COHESION IN SECOND LANGUAGE WRITING

By

Mark Cosgrove Shea

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Second Language Studies

2011

ABSTRACT

COHESION IN SECOND LANGUAGE WRITING

By

Mark Cosgrove Shea

This study investigated the effect of a sequence of pedagogical interventions on the level of textual cohesion in the writing of high-intermediate L2 English learners in a college-level ESL program. Eight sections of a fourth-semester ESL writing course were assigned randomly to the experimental or control groups. The experimental group received no additional instructional time, but the researcher visited the each experimental section for one hour each week over a five-week period to provide a series of pedagogical interventions focused on the use of adverbial connectors, determiner + summary noun constructions, and definitional elements. After attrition, data from n = 46 control participants and n = 47 experimental participants were included in the study, for a total of N= 93 participants.

Each participant contributed three samples of timed writing in a pretest, posttest, delayed posttest design. The texts were rated by three raters, and the mean rater score was used to operationalize writing quality. Additional developmental measurements focused on the fluency and syntactic complexity exhibited within texts and the amount of lexical diversity. The level of cohesion in the texts was operationalized as a combination of sentence and paragraph latent semantic analysis scores as well as measures of adverbial connector use.

The results suggested an effect of treatment. In terms of writing quality, the experimental group scored significantly higher than the control group at posttest, and also produced more and more varied forms of the target structures. The timing and patterns of the effect of instruction

measures, combined with the lack of group differences in broad developmental measures, suggest that the intervention sequence did have a positive effect on experimental participant writing. The results also point to the difficulties of operationalizing lexical cohesion as a construct independent of overall lexical proficiency.

The results of a principal component analysis on the measures of cohesion suggested that cohesion must be operationalized as a multidimensional concept comprising measures of connector use and lexical reference chains. The analysis also suggested that, if latent semantic analysis measures are chosen as operationalization of lexical cohesion, the level of lexical diversity in the text as measured by type-token ratio, will affect the results of the analysis due to an inverse relationship between latent semantic analysis scores and lexical diversity.

To my wife,
Alexis

ACKNOWLEDGEMENTS

In completing this dissertation, I have benefited from the assistance of a number of people, without whom this project would not have been possible. I would like to thank Dr. Charlene Polio, the chair of my dissertation committee, for her guidance and support in this project as well as the beginning of my academic career. I extend my sincere gratitude to the other members of my committee, Drs. Debra Friedman, Shawn Loewen, and Paula Winke for their help and advice during this project. I would also like to thank all the faculty of Michigan State University's Second Language Studies Program, who have helped me grow as a scholar, researcher, and teacher during my four years here. For those four years, the SLS Program has provided me with support in the form of assistantships, a research grant, and a fellowship in my final year, all of which have been of immense help in allowing me to complete this dissertation. I also need to thank Joan Reid, graduate secretary for the SLS Program, for her help and patience with my inability to complete paperwork correctly and/or promptly.

A number of instructors in the Michigan State English Language Center were generous with their time, classrooms, and ideas. I would like to thank Mariah Shafer, Carlee Salas, Andrew McCullough, Alice Poole, Dave Ragan, Justin Cubilo, and Roman Chepyshko for their help. I would also like to thank their students, those who volunteered as participants and those who did not, for their patience and attention.

Finally, I would like to thank my wife, Alexis Allen, for her patience, love, and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION AND REVIEW OF THE LITERATURE

In their review of the literature on cohesion in second language writing, Jimenez Catalan and Moreno Espinosa (2005) identified four major strands of research: (1) the frequency of cohesive devices; (2) the relation between the frequency of cohesive devices, coherence, and writing quality; (3) comparisons between the use of the cohesive devices used by L1 and L2 writers, and between L2 writers of different L1s; and (4) the effect of genre or topic on the types of lexical cohesion used. A wider reading of the cohesion literature confirms a surprising lack of research investigating the effects of instruction on the use of cohesive devices in learner writing.

The present study addressed this gap in the literature by studying the effects of pedagogical intervention on the amount of cohesion in learner writing. Eight sections of a university-level ESL writing course (totaling 93 participants) were assigned to experimental (n = 47) or control (n = 46) conditions. Writing samples were collected before, immediately after, and four weeks after a five-week sequence of instructional interventions presented for one hour each week. A preliminary analysis used principal component analysis to determine whether different cohesive features, namely, lexical and conjunctive cohesion can be treated as a single construct or if cohesion should instead be considered a multidimensional construct. The results indicated that cohesion is indeed a multidimensional construct, and further, that other aspects of lexical proficiency, such as the type-token ratio of a text, may influence the level of cohesion present in a text**.** The texts were rated by three raters on a 90-point, five-category analytic scale as an operationalization of writing quality. The writing of the participants was compared across group and time in order to determine whether the intervention sequence had a significant effect on the

level of cohesion in learner writing, and a second analysis investigated the relationships between treatment effect, level of cohesion, and raters' judgments of writing quality.

Review of the Literature

This section introduces some of the key theoretical constructs used in the present study, introduces some of the prior research on measuring textual cohesion, and provides justification for the choice of intervention target structures.

Cohesion

Halliday and Hasan's (1976) seminal work on textual cohesion is the basis of much of the current theory on the topic. Examining what quality causes a series of sentences to cohere into a single text, Halliday and Hasan identified five cohesive relations that can signal relationships between units of text, a cohesive relation being identified as when one element of a text relies on another for its semantic interpretation (Halliday & Hasan, 1976, 1985).

Three of these relations, *reference, substitution, and ellipsis*, make use of syntactic operations and closed-class words. *Reference* cohesive ties include personal and demonstrative pronouns as well as comparatives (e.g., *I met **a man** on the way to St. Ives. **He** had seven wives*). *Substitution* ties replace a word, a verb phrase, or an entire clause using closed-class words not included in those listed under the *reference* category (e.g., *do* to replace a verb: *She doesn't **like the car** but I **do***.). *Ellipsis* ties refer to substitution by 'zero' (e.g., *She can **drive the car** but I can't* _____). *Lexical cohesion* is created through the repetition of lexical items or use of synonymous items throughout various sections of a text (e.g., *Researchers working on a vaccine are faced with many **difficulties**. The first **challenge** is . . .*). The final type of cohesive relation is *conjunction*, which makes use of coordinating and subordinating conjunctions as well as

adverbial connectors to create explicit connections between propositions (e.g., *The test was ruled a failure.* ***Therefore,*** *the project was scrapped).*

In their original work, Halliday and Hasan (1976) emphasized the more systematic, grammatical means of creating cohesion, devoting less time to lexical cohesion as its idiosyncratic nature rendered it less amenable to theoretical analysis. However, in subsequent work, Hasan (1984) suggested that cohesive ties created by lexical repetition are in fact the true source of cohesion within a text. This idea was further developed by Hoey (1991), who presented a theoretical framework built around the creation of cohesive *chains* which are created by repeated, synonymous, and hyponymous lexical items, as well as reference relationships created by pronoun use. Hoey's framework for analyzing cohesion also simplified the distinctions between Halliday and Hasan's three types of grammatical cohesion by conceptualizing them, along with lexical items, as links in cohesive reference chains. Hoey did not argue that more syntactic relations, such as pronoun reference, were irrelevant, but simply that they did not need to be considered as separate from the creation of cohesive reference chains through lexical repetition.

These cohesive chains refer to particular concepts, entities, or actions, and while a particular referent may occur most often in a single paragraph, some key ideas in a text may occur throughout. In the sample of learner writing in Figure 1, it is possible to see this interaction (the example is not intended to provide an exhaustive representation of all potential reference chains): the argument *South Korea,* the main topic of the essay, appears throughout the first two paragraphs of the text, in all but 2 of the 9 T-units. Compare that with the more localized chain formed by *war* in T-units 1 and 2, in which the writer is providing some historical background

for the country's current problems. In the second paragraph, the country's president becomes a focus, and a new cohesive chain is created between T-units 7, 8 and 9, with T-units 8 and 9 also participating in the *South Korea* chain.

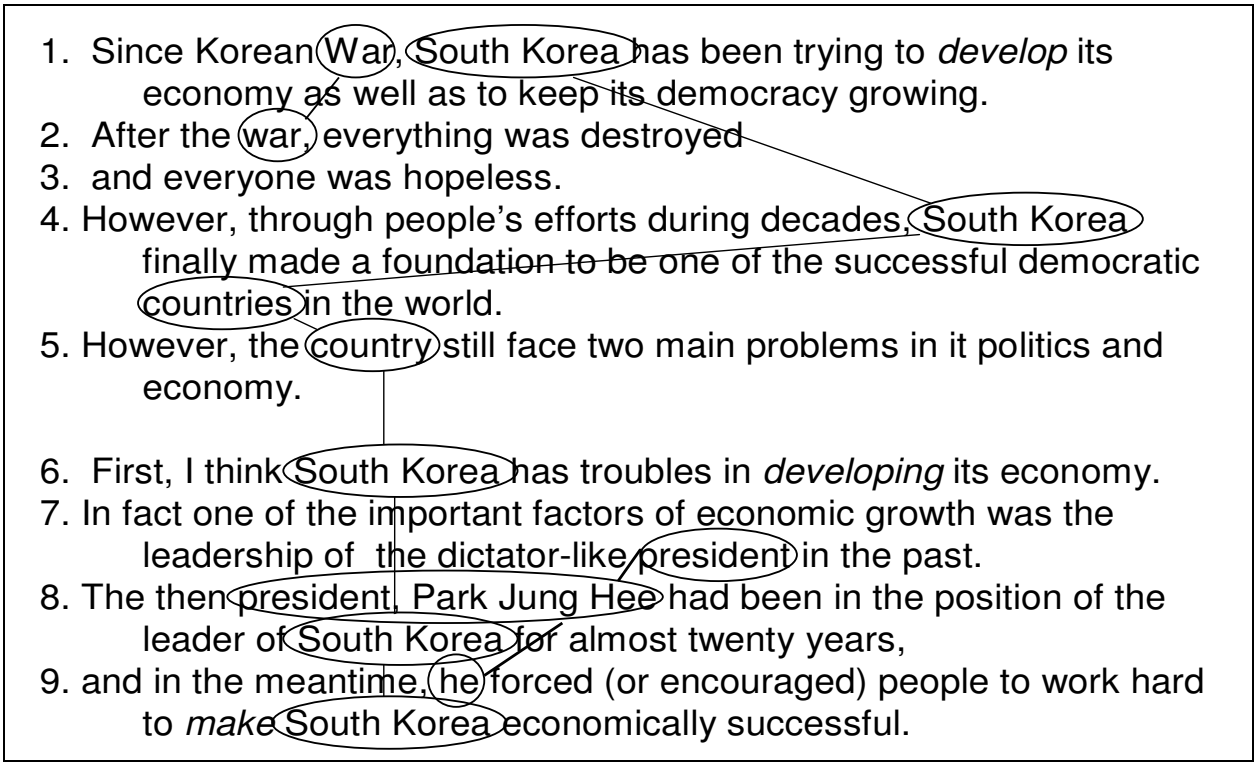

Figure 1: Cohesive chains through two paragraphs of a learner text

This intertwined distribution of cohesive chains makes lexical cohesion much more than a count of how many times a writer repeats a lexical item or how many connector words are employed; the level of cohesion present in a text is affected by the choices writers make in their efforts to organize their thoughts and express their ideas, in the discourse structures they employ and the lexicogrammatical choices they make as they progress from sentence to sentence. A key issue for the use of this framework as a research tool is its ability to be quantified and replicated: a problem discussed in the following sections.

Cohesion and writing quality

The construct of cohesion represents one very specific aspect of a text; thus, a text may contain many cohesive features but still not be considered effective. There is much beyond semantic ties between sentences that goes into creating a meaningful and effective text: genre, text organization and information structure, propositional content, and metadiscourse features, along with lexicogrammatical competence. The term *coherence* is generally used to refer to the combination of all these factors and their interaction with a reader's understanding to create a unified meaning.

Although researchers have adopted various definitions of cohesion and coherence, Hasan's (1984) explanation represents the most commonly used distinction between them: "cohesion is a property of the text, and . . . coherence is a property of the reader's evaluation of the text" (p.12). This distinction characterizes cohesion as a quality that can be measured directly from the text, though researchers have adopted many different ways of doing so, while the quality of coherence must be measured as it is perceived by a reader. By virtue of these definitions, coherence has a clear link to writing quality, since it exists in the mind of the reader, while cohesion rests in the text itself and may be noticed or not noticed by the reader. In addition, to the extent that cohesion is noticed, it may not be regarded as a helpful quality by the reader.

There is a relatively extensive body of research which has investigated the potential connection between the use of cohesive devices in a text and the quality of the text. Several studies have examined the relationship between writing proficiency scores and the use of cohesive devices. The effectiveness of lexical cohesion has received the most support, with

mixed results for grammatical cohesion as described by Haliday and Hasan's (1976) original framework.

In a study of L1 English freshman compositions, Neuner (1987) found that the total number of cohesive ties did not distinguish between a sample of 20 good and 20 weak L1 English essays written by college freshmen, but did find that longer cohesive chains, in addition to other measures of lexical quality, were characteristic of the better essays. Neuner's results suggest that it is not simply lexical ties, but the extent and sophistication of the lexical chains that contribute to stronger essays. This is similar to the results of a study by Ferris (1994), which reported that low rated-learner essays made greater use of lexical repetition than higher-rated essays.

Bae (2001) found that, for young learners (i.e., first and second grade), the amount of referential and lexical cohesion correlated highly with writing quality, which Bae operationalized as the sum of grammar, content, and coherence measures, and that those two types of cohesive device were significant predictors of coherence. Liu and Braine (2005) found that the scores of learner essays correlated with the total number of cohesive devices in a text, and correlated highly with the number of lexical cohesive devices used. However, the researchers pointed out that this result might be a function of the overall higher lexical proficiency of the more competent writers. Grant and Ginther (2000), in a study focusing on the feasibility of identifying differences in L2 writing proficiency through computer-tagging found that two cohesive devices, conjuncts and demonstratives, were used significantly more in essays scoring a 5 on the *Test of Written English* (*TWE)* than essays scoring a 3 or 4. Reynolds (2001), using writing development measures rather than proficiency scores, found that lexical cohesion was the best

predictor of variance in writing development measures in his three-predictor regression model (lexical repetition, L1/cultural background, writing topic). Taken together, these results highlight an important consideration in research on cohesion: it is generally some subset of cohesive features, most often including lexical cohesion, that displays a positive relationship with writing quality measures.

Other research has focused on the perception of cohesive features by essay raters. Chiang (2004) examined the effects of discoursal and grammatical features on the evaluation of learner writing by NS and NNS professors. Chiang found that 27 of 30 raters relied on discoursal rather than grammatical features as a basis for judging "overall essay quality." In addition, 2 of Chiang's 20 cohesive subfeatures: quality of sentence transitions in the absence of junction words and appropriate use of paraphrase and equivalent words, were the best predictors of overall essay quality. Chiang's very specific assessment instrument does however raise the question of whether a rater working without it would be sensitive to the same factors when assessing a learner text. In contrast to Chiang's findings, Watson Todd, Khongput, and Darasawang (2007) found little connection between cohesive breaks in learner texts and feedback given by teachers. Watson Todd et al. used Hoey's framework to identify sentences which had no relationship to other sentences in the text. These were identified as breaks in cohesion, and instructors' written comments were analyzed to determine whether they addressed these breaks.

The biggest difficulty in linking cohesion and writing quality, and an important point to remember when devising pedagogical materials to promote the use of cohesive devices, is that not every good essay is good in the same way. Jarvis, Grant, Bikowski, and Ferris (2004) used

cluster analysis to create profiles of highly rated essays. They found that there is not a single profile of highly rated texts, and while text length is perhaps the most influential factor, types of highly rated essays differed in their relative use of a variety of lexicogrammatical features. Of 8 essay profiles, 2 demonstrated high relative use of demonstratives, and 1 demonstrated high relative use of conjuncts, meaning only 3 profiles included some form of cohesion as a feature. However, the features Jarvis et al. included in their analysis focused on frequency counts of particular parts of speech or grammatical features such as tense or voice. There was no measure that represented the presence, interaction, or extent of cohesive chains.

