Dynamic Effects of Task Type Practice on the Japanese EFL University Student's Writing: Text Analysis with Coh-Metrix

Kyoko Baba & Ryo Nitta *

Kinjo Gakuin University
*Nagoya Gakuin University

Abstract

The longitudinal effects of repeating a timed classroom writing activity on the EFL (English as a Foreign Language) student's L2 (second language) writing development were investigated. Data for 46 students were collected 30 times over one year. The students' compositions were analyzed for fluency and for grammatical and lexical complexity. Text analysis using *Coh Metrix* and visual analysis of non linear individual developmental patterns showed that task repetition has an overall effect on L2 writing development. Grammatical complexity developed more prominently than the other aspects of writing. This counter intuitive result points to the significance of the writers' reflective consciousness towards their own writing.

Introduction

This study investigated the longitudinal effects of repeating a timed writing activity on the EFL (English as a Foreign Language) student's L2 (second language) writing development. Studies on task repetition have shown that the quality of their spoken language improves when the same task is repeated, because learners can pay focal attention to form the second time (Bygate & Samuda, 2005). On the other hand, it is unclear what effects repetition of a writing task brings about, assuming that learners are less pressured in writing.

Unlike most previous studies, this exploratory study throws light not only on a general trend of a group of learners but also on individual developmental patterns, focusing on the non-linearity of writing development. We first assessed which of the three principal aspects of L2 writing (fluency, syntactic complexity, and lexical complexity) develop in one year. Then the textual features that changed significantly during the period in terms of dynamic developmental patterns are examined with a graphical method.

Longitudinal Research on L2 Writing

The significance of longitudinal research on L2 learning is often emphasized, but there is still a dearth of such research. There have been a few longitudinal studies on L2 writing, but they were often either cross-sectional (Henry, 1996), did not look at a variety of text features (Bardovi-Harig, 2002), or focused on a small number of writers (Larsen-Freeman, 2006; Verspoor, et al., 2008). In addition, most studies have dealt with "L2 writers" as a whole, as if there were an average L2 writer at different developmental stages. For educational purposes, however, it is no less important to know how each individual develops than to know how L2 writers in general develop. This is because it is unlikely that every L2 writer develops the same writing features at the same speed. Focusing simultaneously on individual learners and on learners as a whole is an approach that is congruent with the recently advocated complex systems perspective, which sees both the forest and the tree (see below).

To date, no L2 studies have looked at both group trends and individual developmental patterns longitudinally, except that a series of longitudinal studies of L2 adult learner's spontaneous speech have been conducted by Crossley and his colleagues (Crossley, et al., 2009, in press). Crossley et al. (2009) investigated the development of hypernymic lexical knowledge and lexical diversity in the speech of six learners, who were interviewed every two weeks over one year. They showed with a repeated-measures analysis of variance that the learners as a group significantly developed both aspects in their speech over one year. As a supplement to this analysis, they examined whether each of the six learners fit the group trend, and confirmed that most of them did.

We carried Crossley and his colleagues' idea one step further in this study. We also analyzed the developmental trajectories of individuals. However, unlike Crossley and his colleagues who conducted such analysis to "control for random effects" (Crossley et al, 2009, p. 320), we aim to

contribute to an understanding of the dynamic nature of L2 writing development by examining the individual trajectories.

The Dynamic Nature of L2 Writing Development

Two-wave research designs have often been used to investigate the effects of a certain task or teaching method on L2 writing. Usually, a pretest and a posttest are conducted, and a t-test or ANOVA assesses the difference between the two time points (e.g., Shaw & Liu, 1998). However, while such an approach reveals *whether* L2 writers as a whole have developed or not, it is impossible to know *how* they developed their writing. This is because the two-wave research design assumes linear development, which is often not the case with language learning.

To explore the *how* of L2 writing development, this study adopted a complex systems view, in which the non-linearity of language development is underscored (Larsen-Freeman & Cameron, 2008). The exploration of non-linear language development requires multiwave data (Willett, 1994). That is, data is collected from each participant multiple times, which enables us to plot individual growth trajectories. These trajectories can be analyzed with various (especially visual) tools, such as a polynomial trendline, moving min-max graph, and detrended representation of L2 development (Verspoor, et al., 2008).

