influence of errors on essay rating, which suggests the need to “investigate more closely the errors committed at each of the levels represented” (p. 394). Given that punctuation errors interfere with meaning and comprehension, it seems that error analysis seems to merit further application for this specific aspect, in the realm of L2 academic writing as well as in testing.

Operationalization of Sentences, Clauses, and T-units

So far, it has been discussed that punctuation errors are of importance for investigation in academic writing. However, the definition of a sentence and other relevant terms such as *clause* and *T-unit* have been operationalized distinctively by different authors. A widely accepted definition can be found in Hunt (1965) and Tapia (1993), who recognize a *sentence* as a group of words that are delimited with the following punctuation marks that indicate the end of a sentence: period, question mark, exclamation mark, quotation mark, or ellipsis (as cited in Lu, 2010, p. 9; Wolfe-Quintero et al., 1998, p. 70). In regard to a definition of a *clause*, researchers have disagreed; while Hunt (1965) and Polio (1997) consider this to consist of a subject and a finite verb, Bardovi-Harlig and Bofman (1989) argue it also includes a phrase dominated by a verb phrase or a subject, which means that a clause might include fragments that have no overt verb or a subject but only one of them. Since any *sentence* that does not have both can be considered an error of sentence fragment (as discussed in Ferris & Roberts, 2001, p. 169), this paper follows the former definition for a clause. Below are the clause types in English, which are shown in italics (Wolfe-Quintero et al., 1998, p. 71):

a. Independent/main clause:

He is heroic because he saved a child’ s life.

b. Nominal/noun clause (subordinate)

What he has done is heroic because he saved a child from what would have been certain death.

c. Adjective/relative clause (subordinate)

He, *who has never been brave*, is heroic because he saved a child who was drowning.

d. Adverbial clause (subordinate)

Because he saved a child's life, he is heroic.

Among these, the last three are considered to be finite, subordinate clauses[1]. The nominal clause mainly consists of two types: "a statement of fact" introduced by *that*-complementizer (as in "*She told me that I ought to keep quiet*." , and "an indirect question" (as in "*She asked me who went to the game*." (Hunt, 1965, p. 75)). Since this paper considers a clause to have a *finite* verb (as in Hunt, 1965), it does not consider other types (cf. *to*-infinitives and *ing*-constructions as in Beaman, 1984) for the definition of a *clause*.

The relative clause in (c), or more specifically non-restrictive relative clause, is of particular relevance to the present paper, as this type of relative clause necessitates the use of a punctuation mark (i.e., a comma), and contributes to making difference in meaning. This also has been reported to be difficult for L2 learners (Sadighi, 1994).

Another term to be defined is a *T-unit*. T-unit was first put forth by Hunt (1965) for assessing syntactic maturity for young learners (see Wolfe-Quintero et al., 1998 for further discussion). Put simply, this unit is operationalized as "one main [or independent] clause with all the subordinate clauses attached to it" (Hunt, 1965, p. 20). Hunt (1965) also explains a T-unit as follows:

> "These units might be christened "minimal terminable units," since they would be minimal as to length [especially compared with a sentence], and each would be grammatically capable of being terminated with a capital letter and a period. For short, the "minimal terminable unit" might be nicknamed a "T-unit." (p. 21)

Excluding coordination as its part, T-unit focuses on subordination and the sheer number of independent clauses within a sentence. This unit has been used as one of the standard measurements in measuring grammatical accuracy as well as complexity in L2 writing and testing research (e.g., Kyle & Crossley, 2017; Norris & Ortega, 2009; Polio & Yoon, 2018).

Research Question

The present study aims to investigate the following research question:

Whether and to what extent would erroneous punctuation marks (or lack thereof) quantitatively differ among different levels of essays in a university-level English placement test data?

It is predicted that the essays with higher profile levels would contain less punctuation errors, while essays receiving lower scores would include more erroneously punctuated sentences.