Measurement of cohesion

One of the difficulties in synthesizing the research findings on cohesion stems from the fact that researchers have not employed a consistent list of cohesive features in their measurement of cohesion. This is a natural outcome of differing research aims and ambiguity in the reporting of criteria and procedures used to identify cohesive devices. I prefer to attribute the lack of detail to space limitations rather than a lack of rigor, but the effect on subsequent research is the same. It is often difficult to know if the disagreements between study results represent legitimate differences in the data, or are artifacts of differing selection and coding criteria. For example, Liu and Braine (2005) cast rather a wide net when selecting cohesive devices, counting the definite article *the* as a token of a reference cohesive device, which would only be justified in certain contexts, for example those covered by the second-mention pedagogical rule. A second difficulty in interpreting or replicating Liu and Braine's (2005) results lies in the fact that, as written, the study does not make clear whether the *conjuncts* category includes only adverbial connectors or includes coordinating conjunctions as well. In Milton and Tsang's (1993) corpus-

based study of connector use, every token of a connector (e.g., *and*) was counted, a practice which does not differentiate between a token used within a nominal phrase (e.g., *chocolate and vanilla*) and one used to link clauses.

In addition, the hand-coding of lexical chains can become so time consuming and complicated that the effectiveness of the research is severely limited (Hinkel, 2005). Hoey's (1991) framework was developed in a monograph that presented the analysis of just a few texts. Two of the studies that report results clearly supporting lexical cohesion analyzed relatively short texts: for example, Bae (2001) worked with young learners' texts (mean number of words = 67.5) and Reynolds (2001) analyzed timed texts produced by NS and NNS writers (mean number of words = 249). The time required to perform the same coding on extended texts on a scale necessary to create an effectively-sized corpus quickly threatens to become prohibitive.

Beyond logistical constraints, the manual coding of lexical cohesion relationships poses possibly insurmountable challenges to the production of replicable methods and results. An instructive example of the difficulties in this type of coding for cohesion is offered by Morris and Hirst (2005; see also Morris, 2004 for an additional report of this data) who examined how L1 readers' judgments of lexical connections demonstrate some core similarities, but also a wide range of subjectivity.

Morris and Hirst (2005) asked a set of readers to read 1-2 page, general interest texts (*Reader's Digest* articles) and identify word relationships they saw therein. Provided with an array of coding sheets and colored pencils, the participants worked through the texts, first identifying groups of words bearing some semantic relationship (e.g., *police, cop, jail, safety* [examples not taken from Morris & Hirst]) , then identifying word pairs within that group (e.g.,

police and *cop*, *siren* and *police car*), describing the meaning of the word group in the text (e.g., *the side of law-and-order*), and then describing the relationships between the word pairs (*police* and *cop* are synonyms; *a police car* has a *siren*).

Morris and Hirst's (2005) analysis began by including only those word groups identified by at least 4 of their 9 subjects. This numbered 11 word groups, but that fact that they set their cutoff below half the number of their participants suggests that there was likely a large amount of disparity between the word groups chosen by the participants. This is not to criticize the work done by Morris and Hirst, but rather to reiterate the difficulty of using this type of coding as a replicable, quantitative research instrument. The average rate of agreement between all possible pairs of participants in identifying the word groups was 63 percent.

A second step, in which the rate of agreement in identifying word pairs was calculated, indicated that participants had much lower agreement when identifying word pairs. Only 13% of the word pairs were marked by more than 50% of the subjects. However, for any pair of words that was identified by more than one subject, the relationship between those words was found to be reliably identified (86% agreement)

What this suggests is that while a general set of conceptually related words is identifiable, it is a harder task to identify relationships within that set, though once identified, a relationship is generally easy to categorize. However, Morris and Hirst (2005) point out that the majority of the relationships identified were *not* the classic lexical relations of *synonymy, antonymy, hyponymy,* and *meronymy.* Morris and Hirst also report the frustration and fatigue that characterized pilot participants' efforts to identify word pair relationships, and in the reported study, asked participants to focus only on *core* relationships. Lexical relations seem to be crucial to effective

cohesion of a text, but the identification of those relations relies largely on intuitive and associative processes that are difficult to access and discuss explicitly, and while the quality of these relationships may be easily or at least, reliably, identified, the quantity and extent of these relationships might vary considerably between coders

The development of software to analyze cohesion and coherence in texts provides a possible solution to this problem. For example, a software package, Coh-Metrix, designed to analyze the cohesive features of texts, including sentence and paragraph-level LSA scores (McNamara, Louwerse, Cai, & Graesser, 2005), has been developed and made freely available online. Originally used to investigate the readability of texts, recent research has extended the use of the software to evaluate writing, and second language writing in particular (e.g., Crossley & MacNamara, 2009). Results of a comparison of texts produced by L1 and L2 English writers indicate that the repetition of arguments across sentences and the latent semantic analysis measures of sentence relatedness differentiated between L1 and L2 texts.

A key measure used in the automatic analysis of textual cohesion and coherence is Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). LSA is both a theory and a method which has been developed to analyze the usage of words based on they contexts in which they appear. According to Landauer, Foltz and Laham (1998), LSA can be conceptualized in two ways. First, it is a "practical expedient" (p.5) for estimating the relationships between words and the segments of texts (i.e., sentences, paragraphs, and whole texts) within which they appear, as well as the substitutability of words (i.e., how likely it is that a word could replace another word in a particular context). Second, LSA is a model of how the human mind acquires, represents, and uses knowledge. In the proposed study, the focus will be

on the practical expedience LSA offers, rather than the theory of mind it represents, and no claims will be made as to the validity of its representations of knowledge or learning.

Rather than looking at relationships between individual words, LSA investigates the relationships between words and larger local contexts (e.g., sentences or paragraphs) in order to "capture . . . how differences in word choices and differences in passage meanings are related" (Landauer et al., 1998, p.5). It does this by assigning a lexical item a numerical value which represents an "average" of the meanings of all the passages in which the word has appeared. The meaning of a segment of text is then represented by the average of all the words which appear in it. LSA assigns these values by reducing the dimensionality describing a word or passages meaning. Landauer et.al. describe this reduction of dimensions as similar to the practice in linguistics of representing a lexical item as a collection of features (e.g., [+animate, +countable, -human]), but emphasize that there is no concrete connection between these features and the LSA dimensions assigned to a word.

This reduction of dimensionality is carried out through a statistical process similar to factor analysis known as singular value decomposition (SVD). The resulting similarity scores are measured as the cosines between the vectors, with higher scores indicating greater semantic similarity between text segments. The University of Colorado-Boulder maintains a web-based package of Latent Semantic Analysis (LSA) tools, supported by a recently published book on the subject (Landauer, Dennis, McNamara, & Kintsch, 2007). There is a recent and growing body of research on LSA and L2 production that indicates LSA can represent cohesion and coherence in learner production and does correlate with traditional measures of language development. In addition to the study by Crossley and MacNamara (2009) cited above, a longitudinal study of six

English learners by Crossley, Salsbury, McCarthy, and MacNamara (2008) found that the LSA scores of L2 English learners' spoken production increased significantly with time spent studying in a second language context, and that the frequency of negotiation for meaning episodes correlated negatively with LSA scores. In addition, the lexical diversity of the learners increased concurrently with the increase in LSA scores.

The relationship between lexical cohesion, represented by LSA scores, and language development or writing proficiency has received inconsistent support in the literature, however. From a reading perspective, Crossley (2008) suggests that texts with less cohesion promote greater retention for skilled readers, as the breaks in cohesion promote deeper processing of the content. Focusing on cohesion in writing, Foltz (2007) suggests anecdotally that texts with the highest levels of LSA-measured cohesion are often the lowest-rated, as frequent repetition of lexical items will result in high LSA measurements but might be judged excessive by human readers.

Bestgen, Lories, and Thewissen's (2010) results align with Foltz's predictions; they found a small, negative correlation between automatic measures of cohesion (their own LSA measures, confirmed by Coh-Metrix measures) and trained raters' judgments of the coherence of L2 texts according to the Common European Framework of Reference (CEFR) descriptors of coherence in writing. Comparing their results to those of Crossley et al. (2008), Bestgen et al. offer a number of explanations for the fact that their results differ. They suggest that modality (written vs. spoken), proficiency level (intermediate to advanced vs. beginners) and assessment (cross-sectional rating vs. longitudinal development) may all have contributed to differences in findings.

In relation to assessment, one point that Bestgen et al. (2010), mention but that may deserve closer attention is that the CEFR coherence framework places emphasis on a variety of cohesive devices. Bestgen et al. suggest that raters may have been focused on more salient cohesive devices, such as adverbial connectors, and not as focused on the lexical cohesion that LSA measures. In light of the indications of the difficulty in consistently identifying lexical cohesion relationships, research trying to relate LSA to rater's judgments may benefit from a rating instrument that, while not necessarily forgoing explicit judgments of coherence, at least asks raters to provide an impression of overall writing quality. These more global judgments may actually be more effective at capturing the effect of complex interaction of lexical and rhetorical features that comprise the theoretical construct of cohesion.

Teaching Cohesion

While the literature detailing the cohesive features of learner writing is extensive, there is a very small amount of research that has been done on the effective teaching of cohesion. Much of what does exist, though often informed by theory and experience, does not have the benefit of empirical support.

Hinkel's text on academic writing instruction (2004) contains a chapter devoted to the teaching of cohesive devices. Hinkel suggests providing learners with explicit instruction on the topic-comment rhetorical pattern (as presented in Williams, 2002), directing them to generally repeat a word from one sentence to the next to create more extensive lexical chains, and to explicitly teach general nouns. Hinkel also indicates potential areas of difficulty for learners, including parallel structure, inappropriate exemplification and clarification, and the misuse and overuse of adverbial connectors. Swales and Feak (2007), in their textbook aimed at graduate

student learners, address several of the same issues. In particular, they include a section on general nouns used to summarize preceding points, although they present the techniques in their student-facing book less in terms of developing cohesion than as the linguistic traits of a particular discourse community.

Suggestions by McGee (2009), when compared with Hinkel's (2004), highlight some of the difficulties in preparing pedagogical interventions for a topic as fuzzy as cohesion: both Hinkel and McGee recommend instructing learners in the use of hypernyms, or general nouns, but whereas Hinkel also recommends encouraging students to repeat a word from sentence to sentence, McGee repeatedly refers to learners' use, or overuse, of repetition as problematic. These conflicting recommendations are similar to the contradiction between Hinkel's warnings against the common overuse and misuse of adverbial connectors by learners and research (e.g. Grant & Ginther, 2000) that identifies such features as characteristic of more highly-rated essays.

Lee (2002, see 2000 for a description of materials) delivered treatments aimed at improving coherence in learner writing, with cohesion included as one of the six foci of the treatment. The foci, moving from the macro to micro level, are: (1) purpose, audience, and context; (2) macrostructure; (3) topical development and organizing information; (4) propositional development—elaborating, illustrating, exemplifying, (5) cohesion: reference, substitution, conjunctions; (6) metadiscourse: topicalizers, hedges, and attitude markers. The lessons were presented based around text analysis of modified reading passages and the identification of the problematic realization of coherence features in passages.

Lee (2002) describes her study as a preliminary investigation into the feasibility of such treatments, and her detailed description of the treatment and qualitative reports on student

attitudes toward the treatments provide an excellent account of the implementation of the treatments, but stops short of answering the question of whether the treatments had a positive effect on the coherence of the students' writing. One of the key pieces of data missing from her analysis is a comparison of coherence features across first drafts produced over time; Lee confined her investigation to changes across revised versions of the same text, which gives less insight into how students deploy their potentially developed repertoire of coherence features in a new writing task. At the same time, the investigation of revised texts is very desirable from the standpoint of ecological validity, as much of the academic writing that learner's are being prepared for will be of the untimed, revised variety.

Some positive findings from Lee's (2002) study that bear on the design of the proposed study's materials are that the treatments empowered students by giving them specific techniques to use in improving their writing and raised awareness about coherence. Lee also reported that the integration of reading and writing through the text analysis activities was effective. Some negative aspects of the treatments were that some students reported feeling overwhelmed, and in some cases bored, by the extensive text analysis—a finding that was duplicated in the pilot testing of materials for the proposed study and addressed in the experimental materials by emphasizing scaffolded production and reflection over text-analysis. Further, Lee felt that students may have come away from the treatment with the idea that the coherence view of the writing process was the only valid one, a serious problem given Jarvis et al.'s (2004) findings. Finally, Lee's treatment of cohesion relied on Halliday and Hasan's (1976) somewhat complicated taxonomy, rather than the arguably clearer treatment of cohesion as a series of interwoven chains of reference, supplemented by conjunction relations.

Treatment Targets

As the above discussion shows, an intervention designed at improving cohesion in learner writing should meet several criteria. First, it should reflect current theory in the field by privileging lexical cohesion as the primary cohesive tie. The chief difficulty in taking this approach lies in the fact that within the varieties of lexical cohesion, the more effective forms are the more sophisticated and require a greater level of lexical proficiency to employ. To create effective lexical cohesion, a learner needs to employ appropriate synonyms, hyponyms and hypernyms, and part-of-speech transformations. In the absence of these more sophisticated lexical relations, interventions emphasizing lexical repetition risk promoting a more basic writing style that has been connected with lower quality writing (see Silva, 1993, pp.667-668 for a review).

In addition, the intervention should be flexible enough that it can accommodate multiple perspectives on the writing process. It should provide learners with the chance to analyze sample texts with problematic cohesive relationships, and it should do so in a way that is engaging and leaves time for other classroom activities and discussions. Finally, it should be remembered that the goal is not increased cohesion *per se*, but rather the potential increase in writing quality that may result from an increase in effective use of cohesive devices.

Taking these considerations into account, the proposed study chose three main areas of pedagogical focus, based on their pedagogic relevance to the needs of the target population and the likely effect that changes in these areas might have on the level of cohesion in a student text. The language topics and materials were modeled on materials or suggestions provided in Swales and Feak (2007), Hinkel (2004), Lee (2000, 2002), McGee (2009), and Salkie (1995).

Definitional elements. The first focus aims at developing student writers' ability to define technical or key terms used in their texts. *Definitional elements,* as treated in the present study, are a similar, though more specific, concept to Lee's (2002) *propositional development.* In Swales and Feak (2007), this technique is introduced in the context of graduate-level technical or scientific writing. The target population in the current study is not necessarily at the level of academic or linguistic development in which they are called on to write on highly specialized topics. However, all students have a communicative need to provide further definition and clarification for terms they include in their writing. In the example below, taken from an exploratory corpus of student writing collected in preparation for this study, a student uses the term "brunch culture" in a discussion of the increasingly materialistic values in his home country (emphasis added).

> So now in the South Korea, you can easily find the luxury stores and the luxury restaurants everywhere. Also we have *brunch culture* now. People who want to follow these luxury things; they sell the body and borrow money from the capital companies.

It is not immediately clear from the context what the student means by this phrase. At the same time, it is not necessarily true that this is a lexical accuracy error. The student may be using a neologism or translated term to express a meaning that is simply unfamiliar to the reader. The revision needed here may not be a replacement, but rather an elaboration.

In the following text from the introduction to a scaffolded writing created with participants during a pilot intervention session for the present study, note the clarification of the term *communication* (emphasis added).

(1) *Communication is the way we keep in touch with our friends and family*. (2) People live busy and fast lives, and communication is important for people with a fast lifestyle. (3) *Communication can take many different forms such as Facebook, chatting on Yahoo, or even calling your friend on Skype.* (4)These technologies help us stay close even when we are busy or far away. (5) Given this fact, it does not seem possible to say that technology has destroyed communication.

From the perspective of communicative effectiveness, the addition of sentences (1) and (3) arguably improve the quality of the introduction, which originally contained only (2), (4) and (5). From the perspective of an analysis of cohesion, the text now includes more lexical resources available to enter into cohesive chains by explicitly linking the term *communication* to the action *keeping in touch* and the list of technologies in sentence (3).

Figure 2 displays the LSA cohesion measures for each version of the paragraph; the initial version's scores are on the left, and the elaborated versions scores are on the right. Each pairing that includes one of the definitional elements results in a higher LSA score than either of the pairs of the original three sentences. This demonstrates how definitional elements can contribute to higher levels of cohesion in a text.