Since the investigation of intra-individual variability has just begun to draw attention in the L2 developmental field, there are only a few studies that have examined intra-individual development in L2 writing. For instance, Larsen-Freeman (2006) studied the development of five immigrant females' oral and written production over six months in terms of their fluency, accuracy, and complexity. As a group, there was overall progress in all aspects of the participants' writing. However, paths to development widely differed from individual to individual concerning the rate of developmental speed and the aspects that developed.

In this study, the dynamic approach to the study of L2 writing development taken by Larsen-Freeman (2006) is used and extended. Whereas the previous studies focused on a fairly small number of writers, the aim of this study is to make a slightly broader generalization about individual developmental patterns. We explore the group trends as well as individual trends with a larger number of participants than previous studies. Thus, the research questions of this study are: (1) Which textual features of students' writing change in one-year EFL courses through repeating the same writing task? and (2) If there are significant changes in their writing, how do individual students develop each aspect of their writing?

Methods

Participants

The study was conducted at two Japanese universities in two EFL classrooms. Both courses (one-year course) were offered to first-year English major students and taught by the authors. Twenty-three students in each class were engaged in the project regularly until the end of the year. Thus, the compositions of a total of 46 students were used for the analysis. The two courses had different course objectives (one focused on English writing, while the other was oriented more toward integrated skills of speaking, listening, and reading as well as writing). However, there were no significant differences between the two classes in the linear combination of the six text measures (see below) either for the first week, Wilks's $\Lambda = .97$, $F(6, 37) = .21$, p $> .05$, or for the last week Wilks's $\Lambda = .84$, $F(6, 38) = 1.22$, p $> .05$. Therefore, we combined the two classes into one group for our data analysis. We acknowledge that it is worth examining differences, for example, in the rate of development of the two classes, but we decided not to do so due to space limitation.

The student's average TOEIC scores were 351 at the beginning, and 382 at the end of the one-year course. The students had little experience writing in English before entering university, and, even in university, they had little chance to write in English outside the classroom. Therefore, any progress in their writing is largely attributable to the classroom writing activities.

Data Collecting Schedule

The project was conducted over two academic terms consisting of thirty weeks in total. The exact same procedures were followed in the two classes. In every class meeting, the students performed a task called 'freewriting' (Elbow, 1998). Learners simply wrote without stopping for ten minutes about a topic. Each time they were given a topic list with three different topics in order to compensate for differences in their individual experiences and preferences. The topics were selected and adapted from the list of the essay topics for the Regents' Test (Weigle, 2006). We tried to devise such topics that would be easy for Japanese university students to write about and could be written on the basis of their personal experience. Examples of the topics are "What is your favorite source of entertainment? Explain why."; Take up some recent event (some news on the paper), and discuss what you think of it."; and "What do you hope to accomplish within the next ten years?"

Immediately after writing a composition, the students were requested to write reflective comments on their writing in Japanese (L1) about, for example, what they found difficult to write, what they thought during writing, etc. The collected writing was checked by the authors and

returned with some feedback in the next week. The aim of the feedback was to create a sense of audience and to maintain students' motivation for writing every week. Since the freewriting focuses on meaning rather than form, linguistic correction was not offered.

The same list of three topics was used for two weeks. The week after a new topic list was given, students were required to write about the same chosen topic to identify the task practice effects. Then, the next week, another new topic list was given.

Text Measurements

There have been four main categories of measures used to study the development in the quality of texts by L2 writers: fluency, accuracy, and lexical and grammatical complexity (Wolf-Quintero et al., 1998). Since we did not focus on accuracy (we neither told our students to pay attention to grammatical and lexical accuracy nor corrected errors in their writing), we looked at the other three aspects of text quality. In particular, we had expected that fluency would improve to some extent, as task repetition research has shown with spoken data.

To analyze the approximately 1300 compositions that were written by the students over one year, we used the web-based computational tool *Coh-Metrix* (Graesser et al., 2004) available at http://cohmetrix.memphis.edu/cohmetrixpr/index.html (version 2.0). Spelling mistakes in the compositions were carefully corrected manually and also with a spelling checker for analytical purposes. We selected the six most relevant measures of fluency and lexical and grammatical complexity as follows.