Methods

Setting and Materials

Ninety-seven essays were retrieved from English placement test database at a public university in the U.S. The placement test assesses the level of English proficiency for newly admitted international students and provides information on whether and what ESL course(s) each student was required to take upon entrance. It consists of a 90-minute integrated written test and an oral interview. In the written part of this placement test, a student is asked to read six 250-300-word articles, listen to a 10-minute online lecture, and write an argumentative essay with their opinion about a given prompt on the topic, based on the information from those sources.

The errors were manually coded and analyzed by the author of this study, based on the coding scheme described above. The frequency of errors was counted by T-unit and length, with the use of the Web-based L2 Syntactic Complexity Analyzer (L2SCA, https://aihaiyang.com/software/l2sca/). An approximately equal number of essays were retrieved, representing five different scores that range from the highest (A) to the middle (B1 > B2 > C1) to the lowest (C2) level. The essay data from the actual lowest level, D, were excluded from the analysis due to its relatively low number in the database. This rating scale is based on two features, which are argument development and lexico-grammatical features. A-level essays are strong in both features, and D-level ones are weak in both. B- and C-level essays are considered in relative terms, with 1 and 2 indicating strength in argumentation and lexico-grammatical features respectively within the letter-grade scale; B-level essays are overall better than C-level essays; B1 essays are relatively stronger in argumentation than B2, while B2 stronger in lexico-grammatical features than B1. The same holds for C1 and C2 level essays.

Coding Scheme

The categories for error coding were adapted from previous works on punctuation in L2 studies (e.g., Air University Press, 2015; Chan, 2010; Dawkins, 1995; Hart, 2017; Mann, 2003). The error categories were divided into the following: misuse of a period in a main clause (i.e., sentence fragments), misuse of a comma in a main clause (i.e., comma splice), and omission of a comma in a relative clause (non-restrictive). Due to an extremely limited number of occurrences and students' overall proficiency level and their age of L2 English learning, the number of errors in overuse and misplacement of punctuation marks as well as errors with colons, semi-colons, capital letters, and others (e.g., quotation marks, dash) besides commas and periods were extremely low and thus were not considered in this study.

Table 1

Error Categories and Examples

Error Category				**Example**
Category #	**Type**	**Position**	**Mark**	
1 (Sentence fragments)	Misuse	Main Clause	Period	• *Especially for those in the HR department, who carries the important burden to find the most fitted employees for the company so that the firm could be more efficient in using its resources.* ***(period)*** • *Because HRs can be almost the most important parts of the company and a good HR can provide endless potential and valuable people to the company.* ***(period)***
2 (Comma splice)	Misuse	Main Clause	Comma	• *Also some authorized tests are not suitable for hiring process**, for example,** the MMPI is for psychological evaluation.* • *Some people maybe think that personality test is effective and useful for hiring process**, however** in my opinion, it has low practicability and many disadvantages.*
3	Omission	Relative Clause	Comma	• *To be specific, the Myers-Briggs reveals 16 personalities **which** also means thewhole population in theUS are divided into 16 groups.* • *What's more, the personality test like MMPI could greatly avoid the risk of hiring an employee with some mental problem **which** may lower the working efficiency or even cause problems.*

Although some scholars include those as punctuation errors, in the present study, the following types of errors were not considered as errors, as they still provide intended meanings in discourse and do not pertain to punctuation errors per se.

a. **Lacking a comma before an independent clause after the first independent clause**: *Let the test result to be the first impression on interviewee could help the interviewer avoid this bias**(,) and** this can provide a way to get more information about potential hires than an interviewer could gain by him/herself.*

b. **Spacing issues with a comma**:
However people in charge of the hiring can not easily understand the method to use these two tests,and when such tests are administered incorrectly ,or when the results are incorrectly interpreted, they become invalid.

c. **Omission errors in academic register that do not impede meaning**:
To support this opinion, lets say that a company does not have personality test in its hiring process.

d. **Commas used for an emphasis** (as discussed Dawkins 1995):
*However, I believe**,** the combination of the personality test and interview will be much more effective until a perfect personality test system is formed.*