Unrevised LSA score		Sentence		Revised LSA Score
		Communication is the way we keep in touch with our friends and family.	1.	
	1.	People live busy and fast lives, and communication is important for people with a fast lifestyle.	2.	.3
		Communication can take many different forms such as Facebook, chatting on Yahoo, or even calling your friend on Skype.	3.	.24
.1	2.		4.	.17
.08	3.	These technologies help us stay close even when we are busy or far away.	5.	.08
.09		Given this fact, it does not seem possible to say that technology has destroyed communication.		.2
		Mean		

Figure 2: LSA scores of a passage and elaborated passage

Summary nouns. The second treatment focus aimed at increasing the appropriate use of what Swales and Feak (2007) refer to as *summary nouns*, examples of which are *attitude, difficulty,* and *problem*. A writer using the structure *this+ summary noun* (e.g., in sentences (4) and (5) in the example above) is able to refer to more specific entities and propositions in previous or subsequent sentences and thus elaborate and develop their ideas more fully. In a related strand of research, Flowerdew (2003, 2006) has written on the use of *signaling nouns* in academic writing and learner writing in particular. Under this term, Flowerdew collects a variety of more specific noun types referred to by previous writers (e.g., *general nouns* (Halliday & Hasan, 1976); *anaphoric nouns* (Francis, 1986), *metalanguage nouns* (Winter, 1992)). It should be noted that Swales and Feak's term emphasizes the anaphoric use of this type of noun while Flowerdew's emphasizes the cataphoric, though both uses are possible in each version. In this

proposal and subsequent study, Swales and Feak's term *summary noun* will be used for consistency, even when discussing previous research which employs a different term.

As defined by Flowerdew (2003), a summary noun is an abstract noun which does not have a clear meaning without its context. A subsequent study by Flowerdew (2006) found that in a corpus of graded essays written by L1 Cantonese learners of English, the essays receiving the highest grade contained significantly more summary nouns per 100 words (a difference of just under 1 token per 100 words) than the lowest graded essays.

Gray and Cortes (2011) frame their corpus-based study of summary nouns used in published academic writing in terms of a counterargument to prescriptive rules against the use of the pronominal, rather than determiner, forms of *this* and *these* in style manuals (e.g., *APA*, *Chicago*). Gray and Cortes argue that as many advanced L2 writers make use of these guides as a form of writing support, the non-evidence-based guidelines may lead these writers to an inaccurate understanding of academic writing conventions. Their overall finding is that, counter to prescriptive guidelines, roughly 20% of the tokens of *this* and *these* in journals from two academic domains are pronominal.

While Gray and Cortes' (2011) finding is an important one, the converse point, that 80% of the occurrences of *this* and *these* in their corpus were as determiner for NPs, lends empirical support to the inclusion of theses structures in pedagogical interventions designed for intermediate learners. A preliminary investigation of the pilot corpus collected for this study suggests that while student writers do use *this + noun* structures, it is rare for a summary noun to be used to encompass an entire concept or connect a more specific noun to a general concept. Instead, the *this + noun* construction generally repeats a noun from a previous sentence. When

student writers in the pilot corpus attempted to make summarizing connections, they more frequently used the pronoun *it*, resulting in passages similar to the example below (emphasis added).

> [1] They should calm down and think what have done today and whether *it* is right or wrong. [2] *It* is good for their career and helps them to get a high position in your company because you always correct your mistake quickly by usually think alone. [3] I think *it* is also relate to the culture in America. [4] But *it* is quite different from China.

As the passage progresses, it becomes increasingly difficult for a reader to assign a referent to the pronoun *it*: the token in sentence 1 clearly refers to the preceding noun clause, but the token in sentence 2 may refer to the same noun clause or the act of thinking. By sentence 3, the referent seems to have shifted, but to what entity can't be determined with any certainty.

A pedagogical treatment focusing on summary nouns has the advantage of providing learners with the opportunity to create more sophisticated lexical cohesive ties without devoting a large amount of instructional time to topic-specific vocabulary of limited general use. Taking the above passage as an example, such a technique might also have a substantial effect on the quality of a learner's writing if it provides a technique to improve the confusing string of *it*s contained in the passage.

Connectors. The third treatment focus aimed to increase the judicious use of connector words, particularly adverbial connectors. As text linguistics theory has emphasized the importance of lexical substitution, a corresponding dissatisfaction with the importance of conjunctive adverbials can be seen emerging in the pedagogical writing literature. Hinkel (2004) expresses this dissatisfaction with the role these connectors play in student writing:

> The major problem with sentence connectors in L2 writing is that, because these linkers are easy to understand and use, NNS writers employ far too many of them in their text. The second issue with these features of academic prose is that the use of sentence transitions does not necessarily make the L2 academic writing cohesive or the information flow easy to follow (p. 292-293).

Hinkel suggests that a useful activity is to have learners remove all the connectors from a text in order to see how little difference there is. This is almost literally a mirror image of an exercise in Swales and Feak (2007) which invites learners to read two versions of a passage to see the improvement in the passage using connectors. While research has shown that learners often do overuse adverbial connectors, the results are far from conclusive (see Shea, 2009 for a review). Further, the teaching of connectors offers an opportunity to discuss the types of relationships between propositions. This is a less easily described form of cohesion than that created by chains of lexical reference, but one that is no less important to effective writing. Based on their continued inclusion in the theoretical framework of cohesion and the impact an understanding of connectors might have on propositional development in learner writing, they were selected as the third focus of the intervention.

Summary

The theoretical construct of cohesion accounts for connections between sentences and paragraphs within a text. To be considered a cohesive tie, these connections must be explicit in the text rather than created through a reader's interaction with the text. Over the past thirty years, the theory of cohesive relations has shifted towards emphasizing lexical chains running through a text rather than grammatical relations between sentences, and many of Halliday and Hasan's

(1976) original grammatical categories of cohesive tie can be reconceptualized as links in these chains. A survey of the recent literature, however, finds that the empirical research conducted during the same period does not consistently reflect this theoretical change.

There is some ambiguity in the literature regarding cohesion's relationship with coherence or writing quality, but there is enough evidence to warrant further investigation. Often, conflicting research results seem to stem from how granular the concept of cohesion is treated in the study and what cohesive devices are investigated.

While there are a number of studies that present descriptive reports of the cohesive devices used by L2 writers, and many of these studies investigate the link between the use of those cohesive devices and writing quality, only a few studies investigate the effect of pedagogical treatments on learner use of cohesive devices.

Research Questions

In response to the above gaps and inconsistencies in the existing literature on cohesion, the present study investigated cohesion in learner writing, using a framework that emphasized lexical cohesion and integrates the use of connectors. The study addressed the following research questions and associated hypotheses:

RQ1: Can cohesion be represented as a single factor, or should it be treated as a multidimensional construct (i.e., lexical and connective cohesion)?

The first research question is answered by the results of a principle component analysis (PCA), an exploratory statistic, and no a priori hypothesis is associated with it. However, an informal analysis with a small set of pilot data suggests that different forms of cohesion may indeed load onto a single underlying factor.

RQ2. What are the relationships between cohesive devices (lexical and conjunctive) and measures of writing quality?

H2: The overall level of cohesive devices will not correlate with writing quality. More sophisticated forms of lexical and connective cohesion, operationalized respectively as high LSA scores in conjunction with high measures of lexical development and a variety of connector types will correlate with raters' scores.

RQ3: Can learner use of cohesive devices be modified through instruction, and is there a corresponding change in perceived writing quality?

H3: There will be a significant increase in the use of the structures presented in the treatment sessions, as well as a significant increase in the overall use of cohesive devices. There may be a corresponding increase in measures of writing quality.

CHAPTER 2: METHOD

Participants

Context

The participants were all enrolled in the fourth semester (high-intermediate) of an Intensive English Program at a large research university. The students took four classroom hours of English instruction per day, four days per week (total 16 classroom hours). Within the skills-based curriculum, two hours per day were spent in a writing and content class which also incorporated a focus on grammar instruction, although grammar instruction was present throughout the curriculum.

Recruitment and Inclusion

Recruitment was done through intact classes (referred to as "sections" hereafter). First, the section instructors were approached and asked to participate in the study. For those instructors who agreed, the section was randomly assigned to the control or experimental group. The researcher then visited each section to obtain consent from the students to become participants in the study. In addition, data from two sections taught by the researcher, collected as descriptive data prior to the development of the present study, was included in the control group. Data was collected in three timed writing sessions, a pre-test, post-test, and delayed post-test phase. The experimental group received hour long pedagogical interventions between the pre-test and the post-test. In order to be included in the study, participants had to consent to participate and be present for the three timed writing sessions. In addition, experimental participants had to be present for 4 of the 5 pedagogical intervention sessions. To balance this *de facto* attendance requirement, the attendance for the control sections was reviewed and any participant who was absent for more than 5% of the classes (4 classes) was excluded. For the

experimental and control groups, 68 and 67 participants initially agreed to participate. After excluding those participants who did not meet the criteria, 47 and 46 participants remained, for a total of N = 93 participants represented in the reported results.

In addition it should be noted that data from two sections, one control and one experimental, were excluded from this study. In the first case this was due to the fact that the instructor did not administer the agreed-upon prompts, choosing instead to ask students to write in various genres and for shorter or longer amounts of time. In the second case, delays in acquiring the delayed posttest data from the instructor prevented the section's data from being included in the analysis reported here, although the data was ultimately obtained by the researcher.

Language background

In all, participants from eight sections were included in the study, with four sections assigned to the control and experimental groups, respectively. A learner background survey was given to each participant (see Appendix A). The results are presented in Table 1 and Table 2. The majority of participants in the study are L1 Chinese, with L1 Arabic and L1 Korean also comprising substantial percentages of the participants.

Table 1: Participant L1with percentage of group represented

L1	Experimental	Control
Chinese	31 (.66)	21 (.46)
Arabic	6 (.13)	14 (.3)
Korean	7 (.15)	7 (.15)
Japanese	2 (.04)	1 (.02)
Swahili	1 (.02)	0
Turkish	0	1 (.02)

Equality of groups

The decision to use intact sections was made for a number of reasons, most stemming from logistic concerns, the avoidance of attrition, and the balance of instructional time. Ultimately, the hope was that while differences might exist between individual sections, the combination of these sections into larger groups would balance these differences. It is of course important to investigate possible factors that might have influenced the performance of these sections. The equality of sections was examined through a number of measures.

First, the language learning background of the participants was gathered through a survey (Appendix A). A generalized profile of a participant in this study is a recent arrival to the United States, who had studied English for some years in his or her home country though doesn't perceive that much instructional time has been explicitly devoted to writing development, and who considers him or herself an intermediate speaker or English with slightly better speaking/listening skills than reading/writing skills. Table 2 presents a summary of these data, along with the T-statistic and p-value for independent sample T-tests run on the data. The groups did not significantly differ in age of arrival in the United States, years spent studying English, semesters of study in the United States, semesters of a language class focused on writing skills, or self-reported oral or writing ability in English.

Table 2: Participant language learning survey and between groups *T* - test

Group	Age of Arrival in United States Mean (SD)	Years of English Study Mean (SD)	Semesters of Study in USA/SL context Mean (SD)	Semesters of writing study Mean (SD)	L2 Oral Proficiency* Mean (SD)	L2 Literacy Proficiency* Mean (SD)	L3
Experimental	20.5 (5.4)	7.8 (3.4)	1.6 (1.1)	1.9 (1.3)	3.2 (.8)	3.2 (.7)	28 %
Control	19.7 (3.6)	7.1 (2.3)	1.4(.75)	2.4 (1.7)	3.1 (.7)	3 (.6)	26%
T statistic Comparing group means (*p* value)	.82 (.42)	1.3 (.2)	1.1 (.29)	-1.47 (.15)	.97 (.33)	1.25 (.22)	

* measured on a 5-point Likert Scale

The experience of the individual instructors for each section was a second possible source of between-group differences. Table 3 presents the median years of language teaching, semesters teaching at the college level, and semesters teaching writing (see Appendix B for data on individual instructors). Collectively, the instructors of the control sections have more years of experience in language teaching, while the instructors of the experimental sections have more semesters teaching at the college level and teaching college-level writing courses. One instructor (section 1) had 30 years of teaching experience, the majority of it at the college level, and was considered enough of outlier that medians are used to represent central tendencies. Taken as a whole, the instructors in both groups display a similar profile, with 3 instructors in each grouping having a moderate to high amount of teaching experience, and 1 instructor in each grouping (4 and 8) being a relatively new teacher, instructor 4 having received her Masters in TESOL three months before the start of data collection and instructor 8 in the second year of a 2-year MATESOL course.

Table 3. Teacher training and experience

Group	Master's Degree?	Self-Identified as Native-like proficiency	Median Years of Language Teaching	Median Semesters of College-level teaching	Median Semesters of Writing Instruction
Experimental	4/4	4/4	6.5	14.5	10
Control	3/4	4/4	10	7.5	5.5

Procedure

Pedagogical Treatment[1]

It is important to note that the efficacy of *particular* pedagogical methods or techniques for providing instruction in cohesion was not a focus of the present study's research questions; in other words, the study was not designed to compare two different treatments. I adopted a best-practices approach in the development of the treatment materials, integrating a variety of pedagogical activities that I believed would address the target structures, revising them after piloting them first with my own students, and then with a more formal pilot group. I developed an intervention sequence of five lessons that fit into 55-minute blocks and built successively until the final summary session.

The students were not given any homework, as I wanted to maintain an equality of instructional time between experimental and control groups to the extent possible. The powerpoint slides for each session were posted to a wiki after each session after several students asked me for copies. The data metrics for the wiki do not allow me to determine which of the students accessed the wiki, viewed the pages, or downloaded the files, but overall metrics suggest an early peak of interest (approximately 12 unique visitors after the first session) that quickly declined. By the last sessions, there were only occasional visits.

The experimental group participated in five treatment sessions, each lasting one hour, at one-week intervals for five weeks. Each section was scheduled for a 130-minute block, with

[1] In recognition of the fact that I, as the researcher, was the instructor for all pedagogical intervention sessions, the sections dealing with the interventions adopt the first-person voice, rather than the impersonal or passive, which would perhaps be misleading. I also refer to *students* rather than *participants* in this section, as many of the students attending the sessions were ultimately not included as participants in the study.

most instructors providing a 5-10 minute break in the middle of the class. With one exception (section 4) each intervention session was conducted during the second hour of the class, to minimize time on task lost to late arrivals, technology set-up, classroom management, and similar issues. I arrived at the beginning of the class period and sat in the back of the classroom during the first hour, then set up during the break and was prepared to begin immediately after the break.

Over the course of the intervention sequence I introduced three focused strategies, designed to build students' repertoire of writing skills while also increasing cohesion in their writing. The targeted strategies were: (1) defining technical words or key terms, introduced in session 1; (2) *this + summary noun* constructions, introduced in session 2; and (3) effective connector use, introduced in session 4. Sessions 3 and 5 served as consolidation sessions. In addition, two instructional themes addressing global writing concepts were used as a guiding structure throughout the intervention sessions. The first theme related to the structure of an argumentative essay and the way in which subsequent paragraphs added to and developed the idea presented in a thesis. The second theme was the communicative function of writing, in which students were encouraged to view the act of writing as engaging in a dialogue with a reader, in this case, the course instructor.

Overview of instructional activities.

While there was some variation between each intervention session, there was a common structure to each session. The sessions were built around a whole-class activity, the scaffolded writing of an argumentative essay on the prompt: *Do you feel that technology has had a beneficial or a harmful effect on communication between friends and family?* This pedagogical

task, and the specific prompt, was chosen because it was likely to be a familiar writing task and genre for students, based on their preparation for TOEFL examinations and other high-stakes writing assessments. During pilot testing of the intervention materials, I had looked at the possibility of introducing more academically relevant genres, such as a response paper. I found that given the limited instructional time, it was more effective to work within a genre that the students had familiarity with, and with a prompt that didn't require students to incorporate additional texts or sources. In addition, many of the students still had hopes of testing out of their remaining language requirements by retaking the TOEFL or the institution's in-house assessment. In this sense, the genre was considered highly relevant by the students themselves. In addition, the use of a genre that most, if not all, of the students had extensive experience with may have highlighted the effect of the strategies introduced. In other words, because the students already knew what a timed, argumentative essay looked like, they had a reference point by which to evaluate changes made by the introduction of defining language or *this + summary noun* constructions.

In the first session, the initial minutes were used as an introduction to the intervention series and a brief discussion of the instructional goals and objectives. In subsequent sessions, the beginning of the session was spent reviewing the concepts and writing covered in the previous week. This was followed by controlled, sentence-level practice with the target strategy for the day, and then scaffolded and practiced in an extended-discourse context during the group-writing activities. This group-writing activity was introduced in a limited form, using several prompts, in the first session. Beginning with the second session, each class worked on the group essay on the technology and communication prompt.