Fluency. "Fluency" in writing is an elusive concept, but it usually means either how fast or how coherently one writes. The present study used two text measurements to cover both meanings of fluency (speed and coherence). Speed is typically assessed with total number of words in a composition, for skilled writers tend to write longer (Wolfe-Quintero et al., 1998). So, (1) text length (number of words per composition) was used as a measure of speed. To assess coherence, we used (2) LSA (Latent Semantic Analysis). LSA evaluates the similarity of meaning between words, sentences, and passages by analyzing large corpora. We used it as a measure of fluency because similarity of meaning across sentences and paragraphs is likely to contribute to coherence in text. There were three LSA measures on Coh-Metrix, and we chose LSA of all combinations of sentences.

Lexical Complexity. The operationalization of lexical complexity varies with the focus of study, but it generally denotes either lexical sophistication or lexical diversity in text. Therefore, we used two measures of lexical complexity: (3) word frequency values from the CELEX corpus, which has been shown to account for reading difficulty judged by EFL learners (Crossley et al., 2008), and (4) MTLD (the Measure of Textual Lexical Diversity). MTLD was chosen among other lexical diversity measures like type-token ratio and D, because McCarthy (2005) has

revealed that it was the only measure among 14 well-known lexical diversity measures that was not affected by text length. We considered using *vocd* in addition to MTLD, but the correlation with the two measures were strong ($r = .76$), so it seemed redundant to include both measures for the purpose of this study.

Grammatical Complexity. To measure grammatical complexity, we used (5) ASL (average sentence length) and (6) STRUT (sentence syntax similarity, all sentences across paragraphs) (McCarthy et al., 2009). ASL was used as a grammatical complexity measure because Ortega (2003), synthesizing studies on syntactic complexity in L2 writing, found that it largely differed according to L2 proficiency levels. STRUT gauges the degree of similarity in syntactic structures of sentences in a passage by comparing syntactic trees of each pair of sentences. A lower STRUT value possibly indicates greater grammatical complexity. Whereas they did not analyze L2 writers' text, Crossley et al. (2008) revealed that STRUT was one of the three measures that contributed to readability judged by EFL learners.

The correlations among the six measures were all less than $r = .30$, except for the one between ASL and STRUT ($r = -.58$).

Results

A one-way multivariate analysis of variance (MANOVA) was conducted to assess whether the six text measurements significantly changed over one year. The MANOVA compared the compositions that were written in the first and last weeks (Weeks 1 and 30). Significant differences were found among the first and last compositions on the dependent measures, Wilks's $\Lambda = .63$, $F(6, 82) = 8.08$, $p < .001$. The multivariate partial η^2 based on Wilks's Λ was strong, .37. Thus, it was shown that there was a significant change in the students' writing over one year.

Analyses of variances (ANOVAs) on each dependent variable were conducted as follow-up tests to the MANOVA (Table 1). The means and standard deviations for the first and last week are displayed below each measure. To control for Type I error, each ANOVA was tested at the .008 level, using the Bonferroni method. Three measures turned out to be significant: MTLD, $F(1, 87) = 17.97$, $p < .001$, ASL, $F(1, 87) = 19.02$, $p < .001$, and STRUT, $F(1, 87) = 19.98$, $p < .001$.

The results of the ANOVAs suggest that the students improved lexical diversity (MTLD) and grammatical complexity (ASL and STRUT) more markedly than the other text features. It was somewhat unexpected that fluency hardly improved, and yet, syntactic complexity and lexical diversity improved significantly.