Data Analysis

All statistical analyses were conducted in R (R Core Team, 2018). Means, median, and standard deviations were calculated for errors in each coding category and those in all categories combined. For the inferential statistics, a mixed effects modeling was used with Rating (i.e., A-B1-B2-C1-C2) as a fixed effect, and the essay sample number (or a participant number) and the essay topic as random effects. The dependent variable was the number of punctuation errors, which is based on a continuous scale. Therefore, a linear mixed effects model was employed via *lmer()* function from the *lme4* package (Bates et al., 2015). In addition, this modeling was chosen because the variability in raters and essay scores is better handled in this type of analysis. To obtain *p*-values, the Satterthwaite approximation was employed using the *lmer*() function from the lmerTest package as it provides the highest *p*-values and thus helps avoid making type I errors.

First, all error categories were combined, and a linear mixed effects model was run on the data based on the frequency of errors in T-unit and in length respectively. Then, the error data was further categorized into each of the three error categories ((1) misuse of a period in a main clause, (2) misuse of a comma in a main clause, and (3) omission of a comma in a relative clause), and the same linear mixed effects modeling was performed on the data based on the frequency of errors in T-unit and length respectively. In this second analysis, Topic was introduced as a random effect.

The source of any significant effects was investigated further through pairwise comparisons via *emmeans()* function (Lenth, 2018) to compare essay ratings between each pair. The default degrees of freedom for *emmeans*() is Kenward-Roger method; thus, this was manually changed to Satterthwaite method, by specifying the argument as *lmer.df* = "*satterthwaite*". Further, residuals were plotted to check for normality and homoscedasticity, and R^2 values were calculated to compare the goodness of fit for each model.

The frequency of errors was counted by the number of occurrences of punctuation errors for each different error category and then dividing it by T-unit and length (i.e., the number of words). Both the T-unit and the length for each essay were counted on the L2SCA (Web-based, batch-mode). The frequency counts were normed by multiplying 100 and 1000 (words) for T-units and length, respectively.

Table 2

Essay Data

Rating	**Topic 1:**	**Topic 2:**
	Should personality assessments be used in the hiring process?	*Should students consider attending vocational schools over universities?*
A	10 essays	6
B1	10	10
B2	10	10
C1	11	10
C2	10	10

Results

Descriptive Results

First, I would like to provide a brief overview of the data. Again, the errors were divided into three different categories; misuse of a period in a main clause (i.e., sentence fragments), misuse of a comma in a main clause (i.e., comma splice), and omission of a comma in a relative clause (non-restrictive). In each essay, the frequency of each of these error categories was divided by the number of words and T-units. Then, the frequency counts were normed by multiplying 1000 and 100 for words and T-units, respectively. There are five Rating scales (A-B1-B2-C1-C2) and two Topic prompts; the prompt for Topic 1 was “Should personality assessments be used in the hiring process?,” and the one for Topic 2 was “Should students consider attending vocational schools over universities?” (see Table 2 for the number of essays in each Rating scale). A fewer number of A-level essays were found for Topic 2 due to a low number of essays given this rating.

First, the error frequency was counted by the number of Words and then by the number of T-units. The data was first grouped by Rating (i.e., the 5 rating scales) and Topic (i.e., 2 topic prompts). The mean, median, and standard deviations were calculated. Both of these analyses by Word (Table 3) and by T-unit (Table 4) found that mostly the lower rated essays (C1-C2) had a higher error frequency as opposed to higher rated essays (A-B1-B2).