Instruction: Session 1. Session 1 used a discussion of essay macrostructure as an introductory activity, focusing on the use of general statements to introduce topics and ideas. The sentence-level strategy for the session was providing definitions of key terms or technical words in the text, though the classroom practice focused more on defining key terms as the group writing activity was not likely to include technical language. A particular focus was placed on the need to define lexical items that might appear to be unambiguous. The adjective *cold* was used as an example (i.e., what might it mean in August versus January, in describing coffee vs. milk). Several structures for defining terms were introduced, and identification and production exercises followed. The session ended with a whole class activity based on the writing prompt: *Do you agree or disagree with the following statement? "Parents are the best teachers."* Students volunteered ideas on the prompt, and I integrated them into an introductory general statement. (e.g., *We are born knowing very little about the world, and as we grow from children into adults, we need help learning about the world around us. There are many people in our lives who can act as our teachers.*-Section 4). Based on this statement, students identified key terms that might benefit from definitions. I provided four definitions for the word *teacher* (Figure 3) and we discussed which might effectively add to the argument suggested by our general statement.

- A teacher is someone who works in a school.
- A teacher is responsible for the education of less experienced people.
- A teacher gives new knowledge to young people.
- A teacher is an expert in a subject and explains it to other people.

Figure 3. *Four definitions of* teacher *used in Session 1*

Instruction: Session 2. Session 2 introduced the prompt for the scaffolded writing through a review of the General Statement and Definition strategies from Session 1 using the same

activity that closed Session 1. A hand vote indicated that students preferred to take the position that technology had a beneficial effect on communication (this was the case in all four experimental sections). Individually, students wrote general statements regarding the role of technology in communication, and these were combined by the class. The class then identified key words that might need to be defined for the reader (*communication* and *technology*). Individually, the students wrote definitions for the term *communication* which were then combined by the whole class into a one or two sentence definition (Figure 4).

Student Definitions:

- It's a way to connect between people.
- The way people interact with each other by expressing their feelings and thinking
- Gaining or receiving information
- Connecting with each other and transferring information no matter what way is used.

Class Definition:

- It's a way people connect by interacting, expressing feelings and ideas, and exchanging information. It does not matter what way is used.

Figure 4. Defining *communication* (section 4)

When defining the term *technology*, a different technique was used. Rather than provide an explanatory definition, the students were asked to provide specific examples of information technology. This was done through a whole-class discussion, while I entered the terms onto a powerpoint slide. Once an extensive list had been compiled, the class discussed which examples might be effective ones to include in the essay introduction. This exercise had a three-fold purpose: it provided content for the larger writing exercise, it demonstrated the flexible meaning of *definition* that would be used within the intervention sequence, and it modeled the reader expectation that examples of technology used in the introduction would serve as extended examples throughout the text. The general statement and two definitional elements were then

typed onto a Powerpoint slide and the students were asked to individually combine them into an integrated segment of text. I performed the same task simultaneously, and then circulated among the students to provide support and monitor progress. After the majority of the students had completed the task, I displayed my version of the combined ideas and the class discussed differences between the versions, and changes based on their input were made (see Figure 5 for an example of this activity).

General Statement: In the past two decades, communication technology has developed at a very high rate. It has started to make our world feel much closer. Communication: It's a way people connect by interacting, expressing feelings and ideas, and exchanging information. It does not matter what *way* is used. Technology • Skype • Facebook; • Email
Combined In the past two decades, communication technology such as email, Skype, and social-networking websites, has developed at a very high rate. *It* has started to make our world feel much closer by improving the way people communicate, that is, the way they connect by interacting, expressing feelings and ideas, and exchanging information.

Figure 5. Combining general statements and definitional elements

In each class, I combined the ideas in such a way as to begin a sentence with the pronoun *it* or *this*/*these* (see italicized *it* in Figure 5). This was used as a departure point to discuss the *This+summary noun* construction which was introduced and practiced for the remainder of Session 2. Activities included modified versions of those presented in Swales and Feak (2007) and the discussion of student writing examples taken from untimed writings included as part of the control corpus for the present study. Session 2 ended with a return to the combined writing exercise and the insertion of a *this+summary noun* construction (e.g., in Figure 5, *It* → *These advances in Internet-based systems*)

Instruction: Session 3. Session 3 began with a continuation of the work on *This+summary noun* constructions which closed the second session. First, I presented a cloze exercise I developed using examples from the Corpus of Contemporary American English (COCA) (Davies, 2008-2011), accompanied by a list of the twenty-four most frequent summary nouns entering into *This+summary noun* constructions as indicated by a search of COCA. This was followed by additional exercises adapted from Swales and Feak (2007). The purpose of both the COCA and Swales and Feak activities was to provide students with more lexical resources to deploy in *This+summary noun* constructions while also demonstrating the flexibility of and constraints on these constructions (i.e., understanding that a range of summary nouns might be appropriate for an individual referent, and that the same range might not be appropriate for every referent). Appendix C provides a complete list of the summary nouns provided to students through the intervention activities. This list represents a key subset of the lexical items used to in the frequency count of the *This+summary noun* structures appearing in the corpus and is also important in operationalizing measures of effect of instruction. Further discussion on this point follows in the description of analysis.

The second part of Session 3 returned to the group essay begun in Session 2. I presented the students with our combined general statement and definitional elements from Session 2, along with a thesis statement I had added in the intervening week (e.g., *These new systems allow us to make these connections stronger and more meaningful than ever before.*—Section 4). A second slide presented students with a topic sentence for the first body paragraph of the essay. Following some discussion of possible argumentation to support the topic sentence, the students were provided with the outline of a body paragraph in note form (Figure 6).

Topic Sentence: *One way that technology is strengthening communication is by making it easier for friends to maintain contact with each other.*

Support:
1. (Old way) someone needed to be responsible for starting communication
2. (example) Friends are too busy. No call or writing → no communication
3.(New way) Social networking changes old way
4. (elaboration)Friends follow each other like celebrities and get news about each other all the time.
5. (conclusion) This is a good change because following friends is better than following celebrities.

Figure 6. Example of scaffolded paragraph (section 4)

Pilot testing of the materials demonstrated that this was a very effective technique to provide structured opportunities for students to practice combining ideas using cohesive devices. A logistical advantage of this technique is that it provided room for students to deploy individual resources for creating cohesion, while resulting in paragraphs that were similar enough to be discussed in a whole-class setting. A second, related advantage was that it allowed writing practice to be carried out in a limited amount of time by removing the pressure of idea generation from individual students. At the same time, pilot testing indicated that a substantial amount of discussion of the idea chain and modeling of the procedure were necessary for students to make use of the paragraph outline, especially the first time this activity was used.

The students were given a sheet of paper with the introduction printed at the top, followed by blank lines (see Appendix D) and spent approximately ten minutes writing the paragraph. At the end of the session, the students' writing was collected with the promise that it would be returned with comments the following week. During the intervening week, I provided limited feedback on the students' writing, focusing only on areas which we had discussed in the intervention sessions. The majority of the feedback was indirect; an error was indicated by underlining or circling, often accompanied by a question or comment. The texts with feedback

were photocopied, and the originals were returned to the students. As with regard to overall instructional technique, the present study was not designed to investigate or support the use of a particular type of feedback. The choice to provide feedback was made within the context of constructing an instructional sequence based around sound and established pedagogical practices, and feedback was not operationalized separately from the overall effect of instruction.

Instruction: Session 4. Session 4 focused on effective use of connectors to create cohesion. Discussions with students during both the pilot testing and experimental sessions suggested that in some ways, the content of this session was the most familiar, and many students felt that the use of these structures was a source of interest and occasional confusion to them. The session opened with a very brief review of the syntactic and mechanical features of subordination (*Many students use the Internet for research <u>because it is more convenient</u>*), sentence connectors (*The Internet makes a variety of resources available to a student. <u>Furthermore, these resources are available almost instantly</u>*), and phrasal links (*<u>Unlike their parents</u>, students today are comfortable researching papers using the Internet*). Conversations with the section instructors indicated that these grammatical forms had been introduced in each section, and the sections were in the process of practicing and consolidating knowledge of these forms.

I explained to the students that each technique for connecting ideas was a good one, but that I was going to focus on sentence connectors as I had noticed my own students had difficulty with their use. I also emphasized that our focus was not going to be on grammar or mechanics, but rather on the relationships signaled by particular connectors. Basic categories of connectors (addition, cause/effect, contrast, examples, intensification, opposition, and ordering) were

introduced, and cloze exercises adapted from Swales and Feak (2007) were done as a group. This was followed by an acceptability judgment activity consisting of items based on common connector errors identified in Shea's (2009) study of connector use in L2 writing. These activities were designed to provide opportunities for explicit discussion of the type of relationships between propositions (e.g., There is a result or temporal connection between the sentences *It rained while I waited for the bus* and *I got very wet).* I emphasized the fact the fact that different types of connectors were not interchangeable solely because of membership in the same category (e.g., *and* could signal the relationship between the two sentences while *moreover* would not) and that the categories of connector were not mutually exclusive.

At the end of those activities, the students' scaffolded writing from Session 3 was returned with corrections, and examples of effectively and ineffectively used connectors from the students' own writing were reviewed and corrected using the powerpoint slides[2]. The structure of the essay to that point was reviewed, and we prepared to write the second body paragraph. I problematized the decision of what to write next and the class offered suggestions. Many students suggested beginning the second body paragraph with a new point and one of the ordering connectors (e.g., *second, secondly*). This was used to motivate a discussion of elaboration of ideas, and the idea of writing as a dialogue was highlighted.

I introduced the concept of thinking of writing as a conversation between a writer and a reader. I asked the students who they were writing for, generally, and who they imagined the

[2] The use of connectors was not a focus of the writing practice in Session 3, and feedback on connectors was not provided on that writing. The examples of connector use were treated as "found" examples that the students' had produced before the topic was introduced in the intevertion sequence.

reader of our technology and communication essay would be. These questions elicited the response that they generally wrote for their instructors. Using a powerpoint slide, we took a conversational version of our thesis statement (Figure 4; box INT), and imagined what question an instructor might ask us about it. We realized that the first body paragraph (Figure 7; box 1) could be considered a response to an instructor's questioning of our thesis statement (Figure 7; box A).

We then brainstormed what questions or comments an instructor might make in response to our first body paragraph. One possibility was an objection, namely, that the type of communication described (i.e., friends following each other through social networking tools) wasn't actual communication. This was used as the basis for a second body paragraph, which was written following the same scaffolding procedure used in Session 3. Session 4 closed with a discussion of the differences between an enumerated essay and an elaborated essay (although those terms were not used) and the fact that one essay type was not better than the other, but that having two macrostructures in one's repertoire allowed greater flexibility in timed writing, and both were likely necessary to produce a sufficient amount of discussion in longer, untimed assignments.

Figure 7. Powerpoint slide—Writing as a communicative act (section 4). Each text box appeared sequentially during instruction (For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.)

Instruction: Session 5. The final instructional session served as a review of the previous 4 sessions. The students' scaffolded writing from Session 4 was returned with feedback, which included more direct correction and meta-linguistic explanation due to the lexical nature of the target structure. The instructional version of the essay, now including three paragraphs and the blank writing lines, was distributed and discussed. The essay and argumentation to date was reviewed, coupled with powerpoint cloze activities using the essay and focusing on summary nouns and connector words. Identification exercises focusing on definitional elements were also included. The slides presenting the essay as a dialogue were reviewed and possible instructor comments on the 2nd body paragraph were brainstormed but not written due to the hypothetical time constraints of the simulated timed essay the activity was framed as. A model concluding

paragraph that I had written during the intervening week was provided, and the students were asked to find key content phrases in the concluding paragraph and trace them back through the whole essay. This activity was used as a basis for a review of the elaborated essay macrostructure. The key points of the intervention sequence (definitions, summary nouns, connector words, writing as a communicative act) were summarized, and the session ended.

Data Collection and Texts

The corpus of essays used in the present study was collected at three points during a fifteen-week semester, with each participant contributing three essays. Each essay was written within a thirty-minute time limit, and though not graded by the section instructors, was presented as exam practice and written under exam conditions (i.e., without assistance from dictionaries or other language resources and without input from instructors or classmates). The first writing (pretest) was completed during the 2nd week of the semester. The second writing (posttest) was completed during the 11th week of the semester, following the five-week intervention sequence (weeks 5-10). The third writing (delayed posttest) was administered three weeks after the posttest, during the 14th week of the semester. This resulted in a corpus of 279 texts and approximately 82,670 words.

Distribution of Prompts. The researcher provided the prompts and a schedule to the instructors to ensure that the prompts were balanced across time and group. The timed writing sessions were administered by the course instructors, who then provided the researcher with the handwritten texts, which were photocopied and then returned to the instructors. There were three prompts designed to elicit argumentative essays on topics not requiring extensive technical or

content knowledge. Two alternate prompts were also provided after two instructors suggested that the topics of the experimental prompts drew on content knowledge that had been extensively discussed in class activities unrelated to the research (See Appendix F for complete list). Table 4 shows the distribution of prompts across section and time, while Table 5 summarizes the total number of times each prompt was used at each data collection point and the total number of times each prompt was used in data collection. In both tables, the number in parentheses represents the number of participants writing on that prompt at that time.

It should be noted that, although it was counterbalanced in the initial design, the distribution of prompts was not equal across times. For example, in Table 5, it can be seen that prompt A was used in 3 sections at the pretest and delayed posttest, but only once at the posttest. The main reason for these discrepancies is that data collected from 2 sections was excluded, for reasons described above. A secondary factor is the use of the alternate prompts by two of the control sections. This would be a potential cause for concern if differences in rater judgments were associated with particular prompts; however, an ANOVA revealed no significant effect for prompt across the sample ($F = 2.13$, $p = .17$).

Table 4. Distribution of Prompts

Group		Pre	Post	Delayed
Experimental	1 (12)	A	B	C
	2 (9)	C	A	B
	3 (15)	B	C	A
	4 (11)	B	C	A
Control	5 (14)	E	B	A
	6 (13)	E	C	D
	7 (10)	A	C	B
	8 (9)	A	B	C

Table 5. Prompts used by time

Prompt	Pretest	Posttest	Delayed Postest
A	3 (31)	1 (9)	3 (40)
B	2 (26)	3 (35)	2 (19)
C	1 (9)	4 (49)	2 (21)
D	0	0	1 (13)
E	2 (27)	0	0

Preparation of texts. As the texts were collected, each handwritten text was typed by the researcher. Several participants had provided titles or chosen to rewrite the prompt before beginning the essay. These were not included in the typed version. The texts were typed exactly as written, with spelling and punctuation errors left unchanged. Paragraphs were marked by indenting and a line break. After entry, the electronic version was checked against the handwritten document to ensure that spelling errors were present in the original and had not occurred during data-entry. This version of the corpus, essentially identical to the original texts except for handwriting, was the version given to raters.

In order to use corpus analysis tools and other language analysis applications, a second version of the corpus was prepared. The first, and most extensive change, was that the texts had to be spellchecked, with misspelled words and non-standard English words corrected to a form

that would be recognizable by text-processing applications. This was done concurrently with the measurement of the lexical complexity and diversity of the text using the *Vocabprofile* tool on the *Compleat Lexical Tutor* website (Cobb, 2010; Heatly & Nation, 2004). The text file was pasted into the *Vocabprofile* text window, and an analysis was run.

For any word that was not recognized by the program, the following steps were taken. (1) Misspelled words were corrected. For the majority of these misspellings, the writer's intent was recoverable from context. More seriously misspelled words were submitted to the MS Word 2007 spellchecker, and the first suggested option was entered, unless it was deemed wholly inappropriate by the researcher, in which case the second option was used. The spellchecker method was used for 6 tokens out of more than 450 corrections. (2) Neologisms created using derivational morphemes were corrected to the standard form in the same part of speech (e.g., **stableness*→ *stability*). If there was no clear, single-word conversion (e.g., **lucked* → ?*had luck/was lucky*), then the word was changed to a base form (e.g., **lucked* → *luck*). It is important to note that, following the above criteria, misspellings that resulted in another English word (e.g. **He* ***dose*** *it* vs. *He* ***does*** *it*) were not corrected. This decision was made because it was often difficult to ascribe the error to either a mechanical or lexical basis.