While the MANOVA and the ANOVAs examined the general trends for this group of students, a further analysis was conducted to illustrate how each student developed (or did not develop) the three text measures (MTLD, ASL, STRUT). Due to lack of space, we take up STRUT to represent grammatical complexity here, for it was the best

predictor of the student's writing development, and also strongly correlated with the other grammatical complexity measure, ASL ($r = -.58$).
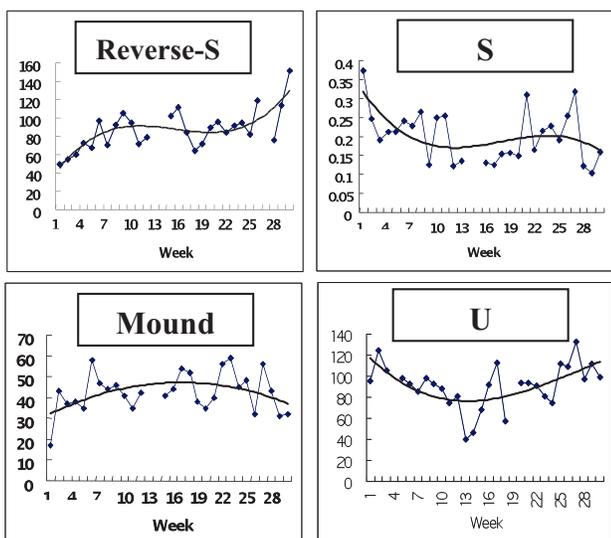

Figure 1. Four non linear line patterns of development (samples).

First, we drew a line graph for each text measure for all individual students. These graphs simply display the change of the measure in the individual students' writing over one year. Next, we added a polynomial trendline of the third degree (Verspoor et al., 2008). This is a common smoother to show the general trajectory of each student's development. Last, by closely examining the trendlines, we identified five line patterns of development: (1) Reverse-S, (2) S, (3) Mound, (4) U, and (5) Line. All the patterns except the Line are non-linear (see Figure 1). Since we focused on the line patterns for this analysis, we ignored the differences in the units of analysis and ranges of scores. In addition to the line patterns, we also identified whether each trendline has an upward, stagnant, or downward trend. For example, the upper left graph in Figure 1 has a Reverse-S pattern with an upward trend. Thus, there were 15 developmental patterns (five line patterns times three up-and-down trends). The authors separately categorized all the students into the 15 patterns. Our inter-rater reliabilities were 93% for MTLD and 96% for STRUT.

After calculating the inter-rater reliabilities, we reached agreement on all discrepancies. The numbers of students who were categorized into each developmental pattern are tabulated for MTLD and STRUT in Tables 2 and 3. To make the two text measures comparable, the up-and-down trends are reversed for STRUT.

Table 2
Number of Students Categorized in Each Developmental Pattern for MTLD

	RevS	S	Mound	U	Line
+	19		5		11
0		1	1		
-		5	3	1	

Table 3
Number of Students Categorized in Each Developmental Pattern for Sentence Syntax Similarity

	S	RevS	U	Mound	Line
-	16	4	9	9	4
0	1				
+		2		1	

The first thing to be noted is that the developmental patterns for most of the students are non-linear, even though we identified the patterns with the trendlines. When we look at the raw line graphs, the fluctuation of the development is more obvious. Tables 2 and 3 also show that more than 75% of the students made progress in both text measures (35 students for MTLD and 42 students for STRUT). For both measures, the largest number of students (about 40%) displayed a Reverse-S pattern with an upward trend. This may mean that in both measures, many students made progress at first, then there was a stagnant period, and then they progressed further. For other patterns, a similar number of students were identified with each pattern (although there was only one student who showed a U pattern for MTLD). That is, whether the learners make progress or not, the trajectories of changes in their writing may differ from learner to learner.

Table 1
Multivariate and Univariate Analysis of Variance F Ratios for Task Practice Effects for Text Measures

		ANOVA					
	MANOVA	TextL	LSA	MTLD	WFreq	ASL	STRUT
Variable	$F(6, 82)$	$F(1, 87)$	$F(1, 87)$	$F(1, 87)$	$F(1, 87)$	$F(1, 87)$	$F(1, 87)$
Time	8.08^{***}	6.13^{*}	1.51	17.97^{***}	.00	19.02^{***}	19.98^{***}
		M SD	M SD	M SD	M SD	M SD	M SD
First		53.36 20.48	21 .16	37.26 11.83	3671 1767	7.69 1 63	23 .08
Last		65.36 24.94	.17 .12	51.38 18.75	3671 1570	10.42 3 83	.17 .06

Note. *F* ratios are Wilks's approximation of *F*s. MANOVA = multivariate analysis of variance; ANOVA = univariate analysis of variance; TextL = text length; Wfreq = raw frequency of content words.
$^{*}p < .05.$ $^{***}p < .001$

A noticeable difference was observed between the two measures. There were 11 students who did not make progress (with a stagnant and a downward trend) in MTLD, while only four students did not make progress in STRUT. This may suggest that it might have been easier for most students to improve syntactic complexity than lexical diversity.