Table 3

By Word				
Rating	Topic	mean	median	sd
A	1	0.374327	0	0.608065
A	2	0.493637	0	0.953424
B1	1	0.946557	0	1.231636
B1	2	0.457601	0	0.909112
B2	1	0.454866	0	1.184083
B2	2	0.510198	0	1.094656
C1	1	1.01563	0	1.741077
C1	2	1.241025	0	1.702293
C2	1	0.664198	0	1.321929
C2	2	1.66819	0.618047	2.143769

Table 4

By T-unit				
Rating	Topic	mean	median	sd
A	1	0.918695	0	1.580471
A	2	1.128606	0	2.206317
B1	1	2.748547	0	3.582245
B1	2	1.183068	0	2.53715
B2	1	1.553465	0	4.227374
B2	2	1.996288	0	4.296158
C1	1	3.214993	0	5.783853
C1	2	5.058885	0	9.238116
C2	1	3.018866	0	7.03629
C2	2	6.34756	1.373438	8.861332

The frequency of errors counted by T-units were further visualized to see whether any differences in frequency counts would exist among those five rating scales, for each Topic. The analysis for Topic 1 is shown in Figure 1, and Topic 2 in Figure 2.

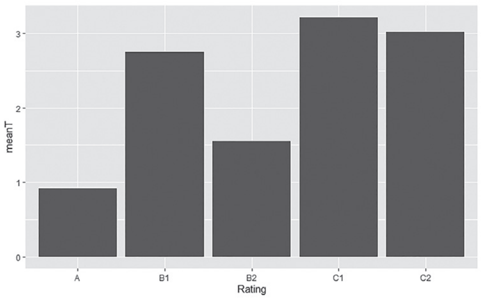

Figure 1

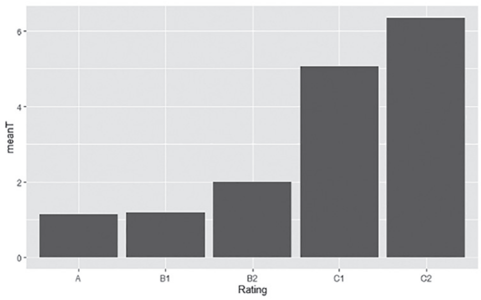

Figure 2

It seems that the frequency counts by T-units are different between the two Topics especially for B1 level, which needs to be taken into account when running inferential statistics, because these tables indicate that Topic might affect the number of punctuation errors. Further analysis was conducted by error Type (Figure 3), by error Position (Figure 4), and by error Mark (Figure 5).

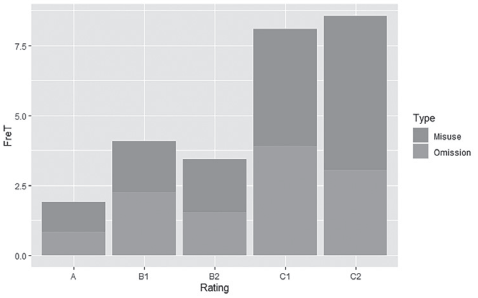

Figure 3

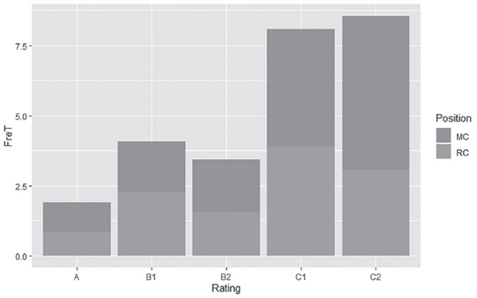

Figure 4

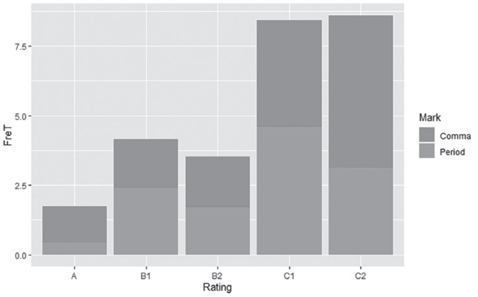

Figure 5

The results each from Figure 3-5 indicate the followings; (1) the misuse of punctuations is slightly more frequent than the omission in almost all of the rating scales; (2) more punctuation errors are found in main clauses rather than in relative clauses in almost all rating scales; (3) erroneous comma punctuations are used made more frequently in some rating scales but period errors are more frequent in the other scales.

Inferential Results

First, all models did not fail to converge, and the normality of residuals and homoscedasticity were checked (see Appendix for the annotated code).