The steps described above were made with some hesitancy by the researcher. It was recognized that some distortion of the data accompanied these textual manipulations, and in some sense, the researcher risked appropriating the writing of the participants. However, automatic calculation of lexical measures is severely affected by misspellings. The *Vocabprofile* program for example, compares tokens in a text to the "first thousand" and "second thousand" word families making up the General Service List (GSL) well as to word families making up the

Academic Word List (AWL) (see e.g., Nation & Waring, 1997 for a description). Words that are not recognized as part of these lists are classified as "off-list." Offlist words might represent technical or content-specific vocabulary or, alternatively, non-standard forms such as slang. For a less developed writer, whose lack of control over the language is manifested at least in part through repeated spelling errors or inconsistent spellings, the software will read that writer's text as containing a wider range of lexical types, including many off-list types, suggesting a greater use of content-specific language when in fact the text may contain only highly common, albeit misspelled, lexical items.

There is also a desire to maintain replicability in coding procedures, both for other researchers interested in expanding on this work, and for future additions to the corpus used in the present study. With these considerations in mind, the decision was made to refrain from correcting clear misspellings that nevertheless resulted in standard English lexical items. In specific cases, this decision does affect the measurements of texts. In the *dose/does* example above, taken from the present corpus, a basic function word is replaced by an off-list content noun. The decision not to correct such errors was a compromise, but one that is easy to replicate. The decision regarding which words are standard English forms, or standard in any written language, can be made using a dictionary. The decision as to whether a particular spelling is what the writer intended, though often obvious, nonetheless represents a judgment call.

Measurement of Writing Quality

Instrument. The quality of writing of each text was operationalized by rating on a five-category analytic scale (*content, organization, vocabulary, language,* and *mechanics*; see Appendix G). Each category was rated on a 20-point continuum. The continuum was broken into

four 5-point proficiency bands as an aid in assigning scores, but the actual point scores rather than band assignment were used for the analysis. When the scores for the 5 bands were totaled for an overall score, the *mechanics* score was divided by 2, reducing its effect on the total score and resulting in a range of 0-90 for the total score.

Norming and rating procedure. In addition to the researcher, two raters, with extensive experience teaching writing to the study population and extensive experience working with a variety of writing assessment instruments, were recruited and paid to participate in the study. Each rater rated each text in the corpus, meaning that each text was rated three times.

Rater training and norming was carried out using texts written during pilot data collection, on the same prompts used in the present study. The raters were told that the entire range of the scale was available to them, but that there was no requirement to use every band when assigning scores. The raters were not told that a single participant had provided multiple texts nor that the texts had been produced at different times during a semester. The norming session lasted until all three raters could consistently assign scores within the same band for each category.

After norming, the raters were given a packet of texts to rate and return to the researcher. There were six rating packets in total, with the initial packets containing fewer texts (30-40) and later packets containing more (60-70). For the initial packets, the scores were reviewed by the researcher to ensure that the raters were still normed. Numbers from the first packet indicated some discrepancy with regard to the *mechanics* subscore, with one rater rating one band lower than the other two raters for a majority of the 35 texts. By email, the researcher reminded the raters of the norming decisions regarding the mechanics subscore and asked them to review their

ratings and resubmit. No information was provided regarding which rater or which texts motivated the feedback. The resubmitted scores did not display the same discrepancy, and the second packet was distributed. The remainder of the packets did not demonstrate wide discrepancies, and the raters were informed that they appeared to still be normed.

The distribution of texts within the packets was pseudo-randomized so that every packet contained texts from each section, time, and prompt. Each packet contained the same texts, but in a different order. In other words, each rater received Packet 1 containing text A, text B, and text C, but texts A, B, and C appeared in a different order in each rater's version of Packet 1. The order of texts was determined using the *randomize* function in MS Excel.

IRB approval and participant consent was obtained to audiorecord discussions during the norming sessions. One point that was particularly salient in the audio recordings is that the raters felt quite clear on the descriptors for each category and were able to separate the features of each when reading a text. However, particularly for weaker texts, they often raised the question of how or whether to separate *content* and *organization* weaknesses when rating. This point will be discussed further in the discussion of results, as well as in the discussion of directions for future research.

Interrater reliability. When all rating was complete, the interrater reliability was calculated. Histograms, Q-Q plots, and Kolmogorov-Smirnov tests indicated that the *total* score and *content* subscore were normally distributed, the *organization, vocabulary,* and *language* subscores approached a normal distribution, and the *mechanics* subscore was not normally distributed. Table 6 presents the Pearsons's correlations for each of the three rater pairings, along with the percent agreement for each subscore. Percent agreement was calculated by taking the

absolute value of the difference between two raters' scores, multiplying it by the percent of the scale represented by one point, and subtracting from 1. Thus on each 20-point subscale, one point represented five percent of the total points available. Two raters who differed by two points would be considered to have 90 percent agreement (1 - 2 x .05). The percent agreement for all rater pairings was at 90% or above for all scores, while the interrater Pearson's correlations varied, but were above .8 for the total scores for two rater pairings and at .78 for the third.

Table 7 presents the Spearman Brown Prophecy values, which represent the reliability across all three raters combined (calculated according to Brown, 2005, p. 187), together with the mean percent agreement scores. Comparing Table 6, which focuses on the reliability between rater pairs, and Table 7, which takes into account the fact that three raters were used in the study, it is possible to see the benefit to the reliability of the rating instrument gained by using more than two raters.

Table 6. Pearson's correlation/percent agreement for interrater reliability

	Subscores					Total
Pairing	Content	Organization	Vocabulary	Language	Mechanics	
Rater 1 Rater 2	.69 /.92	.73/.92	.69/93	.64/.90	.74/.93	.83/.95
Rater 1 Rater 3	.72/.93	.72/.93	.6/.92	.58/.90	.71/.93	.78/.95
Rater 2 Rater 3	.71/.93	.74/.92	.65/.93	.57/.93	.74/.93	.81/.96

Note: Pearson's correlations are reported with percent agreement in parentheses

Table 7. Spearman Brown Prophecy/mean percent agreement for all 3 raters

Subscores					Total
Content	Organization	Vocabulary	Language	Mechanics	
.88/.93	.89/.92	.85/.93	.89/.91	.89/.93	.93/.95

Note: Spearman Brown prophecy statistic is reported with percent agreement in parentheses

Data Analysis

The data for the present study consisted of participant writing collected at three points during a semester of instruction. There are a number of ways that participants' language proficiency and repertoires of writing skills might have changed over the course of that semester. Some of these changes would be expected as the result of a semester of intensive English instruction, in addition to a semester of immersion in an English speaking environment. It would be expected that all participants, regardless of their membership in the experimental of control group, might demonstrate some development in their written language, as measured by standard measures of written language development (e.g., fluency and syntactic complexity). These changes in development might also manifest themselves in higher scores assigned by raters. A second group of changes might be directly attributable to the pedagogical interventions carried out with the experimental group (e.g., increased use of *This+summary noun* constructions or defining language). A third category of changes, increases in the frequency of cohesive devices

and the level of cohesion in texts, might have resulted in part from general language development as well as the specific strategies presented to the experimental group in the intervention sessions.

A number of measures were taken in order to present a clear picture of these various changes in participant writing. The details for each measure are presented in the following sections, followed by a summary that also discusses the statistical tests applied to these measures and the predicted outcomes of those analyses.

General Language Development

Measures of complexity and fluency are often used to provide measures of linguistic proficiency and development (see Ortega, 2003; Wolfe-Quintero at al, 1998 for a review). While there is some discussion regarding the particular measure used to represent each construct (e.g., Norris & Ortega, 2009; Shea, 2011), measures such as the raw frequency of a linguistic unit (e.g., words or T-Units) to assess fluency in timed production contexts, or the length of a particular production unit (e.g., words per T-units) and complexity of a production unit (e.g., clauses per T-unit, T-units per sentence) to assess syntactic complexity, have been used for a wide variety of research aims and contexts. There are analogous measures for assessing lexical development, focusing on lexical diversity (e.g., type/token ratio) and density (a ratio of content words to total words).

In order to provide a developmental context against which to assess changes in participants' levels of written cohesion and use of the strategies presented during the intervention settings, several of these developmental measures were used to measure the participants writing. Fluency was measured by the total number of words (W) and the total number of T-units (TU)

produced during a timed writing. These measures were chosen because number of words is the most straightforward measure, while number of T-units was more analogous to the super-sentential level of interest to the study. Syntactic complexity was measured by the number of words per T-unit (W/TU), and T-units per Sentence (TU/S), in order to reflect both the amount of content contained within individual syntactic units, and within the linguistic units analyzed by LSA software (sentences). Lexical development was measured by a length-adjusted Type/Token ration (Ty/Tok)[3].

Two reviews of the use of these measures in SLA research (Wolfe-Quintero et al., 1998 for a review of all constructs; see also Ortega, 2003 for a research synthesis focusing on syntactic complexity measures) have suggested that there are often not observable effects within a program level or even between adjacent levels. Ortega also found that longitudinal designs might require a year of instruction before effects are detected. Given that the data in the present study were collected over the course of a single semester and from participants within a single program level, it is possible that there would be no significant change in language development

At the same time, it is not unreasonable to expect that all participants in this study would exhibit some change in the broad areas of interlanguage development and second language proficiency represented by these measures. These changes would most likely be attributable to the semester of intensive English instruction the participants were engaged in. Ortega (2003) also noted larger effects for participants in a second language (SL) versus a foreign language

[3] Accuracy is the fourth construct commonly included in discussions of general developmental and proficiency measures. Measures of accuracy require significantly more time to calculate, and are less reliable between coders. Given these limitations, and in light of the fact that general linguistic development is not a focus of the present study, accuracy measures were not used.

instructional context. The effect of an SL environment might have been particularly strong given the fact that the semester of data collection represented the first semester of study in an SL context for a majority of participants in both groups: Experimental: n = 34 (72%); Control: n = 33 (72%).

The Effect of Interventions

There were a number of possible effects of the pedagogical interventions conducted with the experimental groups. An increase in raters' judgments of writing quality, an increase in measures of cohesion, or both, relative to gains made by the control group would serve as indirect evidence for the effectiveness of the pedagogical interventions. The fact that the interventions focused on several explicit, sentence-level rhetorical strategies also provided the opportunity to directly operationalized the effect of the intervention sequence by counting the occurrences of those structures.

There were three strategies that received focus during the intervention sessions: the use of defining language, the use of *This+summary noun* constructions, and the use of connector words and phrases. Using corpus tools, it was possible to measure the changes in the frequencies of these three cohesive devices. An increase in the frequency of some or all of these devices, both within the experimental group from pretest to posttest and relative to gains made by the control group, would provide evidence of an effect for the intervention sequence.

Determiners+summary noun constructions. The cohesive device that required the least amount of interpretation in the search was the *This+summary noun* construction. Using AntConc concordancing software, searches were performed for all occurrences of *this* and *these.* Additional searches were also performed for all occurrences of *that* and *those*. The latter two

determiners are not considered standard forms of the target structure (e.g., Swales & Feak, 2007), but in consideration of the fact that the participants in the study may have had varying degrees of control over the structure, all four determiner forms were included for completeness.

The searches yielded a list of the targets in KWIC (key words in context) format (see Figure 8). For each text, a total number of hits was recorded. No distinction was made between singular or plural forms, but occurrences of *that* and *those* were recorded separately. In subsequent discussion, reference to *this* constructions will include all four forms, unless stated otherwise.

Once the total number of occurrences of *this* were counted, they were categorized according to the following taxonomy, with examples taken from the output shown in Figure 8. Lines 2, 4, 5, 6, and 12 are examples of *pronominal this* (*ProThis*) in which *this* acts as a pronoun. Of the examples in Figure 8, lines 1, 3, 7, 8, 9, 11, and 13 were counted as *Det+summary noun* constructions. Lines 1, 3, 7, 8, 9, 10, 11, 13, 14, and 15 are examples of *determiner this* (*DetThis*), in which *this* acts as a determiner for a noun phrase. DET*this* occurrences were further categorized as *summary noun* constructions or *concrete noun* constructions, in an adaptation of Gray and Cortes' (2011) taxonomy. Of the DET*this* constructions in Figure 8, lines 10, 14, and 15 would be categorized as *concrete noun* constructions, in that the head noun *world* can be identified as a specific semantic concept without making reference to the surrounding text.

Hit	KWIC	File
1	work on Saturday and Sunday. This kind of hard work bings lots	dl02.txt
2	work equal to a lot of money. This is why people work hard becau	dl02.txt
3	ty. Some people disagree with this statement because they are de	dl03.txt
4	ositive influence on society. This is largely because he used a	dl03.txt
5	e of Bill Gates. In addition, this shows that not all rich peopl	dl03.txt
6	e negative impact on society. This is mainly attributed to the :	dl03.txt
7	in economy aspects. Second, this issue is the problem about me	dl04.txt
8	ch money donates for society, this situation could be a good mo	dl04.txt
9	es a society more healthy. In this point, earning money even mu	dl04.txt
10	lity. It is fair to anyone in this world. You have the ability,	dl06.txt
11	y can make the right choices. This ability doesn't belong to al	dl06.txt
12	g to all the people, and also this is the difference between a r	dl06.txt
13	illegally. I have to say that this phenomenon really exists, but	dl06.txt
14	e and what they need to do to this world. All in all, rich peop	dl06.txt
15	ey are making contribution to this world. They deserve to own su	dl06.txt

Figure 8: First 15 lines of search result for *this*. The right-hand column indicates which text contains the token. Four texts (102, 103, 104, 106) from the delayed-posttest are represented.

Gray and Cortes (2010) made a further distinction between examples such as 1, 9, and 13, referring to them as *other*, *adverbial head,* and *shell* constructions respectively. In their taxonomy, only *shell* constructions would be considered analogous to the *This+summary noun* constructions in the present study. However, Cortes and Gray were examining fine-grained distinctions in polished, "expert" texts published in academic journals. The present study focuses on the effect of an intervention within the timed writing of L2 learners, and the technical distinctions made by Gray and Cortes were not part of the interventions. Given the different goals, a decision was made to adopt a more inclusive coding system when counting DET*this* constructions.

If the intervention strategy encouraging the use of *This+summary noun* constructions had an effect, an increase in the number of these constructions within the experimental group from pretest to posttest would be expected, as well as a larger gain in the rate of these constructions

relative to the control group. This increase might manifest itself in a number of ways, and the following measures were taken in order to investigate these potential changes. First, the ratio of *This+summary noun* constructions to the total occurrences of *this* was calculated (SN/This), in order to determine whether participants were more likely to choose the more elaborate structure in contexts which *this* would also be acceptable. Secondly, the ratio of *this+summary noun* constructions to total T-units per text was calculated (SN/TU), to determine whether participants were making more use of the construction to link ideas across cohesive units. These two measures were also calculated using all instances of *DETthis,* or *summary nouns* plus *concrete nouns* (DTh/This; DTh/TU) in order to account for the possibility that some participants may have overgeneralized the strategy to use with any lexical noun.

A third analysis was carried out at the level of the experimental and control corpora. The percentage of DET*this* constructions incorporating one of the summary nouns presented during the pedagogical interventions (Appendix C) was calculated.

Connectors. In order to address the use of connectors in the corpus, a search was conducted using the *AntConc* software. The list of search terms was taken from previous work on connectors by the researcher (Shea, 2009, see Appendix H). The number of connectors per T-unit (Con/T) was calculated per each essay. In addition, the particular connectors, as well as category, were recorded. The overall use of connectors across texts was not predicted to change significantly. However, it was predicted that participants in the experimental group would use a wider range of connectors, from more categories. As with the *This+summary noun* constructions, a comparison of all connectors and those connectors which received attention during the intervention sessions (Figure 8) was conducted.

Therefore	In fact	As a result
On the other hand	However	Consequently
In other words	Nevertheless	In contrast
That is	Otherwise	Actually
For example	Furthermore	Conversely
For instance	In addition	
On the contrary	Moreover	
As a matter of fact	Likewise	

Figure 8. Connectors included in intervention sequence

Definitional elements. The pedagogical focus that is perhaps least amenable to corpus analysis is defining language. Definitional elements can take a wide variety of forms. They can be appositive NPs, embedded relative clauses, or independent sentences that are marked by a connector phrase or unmarked. Thus, identifying a segment of text as a definitional element is a functional, rather than a formal, categorization. The identification of definitional elements was accomplished though an iterative categorization process. During various stages of data processing, including typing the handwritten documents, spellchecking the documents, and the counting of T-units, the researcher noted definitional elements in the texts. This coding was not done during the rating of texts, to avoid possible influence on the researcher's contributions to the ratings. Thus, each text was reviewed three separate times during the data processing procedures. The full corpus was then reviewed a fourth time solely to review and identify any additional definitional elements. A full discussion of the taxonomy and features identified is presented in the results and discussion.