Discussion

This study showed that the EFL students improved their writing over one year, and that among the six fluency and complexity measures, they significantly developed MTLD, ASL, and STRUT. The analysis of individual developmental patterns also shows that more than 75% of the students made some progress in MTLD and STRUT. (It is not reported in the tables, but 74% of the students improved ASL.)

On the basis of findings from previous research in task-repetition effects on speech production, we expected that writing fluency would improve the most out of the three principal aspects of writing skills (fluency, lexical complexity, and grammatical complexity). However, there was not significant progress in fluency, in terms of either speed (text length) or coherence (LSA). This finding may suggest that the development of fluency in L2 writing may take more than one year, especially when the learners write only once a week in EFL settings.

It is difficult to interpret our finding concerning lexical complexity, because lexical diversity (MTLD) did improve significantly over one year, but word frequency in their writing did not. It may suggest that the students came to use a wider variety of words in their writing, but that they produced words with a similar word frequency. This result is partly different from that of longitudinal studies on L2 speech by Crossley and his colleagues, who found that both MTLD and word frequency were improved over one year. The discrepancy may be due to the difference in modes of communication (speaking or writing), in the participant's proficiency level, or in their learning environment (ESL or EFL). This issue remains to be pursued.

A more important finding of this study is that grammatical complexity was the aspect of writing that changed most over one year. Both grammatical complexity measures (STRUT and ASL) improved significantly over one year, and about three fourths of students displayed an upward trend in their developmental patterns for these measures.

It is, of course, debatable whether the improvement in these grammatical complexity measures conclusively proves that the students became more competent in the use of grammatical structures in their writing. One may argue, for example, that the increase in STRUT does not necessarily lead to more skilled writing, for professional writers in a particular genre and in a particular culture tend to use a narrower range of grammatical structures than do native English writers (e.g., Japanese scientists in

McCarthy et al.'s (2007) study who preferred to use syntactically simple sentence structures). Or, longer sentences are apt to be avoided in a particular language. If it is the learner's L1, they may avoid writing longer sentences even when they become advanced L2 writers. These criticisms are humbly accepted. What we can add here is, as mentioned above, that STRUT was useful to distinguish text difficulty for EFL readers (Crossley et al., 2008), and that ASL meaningfully differentiated learners at different proficiency levels in previous studies (Orgeta, 2003). These findings provide some support for the validity of the two indices of grammatical complexity. Moreover, our qualitative data suggests that the students in our study possibly developed or tried to develop their use of grammatical structures in their writing.

The students who were interviewed after the one-year freewriting experience recalled that they had become more reflective about their own writing. They said they aimlessly engaged in this task at first, but gradually became aware of the shortcomings of their own writing by repeating the task (e.g., too short, with too many grammatical and lexical errors). They then intentionally aimed to improve their own writing. One of their attempts to achieve this aim was to deliberately use newly learned grammatical structures.

Why was grammatical complexity significantly developed, if we presume it was, even though the authors did not give the students corrective feedback on their compositions, nor did we instruct them to heed grammatical structures for this freewriting task? The results of this study do not offer a decisive answer to this question. However, one plausible explanation is that grammatical complexity was easier to *consciously* improve than was fluency. For instance, it may be difficult for basic writers to consciously write longer, but it may be possible for them to use a wider variety of grammatical structures by intention. This might be why the students could develop grammatical complexity faster than fluency in writing.