Table 5 and 6 show results on data with all error categories combined. Table 5 based on the units of T-units shows a significant effect of rating for the C-level essays (C1: estimate = 3.067, std error =

1.200, *t*(91) = 2.556, *p* = .012 / C2: estimate = 3.650, std error = 1.213, *t*(92) = 3.009, *p* = .003), indicating the significantly higher error rate in comparison with the reference, or A-level essays in T-units. No such effects were found in B-level essays (*p* > .05). Table 6 based on the unit of length shows similar findings obtained for the analysis in T-units, which shows a significant effect of rating for C-level essays (C1: estimate = .700, std error = .273, *t*(92) = 2.565, *p* = .012 / C2: estimate = .743, std error = .276, *t*(92) = 2.691, *p* = .009) in comparison with the A-level essays. Again, no such effects were obtained with B-level essays.

Table 5

Linear Mixed Effects Model for All Error Categories Combined (By T-Unit)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	1.0333	0.9398	16.4070	1.100	0.28740
RatingB1	0.9325	1.2132	91.5892	0.769	0.44408
RatingB2	0.7416	1.2132	91.5892	0.611	0.54254
RatingC1	3.0666	1.1999	91.4226	2.556	0.01225*
RatingC2	3.6499	1.2132	91.5892	3.009	0.00339**

Random effects	**Variance**	**SD**
Essay	3.3273	1.8241
Topic	0.1286	0.3586

*Note. .p < .1, *p < .05, **p < .01, ***p < .001.*

Table 6

Linear Mixed Effects Model for All Error Categories Combined (By Length)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	0.42354	0.20983	20.23427	2.019	0.05699
RatingB1	0.27854	0.27599	91.67321	1.009	0.31553
RatingB2	0.05899	0.27599	91.67321	0.214	0.83123
RatingC1	0.70027	0.27301	91.49084	2.565	0.01194*
RatingC2	0.74265	0.27599	91.67321	2.691	0.00847**

Random effects	**Variance**	**SD**
Essay	0.045588	0.21351
Topic	0.003343	0.05781

*Note. .p < .1, *p < .05, **p < .01, ***p < .001.*

Table 7 through 12 show the results of analysis based on each of the three error categories (namely, (1) misuse of a period in a main clause, (2) misuse of a comma in a main clause, and (3) omission of a comma in a relative clause). Table 7 through 9 are based on the analysis of errors in T-units, and 10 through 12 in the unit of length. Overall, varying effects of rating were revealed.

For the analysis based on T-units, while Table 7 (sentence fragments, or misuse of a period in a main clause) and 9 (omission of a comma in a relative clause) show a marginal effect of rating in C1-level essays in comparison with the reference level A-essays, Table 8 (comma splice, or misuse of a comma in a main clause) shows a significant effect of rating for C2-level essays (estimate = 6.127, std error = 1.934, *t*(92) = 3.169, *p* = 0.002).

Table 7

Linear Mixed Effects Model for the First Error Category (sentence fragments) (By T-Unit)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	0.6157	1.752	7.6537	0.351	0.7347
RatingB1	1.7974	2.0682	91.3476	0.869	0.3871
RatingB2	1.0864	2.0682	91.3476	0.525	0.6006
RatingC1	3.9974	2.0449	91.241	1.955	0.0537.
RatingC2	2.5182	2.0682	91.3476	1.218	0.2265

Random effects	**Variance**	**SD**
Topic	1.372	1.171

*Note. .p < .1, *p < .05, **p < .01, ***p < .001.*

Table 8

Linear Mixed Effects Model for the Second Error Category (comma splice) (By T-Unit)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	1.7167	1.4413	92	1.191	0.23668
RatingB1	-0.497	1.9337	92	-0.257	0.79773
RatingB2	0.3493	1.9337	92	0.181	0.85705
RatingC1	2.0796	1.9131	92	1.087	0.27986
RatingC2	6.1273	1.9337	92	3.169	0.00208**