Because many texts contained no definitional elements, and many others contained only one or two of these features, texts were grouped into those containing no definitional elements, 1 -2 definitional elements, and 3 or more definitional elements. The raw frequencies were retained

to aid in interpreting the results, as well as the gain scores exhibited by participants were used in analyzing the results.

Global effects of instruction. It is important to note that in addition to the three explicitly taught strategies, the intervention sessions were organized around two themes focusing on essay macrostructure and the communicative function of writing. These themes were included in order to provide context for the three sentence-level strategies, and also because an awareness of global coherence is in some ways a necessary part of a writer's understanding of cohesion. However, changes resulting from participants' attention to these themes would not necessarily be marked by explicit changes to textual features. If the experimental group demonstrated an increase in measures of cohesion or in raters' scores from the pretest to posttest relative to the control group, such an increase could be attributed to these less explicit features of the interventions. Similarly, if raters' scores of the experimental groups writing increased relative to the control group, but without an accompanying relative increase in general measures of language development, that would constitute indirect evidence of the effect of the pedagogical intervention.

Measuring cohesion.

As described in the review of the literature, the construct of cohesion is very likely a multidimensional one, representing the interactions of several features of texts. Both the theoretical and research literature suggest that the creation of *complex*, *interacting* lexical chains is a central factor in the creation of cohesive texture. However, focusing on the amount of lexical cohesion will not account for the fact that highly, or overly, cohesive texts are often perceived as less effective by readers. It is likely that the amount of lexical cohesion interacts with the lexical

diversity and density of a text in the creation of effective local coherence. A second factor contributing to cohesion is the use of connectors, but again, research suggests that the quality as well as the quantity of connector use must be considered.

Lexical development measures. After the spell-checking and other data cleaning procedures were completed, the text was resubmitted to the *Vocabprofile* program. The results of the analysis were used to create a context against which the lexical cohesion of a text could be evaluated. The following lexical measures were recorded. (1) Tokens (total # of words) and (2) Types were used to calculate (3) a length-adjusted type/token ratio. Texts with a lower type/token ratio were likely incorporating more simple repetition.

Latent Semantic Analysis. The review of the literature provided a discussion of research findings on cohesion and coherence using LSA-based methods. A more detailed discussion of the technical aspects of LSA is presented here.

The first step in an LSA analysis is the creation of a semantic space for the analysis. The following example of this process is a paraphrase Martin and Berry (2007, citing Witter & Berry, 1998). A corpus of *documents* matching the particular semantic domain of interest is collected. In the creation of a vector space model, the term *document* can refer to a unit of text, whether it be a sentence, paragraph, or entire text. In this case, the *documents* are the keywords in titles for topics on music and baking. Table 8 displays a list of these titles, with the keywords, which will be the only words included in this example corpus, italicized.

Once the corpus is collected, the types and documents are used to create a type-document matrix, in which each row represents a type (word) appearing in the training corpus and each

column represents a document included in the corpus. Each cell in the matrix is marked with the frequency that each type appears in each document (Table 9).

Table 8. LSA example: music and baking titles

Document Label	Title
M1	*Rock* and *Roll Music* in the 1960's
M2	Different *Drum Rolls*, a *Demonstration* of Technique
M3	A Perspective of *Rock Music* in the 90's
M4	A Perspective of *Rock Music* in the 90's
M5	*Music* and *Composition* of Popular Bands
B1	How to Make *Bread* and *Rolls*, a *Demonstration*
B2	*Ingredients* for Crescent *Rolls*
B3	A *Recipe* for Sourdough *Bread*
B4	A Quick *Recipe* for *Bread* Using Organic *Ingredients*

The type-document matrix is generally a sparse matrix (i.e., most cells have a value of zero). This is due to the fact that the majority of words will not occur in the majority of texts, although the example above has a non-zero value for roughly 25% of its cells

A weighting transformation is commonly done on the matrix to weight the types based on how well they differentiate between the documents. *Global weighting* functions represent how frequent the type is throughout the corpus; a very frequent type will likely appear in a large number of texts and thus not differentiate between texts well. Local weighting functions represent how frequent a type is within a particular document; a type that appears frequently within one document is more likely to be related to that
document's meaning, and a type that appears frequently in one document but not in others is likely to be useful in differentiating between the semantic content of different documents. These two weighting functions, global and local, are then combined to weight each cell in the matrix. A commonly used weighting function, and one employed by the LSA applications used in the present study, is *log-entropy weighting*, which decreases the effect of large differences in local

frequencies while also decreasing the influence of types common across the corpus. Table 10 presents the weighted version of the matrix in Table 9.

Table 9. Type-document matrix with frequencies corresponding to Table 8

Types	Documents								
	M1	M2	M3	M4	M5	B1	B2	B3	B4
Bread	0	0	0	0	0	1	0	1	0
Composition	0	0	1	0	1	0	0	0	0
Demonstration	0	1	0	0	0	1	0	0	0
Dough	0	0	0	0	0	0	0	1	1
Drum	0	1	1	0	0	0	0	0	0
Ingredients	0	0	0	0	0	0	1	0	1
Music	1	0	0	1	1	0	0	0	0
Recipe	0	0	0	0	0	0	0	1	1
Rock	1	0	0	1	0	0	0	0	0
Roll	1	1	0	0	0	1	1	0	0

Table 10. Type-document matrix with frequencies corresponding to Table 9

Types	Documents								
	M1	M2	M3	M4	M5	B1	B2	B3	B4
Bread	0	0	0	0	0	.474	0	.474	0
Composition	0	0	.474	0	.474	0	0	0	0
Demonstration	0	.474	0	0	0	.474	0	0	0
Dough	0	0	0	0	0	0	0	.474	.474
Drum	0	.474	.474	0	0	0	0	0	0
Ingredients	0	0	0	0	0	0	.474	0	.474
Music	.347	0	0	.347	.347	0	0	0	0
Recipe	0	0	0	0	0	0	0	.474	.474
Rock	.474	0	0	.474	0	0	0	0	0
Roll	.256	.256	0	0	0	.256	.256	0	0

In the example above, no type appeared in a document more than once, so the local weighting is the same for each cell. The more documents a type appears in, the less unique it is, and the more its value is reduced by the global weighting function, resulting in *roll* which appears in 4 different documents, receiving the lowest value in the matrix.

Using these weighted values, the matrix is then decomposed using a statistical procedure known as Singular Value Decomposition (SVD), which is a form of factor analysis. Essentially, this assigns values to a particular word for a large (100-500) number of factors relating to which semantic contexts it is likely or unlikely to appear in. It is intuitively useful to imagine these factors as representing semantic concepts; thus, *ingredients* might be thought of as loading heavily onto factors related to food and cooking, but not loading heavily onto factors representing music and music theory. However, it is important to bear in mind that these factors are mathematical abstractions, and would not correspond to semantic categories in any recognizable way.

A second important point is that these procedures do not describe the analysis done on the target texts (i.e., the data for the present study). These are the steps taken to build a particular semantic space, which is then used to evaluate the semantic information of target texts. The particular semantic space chosen is an important feature, as some spaces may not appropriately account for the semantic content of the target text. For example, while the sample space above would be effective at discriminating between music and baking texts, it might misclassify geology texts as similar to music texts based on the word *rock*.

On the LSA website, a variety of semantic spaces are available as options. All analyses included in the present study were carried out within the *College Level General Reading* semantic space, consisting of 37,560 documents and 92,409 unique lexical types drawn from a cumulative progression of reading levels from 3rd grade to college level and a range of subject areas (see Dennis, 2007, pp.69-70 for a complete description).

LSA applications. After the lexical measures were obtained using the *Vocabprofile* tool, the texts were then analyzed using two LSA applications available on the LSA website maintained by the University of Colorado, Boulder (http://lsa.colorado.edu/). A summary of those applications is presented here. A complete review and explanation of the tools provided by this website is available from Dennis (2007).

The first application used is the *Sentence Comparison* tool. This tool calculates the cosine between the LSA vectors of adjacent sentences, with higher cosines representing more semantically related sentences. Foltz, Kintsch, and Landauer (1998) have reported that the mean of cosines between adjacent sentences in a text can provide an approximate representation of the coherence of that text. Using this application, each text in the corpus was given a mean cosine measure. The standard deviations for these means were also recorded.

The second application, *Matrix Comparison*, was applied to the paragraphs of a text, and provided a matrix representation of the semantic relatedness between each paragraph and every other paragraph in the text. Again, the mean of these cosines was recorded for each text, with higher means representing texts whose paragraphs were more semantically related to each other. Using the same matrix, the average cosines between the first paragraph and each of the other paragraphs in the text was recorded. This measure was taken to identify texts in which individual body paragraphs had a low level of relation, but each related back to the introductory paragraph. Such a relationship was thought to be more characteristic of enumerated, rather than elaborated, texts. However, the two measures were almost identical and so only the overall paragraph to paragraph score is reported.

Connectors. LSA measures were used to represent the presence of lexical reference chains throughout a text. The second component of cohesion, the use of connectors to signal relations between propositions, was measured using corpus analysis tools. Unlike, lexical cohesion, the use of connectors was a direct target of instruction. It was measured in the same way as when ascertaining the effectiveness of the pedagogical interventions. A list of connectors, compiled by Shea (2009) was used as a search list (Appendix H). Because connectors are used to connect syntactic units, the raw frequency of connectors in each text was divided by the number of T-units to create a connector per T-unit ratio (CON/T). Connectors were also classified according to type (*additive, appositive, causative, contrastive, enumerative, summative, transition).*

Analysis

Rater Scores. In preparation for the analyses which directly addressed the research questions, several preliminary analyses were performed. The first was conducted on the group means of the scores assigned by the three raters on the 90-point assessment instrument. A repeated measures factorial ANOVA was run on the average total rater score, treating group (control and experimental) as a between subjects variable and time (pretest, posttest and delay) as a within-subjects variable. The assumption of sphericity was violated, so the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .998$) as recommended by Field (2006).

Planned contrasts were included in the factorial ANOVA in order to identify the points of difference between the groups. Following the recommendation of Fields (2006, pp. 460-463; 473-478; 489) planned *repeated measures* contrasts were selected in SPSS, which compared the

main effect for Time and the interaction effect for Time and Group from pretest to posttest and from posttest to delayed posttest. A second analysis was run using with *simple* contrast selected to obtain a contrast for pretest and delayed posttest.

For the purposes of the study, the interaction between Group and Time were the important contrasts. The analysis indicated which contrasts represented significant differences in the performance of the two groups one stage of data collection to the next, but the direction of those differences (i.e., whether the experimental group performed better than the control group) was not indicated. The results were thus interpreted in conjunction with the graphic and numeric representations of the data.

The same statistical procedure was used to analyze measures of fluency (total words, total T-units)

Analyses for Research Questions

Three statistical analyses that were conducted in order to address the research questions are presented below, organized by the particular research question they address.

RQ1.: Can cohesion be represented as a single factor, or should it be treated as a multidimensional construct? Preliminary analysis on a smaller test corpus suggested that when included in a principal component analysis (PCA) with measures that tap well-established writing constructs such as fluency and complexity, measures of LSA and connector use load onto a single factor with an eigenvalue above 1, which can be conceptualized as *cohesion*. This analysis was replicated in the present study, including the measures of lexical diversity with measures of cohesion (LSA measures and connector use) in a direct oblimin rotated solution. If the result is replicated, and the lexical and conjunctive cohesion measures load onto a single

factor, that factor score will then be used to operationalize cohesion in subsequent analyses. The unit of analysis is individual texts.

The principal component analysis was carried out using a direct oblimin rotation, suitable as the factors were unlikely to be completely independent (see Field, 2006). Following Field, an eigenvalue of greater than 1 was chosen as a conservative measure of an independent factor, Regarding sample size, Field (2006) discusses two suggested guidelines for determining adequate sample sizes: 10-15 participants per variable or an overall sample size of 300 (pp 638-641). In this analysis, the unit of analysis was the text, of which there were 279. This is more than 15 times the number of variables included in the final analysis (8), and close to the 300-participant mark suggested by Field. The final factor model consisted of 8 variables: 3 lexical measures (type-token ratio, measure of textual, lexical diversity (MTLD) and voc_d), three LSA measures (sentence-level, paragraph-level, and the standard deviation of sentence to sentence) and 2 measures of connector use (connectors per 100 T-units, number of connector categories). The overall model and each variable reached the minimal level of sampling adequacy (KMO > .5).

RQ2.. What are the relationships between the level of cohesion within a text (lexical and conjunctive) and measures of writing quality? It was predicted that cohesion, conceptualized as a construct consisting of lexical cohesion and connector use, would interact with the lexical development within a text. For those texts demonstrating a higher level of lexical variety, the level of cohesion will correlate with raters' scores. For texts with a lower level of lexical variety, the level of cohesion will not correlate, or will correlate negatively with raters' scores.

The factors identified in the PCA carried out for RQ 1 will be entered into a Spearman's ρ non-parametric correlation analysis. It was predicted that the cohesive factors would correlate positively with mean total scores.

RQ3: Can learner use of cohesive devices be modified through instruction, and is there a corresponding change in perceived quality? Unlike the analyses conducted under RQ1 and RQ2, the unit of analysis is the participant. The effect of instruction is operationalized as the frequency of summary nouns, connector use, and definitional elements as well as the variety of use of these structures.

The use of inferential statistics to address this research question was potentially problematic, due to the nature of the data. The structures studied and the writing tasks were not such that it was necessary to produce the target structures to successfully complete the task. While nearly all participants produced some adverbial connectors, for example, many texts did not include any determiner+ summary noun constructions or definitional elements. This led to data which difficult to interpret using measures of central tendency, a foundation of inferential statistical analysis.

Non-parametric statistics were more appropriate to use with this data; to determine the effect of instruction, Friedman's ANOVAs were used to determine within group differences across time, with Wilcoxon signed-ranks tests used as post-hoc tests to identify specific points of difference when appropriate. To investigate the relationship between rater scores, cohesion measures, and treatment targets, a Spearman's ρ correlation was conducted.

Summary.. Table 11 provides a summary of the various measures used in the present study, giving information on the type of measure (e.g., frequency count, ratio), the purpose of the

measure (i.e., how it contributes to an investigation of the research questions), and the predicted results of the measure, both from prettest to posttest and between control and experimental group. Some preliminary analyses (e.g., t-tests to establish initial equality between the control groups) are not included.

Table 11. Summary of measures in present study

Measure	Type	Purpose & Predicted Findings
Measures of Writing Quality		1. Assess potential effect of treatment on writing quality
Rater Scores		2. Investigate relation between cohesion and coherence
5 Subscores	Mean (3 raters)	Predictions:
Total Score	Mean (3 raters)	Higher posttest scores for EG relative to CG and pretest scores
Measures of Development		1. Provide context for increase in cohesion within general language development
Fluency		2. Demonstrate equality of EG and CG in terms of general language development
Words	Frequency	
T-units	Frequency	
Complexity		3. (*Lexical measures only*) Provide context for differential effect of high level of lexical cohesion
Words/T-unit	Ratio	
T-units/Sentence	Ratio	Predictions:
		1. Potential main effect for Time; no main effect for Group.
Lexical Development		2. Texts with High Lexical Development and Lexical Cohesion rated more highly than High Lexical Development and Low Lexical Cohesion. Texts with Low Lexical Development and Lexical Cohesion possibly rated more highly than texts with Low Lexical Development and High Lexical Cohesion
Type/Token	Ratio	
Measures of Treatment Effect		1. Direct evidence of the effect of intervention sequence
Summary Nouns		2. Establish relation between treatment targets and writing quality (with WQ measures)
Determiner *this* constructions	Frequency	
Summary Noun tokens and types	Frequency	3. Establish relation between treatment targets and cohesive elements (with Measures of Cohesion measures)
Determiner + summary noun		Predictions:
Determiner + concrete noun	Frequency	1. EG demonstrates higher rate of SN use and frequency of DEF
Change in target summary nouns produced	Frequency Gain score	2. EG demonstrates more varied CON 3. Correlation between SN, CON, and DEF and WQ 4. Correlation between SN, CON, and DEF and LSA
Table 11 Continued	Ratio	

Table 11 Continued

Connector Use		
Connectors/T-unit	Ratio	
Connector Categories	Frequency	
Enumerating connectors/all connectors	Ratio	
Text by number of connector categories	Distribution	
Definitional Elements		
Number	Frequency	
Texts by number of summary noun types	Distribution	
Measures of Cohesion		1. Demonstrate relation between lexical cohesion and WQ
Lexical cohesion (LSA)		2. Demonstrate relation between variety of CON and WQ
Sentence-to-sentence	Mean (all adjacent pairs)	Predictions:
Paragraph-Paragraph	Mean (all combinations)	1. Correlations between LSA measures and WQ
		2. Correlations between CON and WQ
Connector Use		
Connectors/T-unit	Ratio	
Connector Types	Frequency	

Note. EG= Experimental Group; CG = Control Group; WQ = Writing Quality; SN = Summary Nouns; CON = Connector Use; DEF = Definitional Elements

CHAPTER 3: RESULTS

The organization of the results section is as follows. For all reported analyses, both between and within-group differences are discussed. The rater scores are first reported in order to determine if there was indeed any change in participant writing quality during the course of the data collection. This is followed by a report of fluency and syntactic complexity measures, which are provided before the results of the main analyses for context in interpreting the results. The results pertaining to each of the three research questions are then discussed in order.