The present study has also shown that, the student's developmental trajectories were widely varied, though there was one most popular pattern (Reverse-S). In addition, the majority of students showed non-linear developmental patters in the two measures described above. Thus, even when the EFL students made some progress in writing, they may not follow a sole, straight developmental path. We could say that the task (repetition of the freewriting) was flexible enough to allow most of the learners to develop their writing in their own way. However, *how* this was possible should be explored more extensively in future research.

Conclusions

The findings of this study have shown that one-year task-type practice had a general effect on the development of EFL learners' writing. This effect might have been even more noticeable because the students in the study were all basic L2 writers. The strength of the impact seemed to vary

with the aspect of writing. Contrary to our prediction, we found that grammatical complexity grew faster than did fluency and lexical complexity. This may imply that the development of fluency and lexical complexity in EFL writing should be investigated over a longer span, or that it may be necessary to offer a more intensive pedagogical intervention to develop these aspects (e.g., corrective feedback).

Since this study is only in its initial stage, it has some limitations. First, with regard to the analysis of text, it may be desirable to look at a wider variety of text measures. For example, we used STRUT as a grammatical complexity index, but MED could be used instead, since McCarthy et al. (2009) have shown that MED evaluated paraphrase quality better than STRUT. Second, there are various other ways to analyze developmental patterns than the one used in this study. In particular, other visualization methods such as a min-max graph may be useful in capturing a critical *phase shift* (Larsen-Freeman & Cameron, 2008) in the EFL student's writing behavior. Third, although this study has significant implications for L2 writing research in that it investigated the changes in writing with data that was collected with high frequency over a long period (30 times over one year), one year may not have been long enough to describe the changes in certain aspects of L2 writing, as the results above suggest.

To our knowledge, this study was the first to classify longitudinal developmental patterns in L2 writing by focusing on its non-linear trajectories. Therefore, it is far from describing the system in which learners learn and develop their L2 writing. In addition to overcoming the limitations mentioned above, future research should delve more into agentivity of writers, such as their motivation to write, what the writers are attempting every time they write, and how their perceptions toward L2 writing change over time.

References

Bardovi-Harlig, K. 2002. A new starting point?: Investigating formulaic use and input in future expression. *Studies in Second Language Acquisition, 24*, 189-198.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475-493.

Crossley, S., Salsbury, T., & McNamara, D. 2009. Measuring L2 lexical growth using hypernymic relationships. *Language Learning, 59*(2), 307-334.

Elbow, P. 1998. *Writing with power: Techniques for mastering the writing process*. New York: Oxford University Press.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (in press). The development of polysemy and frequency use in English second language speakers. *Language Learning, 60*(3).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, &*

Computers, 36(2), 193-202.

Henry, K. 1996. Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *The Modern Language Journal, 80*(3), 309-326.

Larsen-Freeman, D. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-619.

Larsen-Freeman, D., & Cameron, L. 2008. *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Martin, B., & Samuda, V. (2005). Integrative planning through the use of task repetition. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 37-74). Amsterdam: John Benjamins.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD).* Unpublished doctoral dissertation, The University of Memphis.

McCarthy, P. M., Cai, Z., & McNamara, D. S. 2009. Computational replication of human paraphrase assessment. In C. H. Lane & H. W. Guesgen (Eds.), *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference* (pp. 266-271). Menlo Park, CA: The AAAI Press.

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y., & McNamara, D. S. 2007. A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes, 6*, 46-77.

Ortega, L. (2003). Syntactic complexity measures and theire relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

Shaw, P., & Liu, E. T.-K. 1998. What develops in the development of second-language writing? *Applied Linguistics, 19*(2), 225-254.

Verspoor, M., Lowie, W., & van Dijk, M. 2008. Variability in second language development from a dynamic systems perspective. *The Modern Language Journal, 92*(2), 214-231.

Weigle, S. C. (2006). Investing in assessment: Designing tests to promote positive washback. In P. K. Matsuda & C. Ortmeier-Hooper & X. You (Eds.), *The politics of second language writing: In search of the promised land* (pp. 222-244). West Lafayette, IN: Parlor.

Willett, J. B. 1994. Measurement of change. In T. Husen & T. N. Postlethwaite (Eds.), *The international encyclopedia of education* (2nd ed., pp. 671-678). Oxford: Pergamon.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.