Random effects	**Variance**	**SD**
Topic	4.941e-17	7.029e-09

*Note. .p < .1, *p < .05, **p < .01, ***p < .001.*

Table 9

Linear Mixed Effects Model for the Third Error Category (relative clause) (By T-Unit)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	0.8247	1.2598	92	0.655	0.5143
RatingB1	1.4398	1.6902	92	0.852	0.3965
RatingB2	0.7317	1.6902	92	0.433	0.6661
RatingC1	3.0763	1.6723	92	1.84	0.0691.
RatingC2	2.247	1.6902	92	1.329	0.187

Random effects	**Variance**	**SD**
Topic	0.00	0.000

Note. .*p* < .1, **p* < .05, ***p* < .01, ****p* < .001.

The analysis based on the unit of length (Table 10 through 12), on the other hand, shows somewhat different results. Table 10 shows a significant effect of rating for C1 level essays (estimate = .865, std error = .424, *t*(91) = 2.042, *p* = .044) in data based on the first error category (i.e., sentence fragments), but no such effects for the essays of this level were found in the other two error categories (*p* > .05). In the results based on the second category (i.e., misuse of a comma in a main clause) as seen in Table 11, the effect of rating was found only for C2 level essays (estimate = 1.336, std error = .514, *t*(92) = 2.598, *p* = 0.011). No effects of rating were found for any essay levels in the data for the third error category (i.e., omission of a comma in a relative clause) as seen in Table 12.

Table 10

Linear Mixed Effects Model for the First Error Category (By Length)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	0.2242	0.3496	10.0844	0.641	0.536
RatingB1	0.5733	0.4284	91.425	1.338	0.184
RatingB2	0.2195	0.4284	91.425	0.512	0.61
RatingC1	0.8651	0.4236	91.2974	2.042	0.044*
RatingC2	0.5067	0.4284	91.425	1.183	0.24

Random effects	**Variance**	**SD**
Topic	0.04003	0.2001

Note. .*p* < .1, **p* < .05, ***p* < .01, ****p* < .001.

Table 11

Linear Mixed Effects Model for the Second Error Category (By Length)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	0.6665	0.3833	92	1.739	0.0854.
RatingB1	-0.1898	0.5142	92	-0.369	0.7129
RatingB2	-0.0555	0.5142	92	-0.108	0.9143
RatingC1	0.5997	0.5087	92	1.179	0.2415
RatingC2	1.3359	0.5142	92	2.598	0.0109*

Random effects	**Variance**	**SD**
Topic	0.00	0.000

Note. .*p* < .1, **p* < .05, ***p* < .01, ****p* < .001.

Table 12

Linear Mixed Effects Model for the Third Error Category (By Length)

Fixed effects	**Estimate**	**SE**	***df***	***t*-value**	***p*-value**
(Intercept)	3.93E-01	3.31E-01	9.20E+01	1.187	0.238
RatingB1	4.40E-01	4.44E-01	9.20E+01	0.99	0.325
RatingB2	4.00E-04	4.44E-01	9.20E+01	0.001	0.999
RatingC1	6.26E-01	4.39E-01	9.20E+01	1.426	0.157
RatingC2	3.73E-01	4.44E-01	9.20E+01	0.84	0.403

Random effects	**Variance**	**SD**
Topic	0.00	0.000

Note. .*p* < .1, **p* < .05, ***p* < .01, ****p* < .001.

The R^2 value results for the models above show the best fit for the models analyzing the second error category (i.e., the misuse of a comma in a main clause, or comma splice), which is summarized in Table 13 below:

Table 13

R^2 Values for Models Based on Each Error Category

Error Data	**Unit**	***R2m***	***R2c***
All errors combined	T-unit	0.057147	0.156952
-	Length	0.047503	0.071512
1st error category (Sentence fragments)	T-unit	0.043317	0.076758
-	Length	0.049259	0.072101
2nd error category (Comma splice)	T-unit	0.15253	0.15253
-	Length	0.123605	0.123605
3rd error category (Non-restrictive relative clauses)	T-unit	0.043413	0.043413
-	Length	0.03425531	0.03425531

The post-hoc pairwise comparison provided results revealing significant contrasts between A/B level essays and C level essays, mainly in the data for all errors combined and the second error category (i.e., comma splice errors). No significant contrasts in essay levels were found in the first (misuse of a period in a main clause, or sentence fragments) and the third (omission of a comma in a relative clause) error category data. Table 14 below is the summary of post-hoc results, based on the degrees of freedom with Satterthwaite method.