The first research question asked whether cohesion could be thought of as a unified construct, or whether its different components, namely lexical cohesion and connector use, need to be considered separately. Before reporting the main analysis for RQ1, the analyses of LSA measures are reported. These initial analyses are followed by the results of the PCA. This is followed by the results of the analyses relevant to RQ 2, which asked if cohesion measures could be related to measures of writing quality. The third research question asked if it was possible to affect the level of cohesion in participant writing through a pedagogical intervention. These results are presented and interpreted in light of the results from RQ2.

Rating

Table 12 presents the mean scores of writing quality for each group, which were calculated for each text by taking the mean of the three raters' total scores on the 90-point analytic scale. These means are represented graphically in Figure 10.

Table 12. Mean total rater scores

Time	Mean	*SE*	95% Confidence Interval	
			Lower Bound	Upper Bound
		Control		
pretest	53.03	1.03	51	55.08
posttest	50.89	.98	48.94	52.83
delayed	55.55	1.04	53.48	57.62
		Experimental		
pretest	50.33	1.02	48.3	52.35
posttest	56.48	.97	54.56	58.4
delayed	56.46	1.03	54.41	58.51

The analysis indicated a significant main effect for time, F (2, 181.65) = 15.39, $p < .001$. Contrasts indicated that at each time, the mean total score rose significantly (Table 13).

Table 13. Planned contrasts examining main effect for Time (rater scores)

Time	Mean difference	*F*	*df*	*p*	*r*
pre-post	2	6.31	1, 91	.014	.25
post-delay	2.32	9.09	1, 91	.003	.3
pre-delay	4.33	31.32	1, 91	<.001	.51

These results indicate that, as a whole, the quality of the participants' writing went up over the course of data collection. Given that all participants were enrolled in intensive English program and that data collection spanned a semester, this result was expected.

There was no main effect for group, F (1, 91) = 1.29, p = .26, r = .12. This indicated that, when time was not taken into account, there were no differences between the control and experimental groups.

There was a significant interaction between group and time, F(2, 181.65) = 14.19, $p <$ .001. Table 14 presents the results of the planned contrasts investigating these interactions.

Table 14. Planned contrasts examining interaction of Time*Group (rater scores)

Time	*F*	*df*	*p*	*r*
pre-post	27.1	1, 91	<.001	.48
post-delay	9.21	1, 91	.003	.3
pre-delay	5.47	1, 91	.022	.24

Looking at Figure 10 to interpret these results, the most highly significant and largest effect occurred between pretest and posttest, during which time the experimental group mean increased by approximately 6 points, while the control group decreased 3 points. The second significant effect occurred between the posttest and delayed posttest, during which the control group increased by just less than 5 points while the experimental group remained largely unchanged. From pretest to delayed posttest, there was a smaller, significant difference which the graph suggests is due to the experimental groups' larger overall gain of 6 points compared to the control group's 2.5

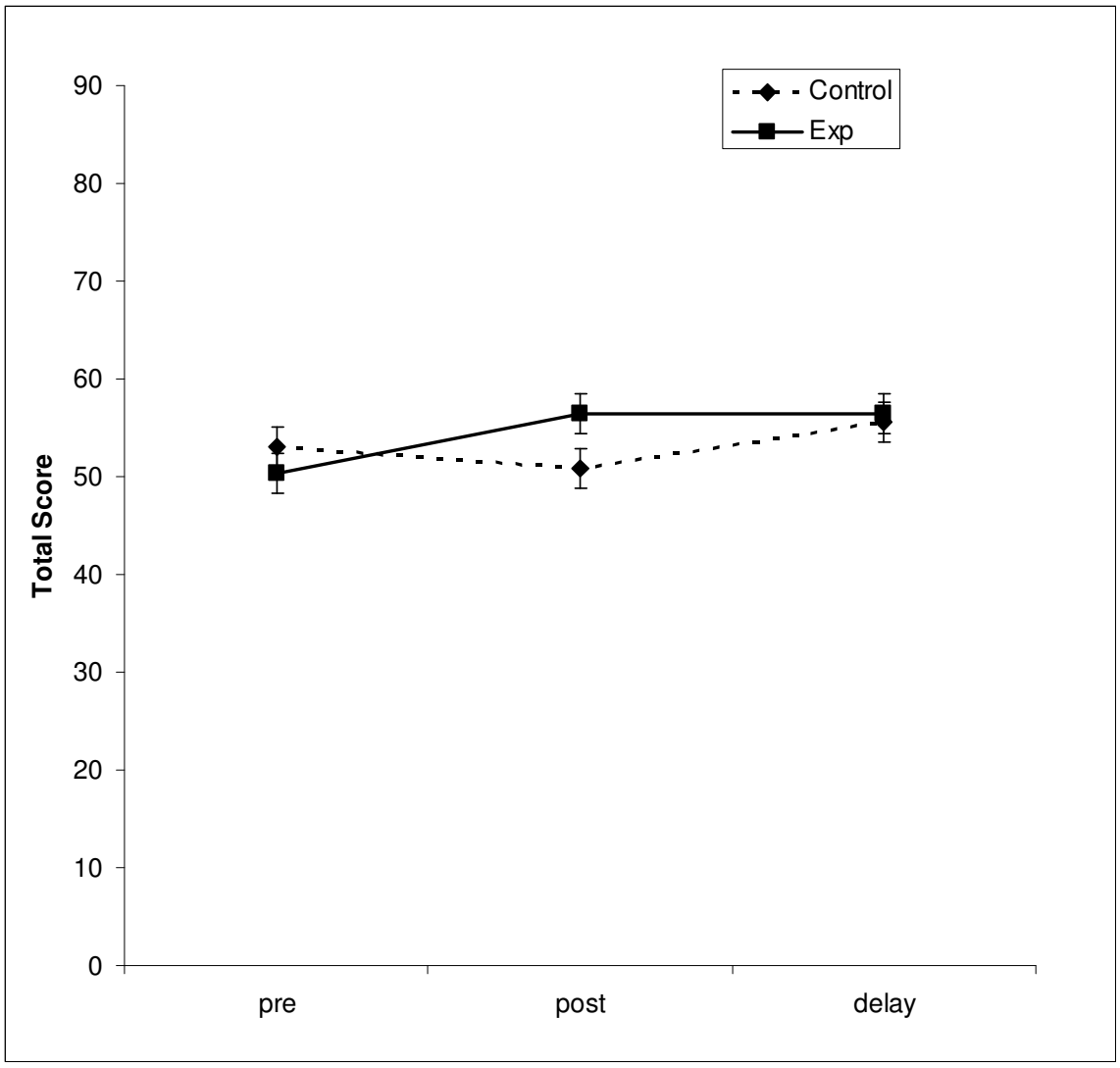

Figure 9. Mean Rater Scores

The two groups were not significantly different in their scores at pretest or delayed posttest. The experimental group performed significantly better than the control group at the posttest. What needs to be investigated then, are possible explanations for the early increase in the quality of the experimental group's writing. Of particular interest is whether these differences could be associated with the treatments administered as part of this study.

Development

Before arguing that the posttest difference in total mean scores was the result of the intervention sequence, it was necessary to look at the within and between group measures considered to represent core language development. Table 15 presents the descriptive data for all developmental measures by group and time.

Fluency. To measure the development of fluency both the number of words produced (Figure 11) and the number of T-units produced (Figure 12) were calculated and analyzed using a repeated measures factorial ANOVA (see Table 15 for descriptive statistics). As suggested by the figures, there was little difference between the two groups.

In both analyses, a main effect was found for time, (F_{words}(1.85, 168.7) = 20.74, p <.001; $F_{T\text{-}unit}$ (2, 182) = 13.89, p < .001) but no main effect was found for group, F_{words} (1, 91) = .07, p = .79, r = .02; $F_{T\text{-}unit}$ (1, 91) = .4, p = .53, r = .06). The interaction effect between time and group was also found to be non-significant (F_{words}(1.85, 168.68) = .72, p <.48; $F_{T\text{-}unit}$ (2, 182) = .88, p < .42).

Table 15. Descriptive data for fluency, complexity, and lexical developmental measures

	Words			T-unit			Word per T-unit			Type-Token Ratio		
Time	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
						Control						
Pretest	271.83	83.54	383	21.48	6.48	28	12.8	1.98	8.54	5.51	.74	2.82
Posttest	301.59	78.53	354	25.26	8.35	41	13.05	3.51	18.63	5.24	.6	2.51
Delayed	317.33	81.12	396	25.43	8.94	44	12.54	2.51	10.32	5.34	.67	3.03
						Experimental						
Pretest	258.26	86.41	372	24.6	9.62	44	13.14	3.66	20.36	5.36	.65	2.97
Posttest	307.77	86.87	449	20.06	6.66	32	13.15	2.33	9.23	5.55	.68	2.84
Delayed	313.28	76.53	366	24.02	8.39	45	13.87	3.65	19.04	5.51	.66	2.87

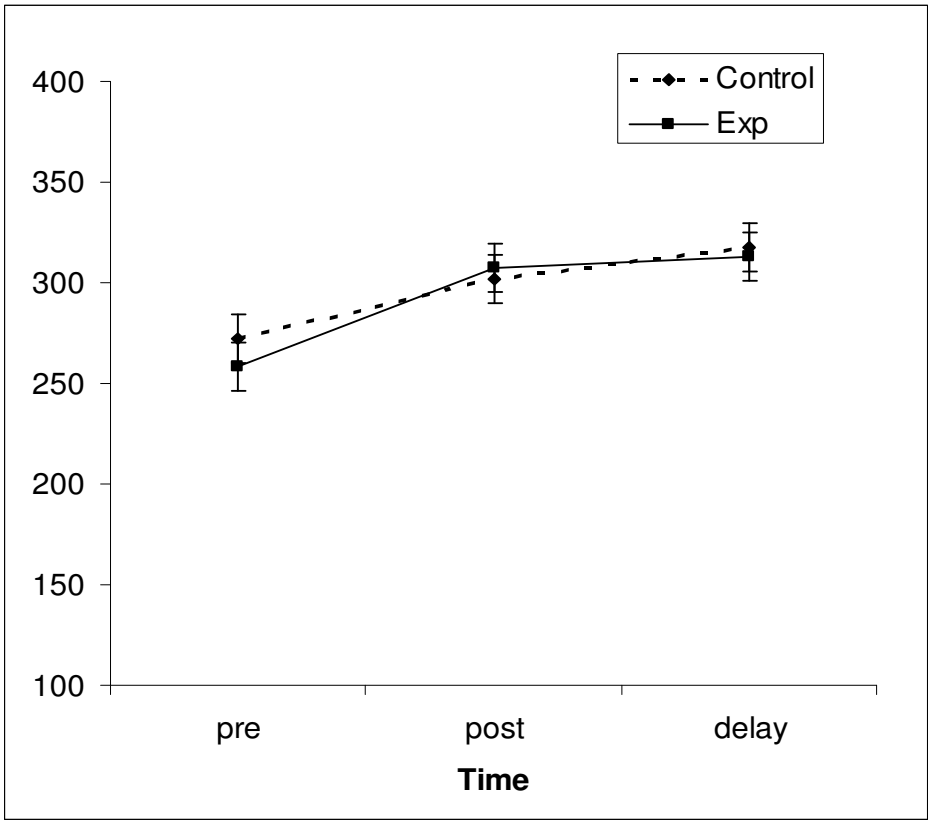

Figure 10. Mean number of words

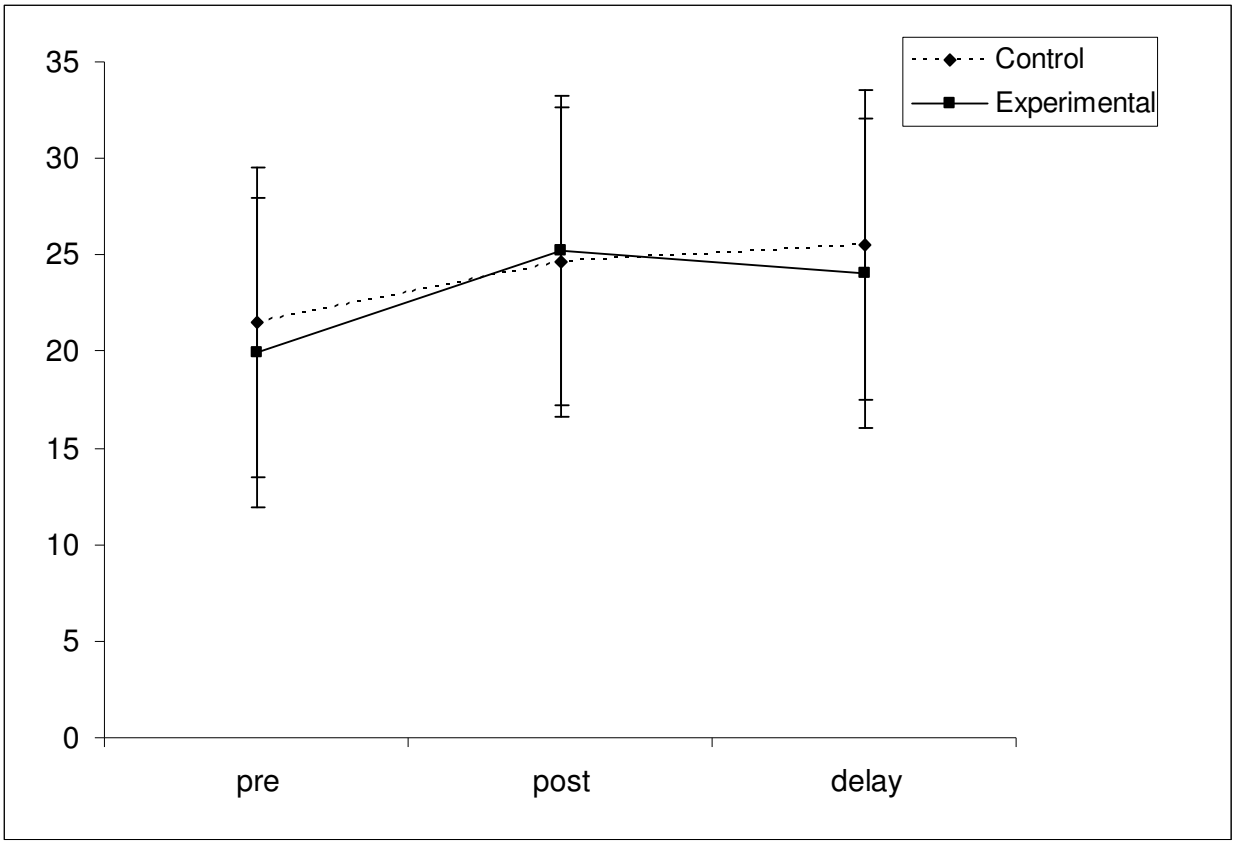

Figure 11. Mean number of T-units

Planned contrasts indicated that the significant increase occurred from pretest to posttest and was maintained at the delayed posttest; in other words, the pretest measures were significantly different from both the posttest and delayed posttest, which did not themselves differ; Table 16 presents the results of these contrasts.

These results were as expected, and indicate that the entire study population exhibited language development, in terms of fluency, over the course of data collection, and did so in a way that did not differ significantly between groups. As the intervention sequence was not designed to affect fluency, this result strengthens the argument that any differences in rater scores or evidence of treatment effect was attributable to the intervention sequence itself, rather

than language development or instructional experiences that occurred concurrently with the data collection.