Table 14

Summary of the Significant Post-Hoc Pairwise Comparisons (Satterthwaite method)

Category	**Significant comparison**	***p*-value**
All errors (T-units)	A - C1	0.0875.
-	A - C2	0.0273*
-	B2 - C2	0.0897.
All errors (length)	A - C1	0.0856.
-	A - C2	0.0631.
-	B2 - C2	0.0735.
Comma splice (T-units)	A - C2	0.0173*
-	B1 - C2	0.0041**
-	B2 - C2	0.0173*
Comma splice (length)	A - C2	0.0791.
-	B1 - C2	0.0185*
-	B2 - C2	0.0398*

*Note. .p < .1, *p < .05, **p < .01.*

Discussion and Conclusions

When all errors were combined, C2 level essays fared worst in comparison with any other rating scale level. This is an interesting observation, as briefly explain above in Methods, it is the C2 level essays that are considered to be relatively stronger in lexico-grammar features; therefore, although C1 essays were expected to fare worse than C2 essays in the quantity of errors, it seems that C2 level essays also contain a good number of punctuation errors.

However, it is also important to note that errors in the first error category (i.e., misuse of a period in main clause) were observed greatly in C1 essays, indicating that this error category (i.e., sentence fragments) might be particularly important in deciding whether essays will be evaluated as C1 or C2 and thus might be influential in determining whether an essay is strong or weak in lexico-grammatical features. In turn, this can also indicate that the other two error categories related to the use of commas (i.e., comma splice and a comma in non-restrictive relative clause) might not contribute much to the decision on whether an essay is strong in the lexico-grammar feature during the process of rating. This corroborates the findings of some recent studies showing that incomplete sentential structure such as sentence fragments can significantly impact one’s parsing of a sentence as shown in L2 self-paced reading studies (Roberts, 2016), possibly because it is indeed lacking information within one same sentence. This is a situation that differs in the comma splice (2nd error category) and the non-restrictive relative clause comma errors (3rd error category), in that the sentences in these two categories still at least include information to be communicated to readers. Yet, this finding needs to be reconsidered, as errors with comma splice (i.e., the 2nd error category) show the best goodness of fit, which is in line with literature pointing out the severity of comma spice issues and its impacts on essay evaluations (cf. Bakla, 2019).

Overall, no significant effects of rating or pairwise differences were found between A- and B-level essays, and it can be deduced that it is the argument development that distinguishes between these two more advanced levels. Different linguistic features differentially distinguish learner levels (Ishikawa, 1995), and the results of this study might indicate that the punctuation use might not be a good indicator for essay rating but might become more effective for low level essays, as shown in the results revealing significant differences between C-level and B/A-level essays. As Bardovi-Harlig and Bofman (1989) explained, L2 learners at advanced levels seem to not have issues with dealing with *global* errors (i.e., errors that impede comprehension such as syntactic errors) unlike lower-level students who still struggle with grammatical issues that contribute to differences in meaning.

The present study investigated provides pedagogical implications as to what punctuation features to be prioritized in classroom teaching for L2 writing depending on different proficiency levels. Especially for lower-level learners, materials dedicated to punctuation use might merit instructional focus.

Notes

[1] Some scholars view only adverbial clauses, or adverbial and adjectival clauses as subordinate (Wolfe-Quintero et al., 1998, p. 72).

References

Air University Press. (2015). Grammar and punctuation. *Style and Author Guide*. (2nd ed.). (pp. 61-84). Air University Press.

Bakla, A. (2019). A mixed-methods study of tailor-made animated cartoons in teaching punctuation in EFL writing. *ReCALL*, 31(1), 75-91. https://doi.org/10.1017/S0958344018000046

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*(1), 17-34. https://www.jstor.org/stable/44487471

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. http://arxiv.org/abs/1506.04967

Bayraktar, M., Say, B., & Akman, V. (1998). An analysis of English punctuation: The special case of comma. *International Journal of Corpus Linguistics, 3*(1), 33-57. https://doi.org/10.1075/ijcl.3.1.03bay

Chan, A. Y. (2010). Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners. *TESOL Quarterly, 44*(2), 295-319.

Dawkins, J. (1995). Teaching punctuation as a rhetorical tool. *College Composition and Communication*, 46(4), 533-548. https://www.jstor.org/stable/358327

Ellis, R. (1996). *Second language acquisition research and language teaching*. Oxford University Press.

Ferris, D. (1999). The case for grammar correction in L2 writing classes: A response to Truscott (1996). *Journal of Second Language Writing, 8*(1), 1-11.

Ferris, D. (2002). *Treatment of error in second language student writing* (1st ed.). University of Michigan Press.

Ferris, D., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be?. *Journal of Second Language Writing, 10*(3), 161-184.

Ginther, A., & Grant, L. (1997). The influence of proficiency, language background, and topic on the production of grammatical form and error on the Test of Written English. *Current developments and alternatives in language assessment*, 385-97.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. (Eds.), *Studies in Syntax and Semantics III: Speech Acts* (pp. 183–198). Academic Press.

Hart, S. (2017). *English Exposed: Common Mistakes Made by Chinese Speakers*. Hong Kong University Press.

Hinkel, E. (2013). Research findings on teaching grammar for academic writing. *English Teaching*, 68(4), 3-21.

Hinkel, E. (2017). Prioritizing Grammar to Teach or Not to Teach. *Handbook of Research in Second Language Teaching and Learning, 3*, 369-383.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. National Council of Teachers of English.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*(1), 51-69. https://doi.org/10.1016/1060-3743(95)90023-3

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user’s guide. *Studies in Second Language Acquisition*, 37(1), 1-32.

Kolln, M., & Gray, L., & Salvatore, J. (2016). *Understanding English Grammar* (10th ed.). Pearson Education Inc.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513-535.

Larsen-Freeman, D., Celce-Murcia, M., Frodesen, J., White, B., & Williams, H. A. (2016). *The grammar book: Form, meaning, and use for English language teachers*. National Geographic Learning, Heinle Cengage Learning.

Lenth, R. (2018). Emmeans: Estimated marginal means, aka least-squares means. R Package

Version 1(2). https://cran.r-project.org/web/packages/emmeans.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu

Mann, N. (2003). Point counterpoint: Teaching punctuation as information management. *College Composition and Communication, 54*(3), 359-393. https://www.jstor.org/stable/3594170

Moore, N. (2016). What’s the point? The role of punctuation in realising information structure in written English. *Functional Linguistics, 3*(1), 1-23. https://doi.org/10.1186/s40554-016-0029-x

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

O’Conner, P. T. (2010). *Woe Is I: The Grammarphobe’s Guide to Better English in Plain English*. (3rd ed.). Penguin Publishing Group.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143. https://doi.org/10.1111/0023-8333.31997003

Polio, C., & Yoon, H. J. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics, 28*(1), 165-188. https://doi.org/10.1111/ijal.12200

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roberts, L. (2016). Self-paced reading and L2 grammatical processing. In Mackey, A. & Marsden, E. (Eds.), *Advancing Methodology and Practice. The IRIS repository of instruments for research into second languages* (pp. 58-72). Routledge.

Sadighi, F. (1994). The acquisition of English restrictive relative clauses by Chinese, Japanese, and Korean adult native speakers. *International Review of Applied Linguistics in Language Teaching*, 32(2), 141-153.

Tapia, E. (1993). *Cognitive demand as factor in interlanguage syntax: A study in topics and texts*. Indiana University.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*, 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

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