Table 16. Planned contrasts examining main effect for Time on fluency measures

Time	Mean difference	*F*	*df*	*p*	*r*
			by Word		
pre-post	39.64	19.84	1, 91	< .001*	.42
post-delay	10.62	2.31	1, 91	.13	.04
pre-delay	50.26	33.71	1, 91	< .001*	.52
			by T-unit		
pre-post	4.23	20.51	1, 91	< .001*	.43
post-delay	.15	.03	1, 91	.86	.02
pre-delay	4.08	18.13	1, 91	< 001*	.41

Complexity. The two groups did not differ significantly with regard to W/T-unit a general complexity measure over the course of the semester (see Table 15 for descriptive statistics). A repeated measures factorial ANOVA found no significant main effect for time, F (2, 182) = 1.1, p = .34, for group, F(1, 91) = .4, p = .59, or for an interaction between the two factors, F(2, 182) = 1.53, p = .22. The results of these analyses indicate that, in terms of overall syntactic development, there were no group differences and no change in either group or the overall sample over the course of data collection. This lack of change was expected, based on the results of Ortega's (2003) syntactic development meta-analysis of complexity measures, which suggested that a minimum of a year of instruction is necessary before significant differences are able to be identified. The lack of significant group differences in syntactic complexity, however, again lends support to the argument that group differences in rating are related to the experimental intervention sequence.

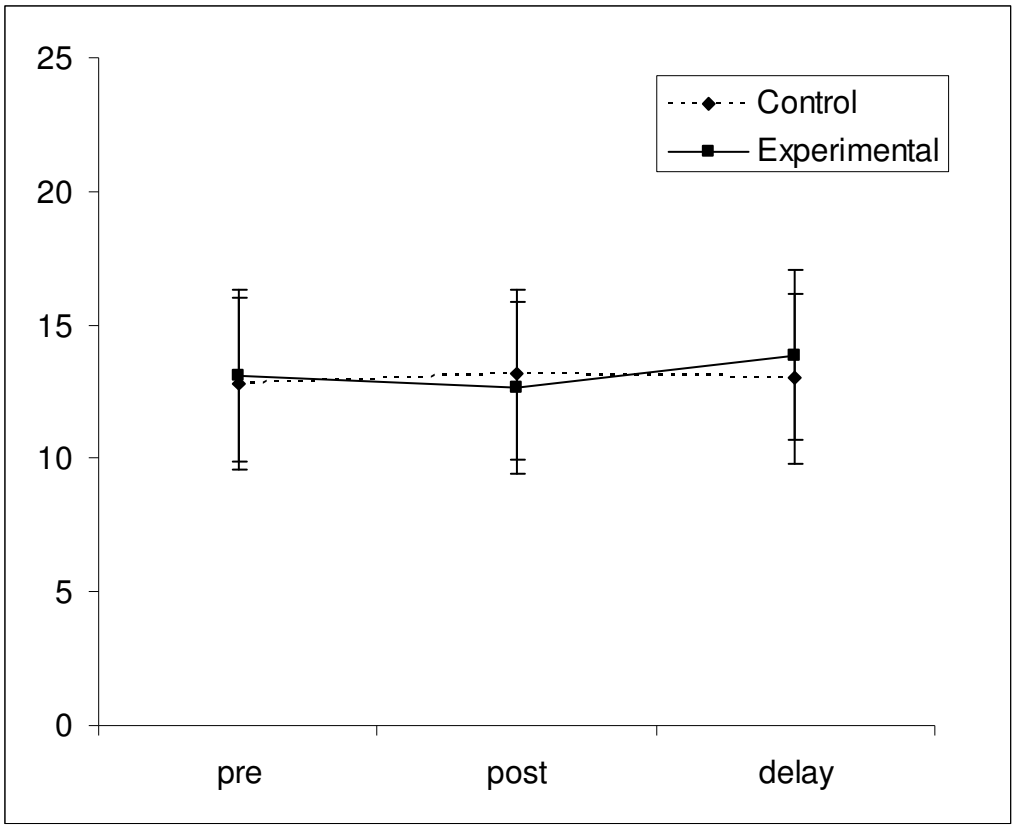

Figure 12. Words per T-unit by group and time

Lexical Diversity. The corrected type token ration (TTR), using a formula to account for text length, was calculated for each text (type/√[2*tokens]; Carroll, 1967 as cited in Wolfe-Quintero, Inagaki & Smith, 1998). The descriptive statistis are presented in Table 15. As can be seen in Figure 14, there was very little change for either group over the course of the semester.

A repeated measures factorial ANOVA confirmed this, as there was no main effect for time, $F(2, 182) = .17$, $p = .85$. There was also no main effect for group, $F(1, 91) = .88$, $p = .35$, indicating that the groups did not differ across the entire sample. However, there was significant interaction between group and time $F(2, 182) = 6.08$, $p = .003$. Table 17 shows the results of planned contrasts analyzing this difference. The significant differences occurred between pretest and posttest, and between pretest and delayed posttest.

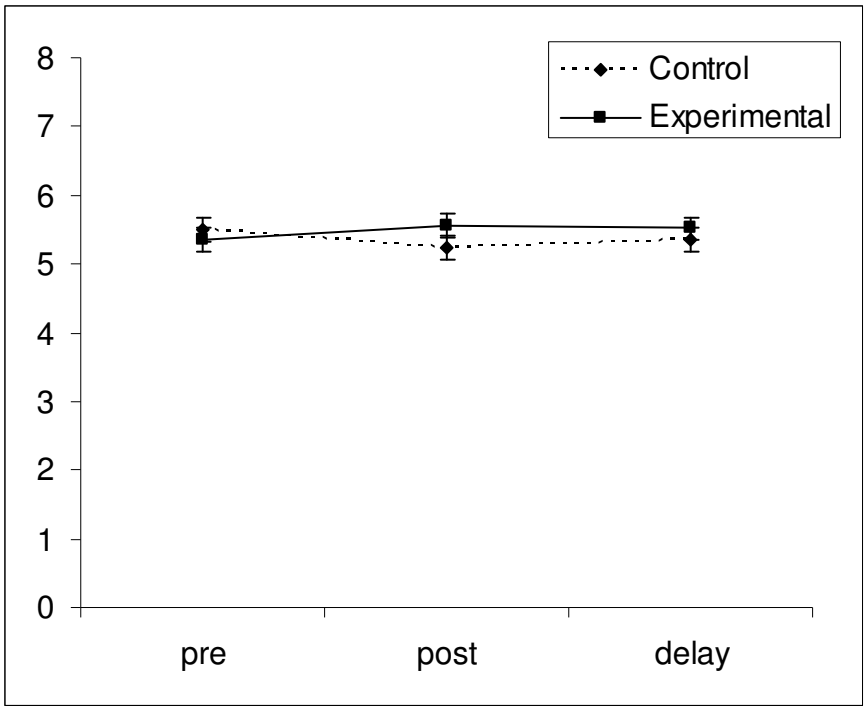

Figure 13. Type-token ratio by group and time

Table 17. Planned contrasts investigating effect of Group*Time (Type-token ratio)

Time	*F*	*df*	*p*	*r*
pre-post	10.31	1, 91	< .002*	.32
post-delay	1.18	1, 91	.28	.11
pre-delay	2.37	1, 91	< .02*	.16

An examination of the group means indicates that this difference was the result of a slight pretest to posttest decrease in type-token ratio for the control group, as well as a pretest to posttest increase for the experimental group that was maintained at the delayed posttest. The control group's TTR did increase again from posttest to delayed posttest, but not to the level of the original pretest score. These measures suggest that, after the pretest, the experimental participants used a wider variety of lexical types in their writing. The control participants actually used as less varied set of lexical tokens at posttest

Connections to Quality.Table 18 presents a non-parametric correlation matrix of the relationships between the mean Total scores and the developmental measures discussed in this section [*Note:* the correlation tables presented here and in the following sections of the results reporting were conducted as part of one, large analysis. The complete table can be seen in but for clarity and ease of interpretation is only presented in excerpted form in this discussion]

Table 18. Spearman's ρ for Rater Score and developmental measures

	Words	T-units	W/T	TTR
Total Score	.4**	.36**	.004	.29**
Words	-	.78**	.1	.36**
T-units		-	-.49**	.3**
Words/T-unit			-	.03
** $p < .0001$				

The significant correlations between these measures and the raters' judgments are certainly not surprising, as extensive research has established constructs such as fluency as central measures of writing development and proficiency. There are two results that need further explanation. The first is that the complexity measure is uncorrelated with measures of quality, which is likely an effect of the relative lack of difference in complexity among texts in the sample, rather than an indication that syntactic complexity is not a component of writing proficiency. The second is that the adjusted type token ratio correlated significantly and positively with the fluency measures, as the base type token ratio has been shown to vary inversely with length.

Developmental measures: Summary. In terms of broad-focus measures of language development, there was very little difference between the control and experimental groups. Both groups showed a significant increase in fluency over the course of the semester of data collection. Neither group increased their mean length of T-unit, suggesting that the level of syntactic complexity of their writing did not change. The lexical diversity of the texts, measured

by TTR, suggested that at posttest, the experimental group used a greater variety of lexical types than the control group, and that this difference persisted, but was somewhat reduced, at delayed posttest. This pattern is of interest as it is similar to that of the mean Total scores for the two groups.

The possibility of the group differences being driven by lexical characteristics of the texts offers intriguing connections to the present study's research questions. As discussed in the review of the literature, lexical cohesion may be the most influential factor in creating effectively cohesive texts. Furthermore, research has found that the most effective forms of lexical cohesion are cohesive chains created through complex repetition and paraphrase, rather than simple repetition. Texts which use a variety of terms to refer to key content, rather than repeating the same tokens throughout, would likely have a higher TTR than texts relying on simple repetition. The fact that the TTR measure correlated significantly with the mean total scores suggest that, if these complex lexical relations were indeed what were driving the group differences in TTR, then they were judged to be effective by the raters.

Research Question 1

RQ1: Can the cohesion be represented as a single factor, or should it be treated as a multidimensional construct (i.e., lexical and connective cohesion)?

The first research question sought to determine if the various measures of cohesion, particularly the LSA and connector measures, could be thought of as representing a single underlying construct. In one sense, it seemed very unlikely that the two types of measure would load onto a single factor, as they reflected very different features of the text. On the other hand, connector use has been considered a component of the construct of cohesion since Haliday and

Hasan's (1976) work on the subject, and it would be of considerable interest to determine the relationships between the two components.

Table 19 presents the results of the PCA analysis. Factor loadings of above .4 were considered relevant to the analysis. Factor loadings that did not meet that threshold are indicated in grayscale text in the table. The analysis showed that there were three distinct factors with eigenvalues greater than 1. The factor loadings suggested that these corresponded to (1) lexical diversity, (2) connector use, and (3) lexical cohesion.

Table 19. Results of principal component analysis of cohesive element measures

	Lexical Diversity	Connector Use	Lexical Cohesion (LSA)
Type-Token	.9	.09	-.02
Voc_D	.96	.02	.02
MTLD	.86	-.03	.01
Sentence-level LSA	-.21	.01	.75
Paragraph-level LSA	-.38	.1	.48
Sentence-level LSA *SD*	.18	-.05	.77
Connectors per 100 T-units	-.07	.9	-.07
Categories of Connector	.07	.89	.04
Eigenvalue	3.3	1.66	1.12
Variance Explained	41.26%	62.04%	76.14%
Determinant = .026 KMO & Bartlett's = .71, $p < .0001$			

One aspect of the factor loadings needs further explanation. The two main LSA measures, sentence and paragraph-level cohesion, load positively onto the LSA factor, but also load negatively onto the lexical diversity factor. These loadings indicated an inverse relationship between lexical cohesion and lexical diversity, when lexical cohesion is measured by LSA. The

inverse correlation between these two factors has implications for the use of LSA measures to evaluate the lexical cohesion of texts, which are investigated further in the following section.

Research Question 2

RQ2. What are the relationships between cohesive devices (lexical and conjunctive) and measures of writing quality?

For the second research question, the hypothesized result was that a combination of high LSA and high lexical development scores would correlate with raters' judgments. However, the results of the PCA suggested that there was a direct, inverse relationship between measures of lexical diversity and the LSA measures. Before discussing the main analyses, an analysis of between and within group differences in the level of cohesion is presented.

LSA Measures. Figures 15 and 16 display the mean LSA measures, both the average of the vector cosines of adjacent sentences across the text and the average of the vector cosines between all paragraphs in a text; the associated descriptive data are presented in Table 20. Both measures displayed a slight upward trend over the course of the study, but a pair of repeated measures factorial ANOVAs found no significant differences for time at the sentence level, F_{sent} (1.91, 156.28) = 2.593, p = .08, group, no main effect for group, F_{sent} (1, 82) = 2.19, p = .14; F_{pgh} (1, 82) = 2.7, p = .1, and no interaction between the time and group, F_{sent} (1.91, 156.28) = 2.72, p = .07; F_{pgh} (2, 162) = 1.19, p = .31.

Table 20. Descriptive statistics for sentence and paragraph LSA measures

	Sentence LSA			Paragraph LSA		
Time	Mean	SD	Range	Mean	SD	Range
			Control			
Pretest	.19	.05	.21	.46	.13	.56
Posttest	.29	.08	.38	.48	.13	.43
Delayed	.32	.1	.37	.53	.13	.5
			Experimental			
Pretest	.2	.04	.21	.52	.11	.44
Posttest	.28	.07	.33	.52	.12	.51
Delayed	.27	.08	.37	.53	.11	.37

There was a significant main effect for time at the paragraph level, F_{pgh} (2, 162) = 2.97, p = .05. The significant result for a main effect for time for the LSA paragraph measure represents a small rise in the level of semantic relatedness of paragraphs over the course data collection for all participants

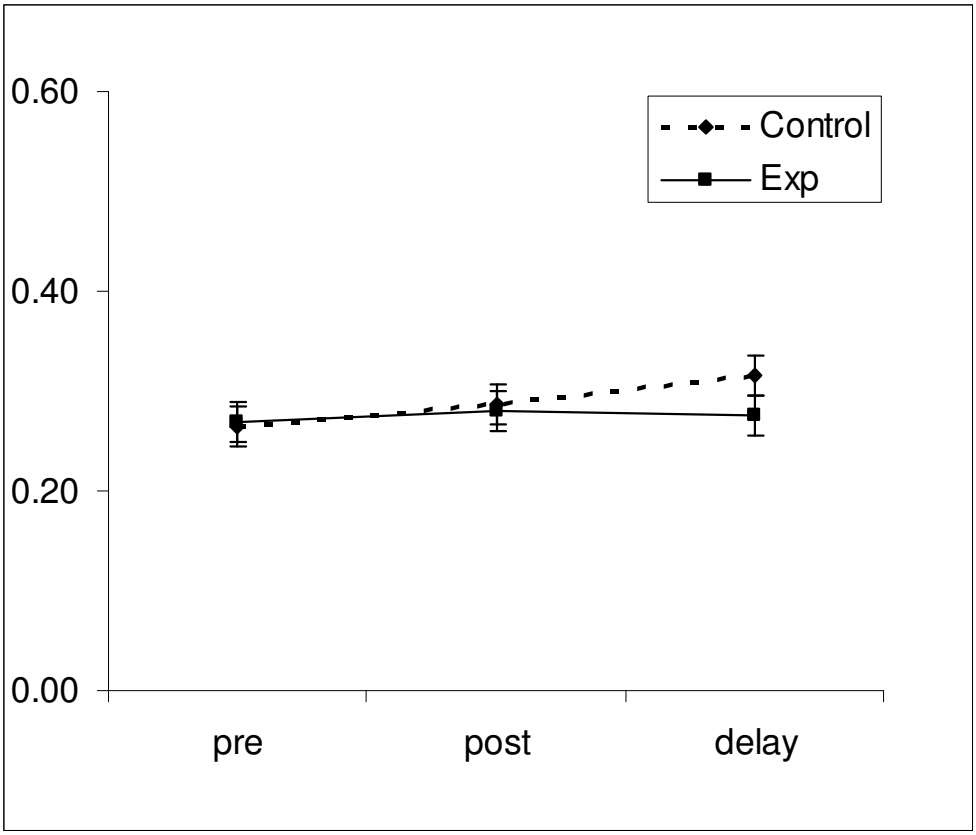

Figure 14. Mean sentence-level LSA measure

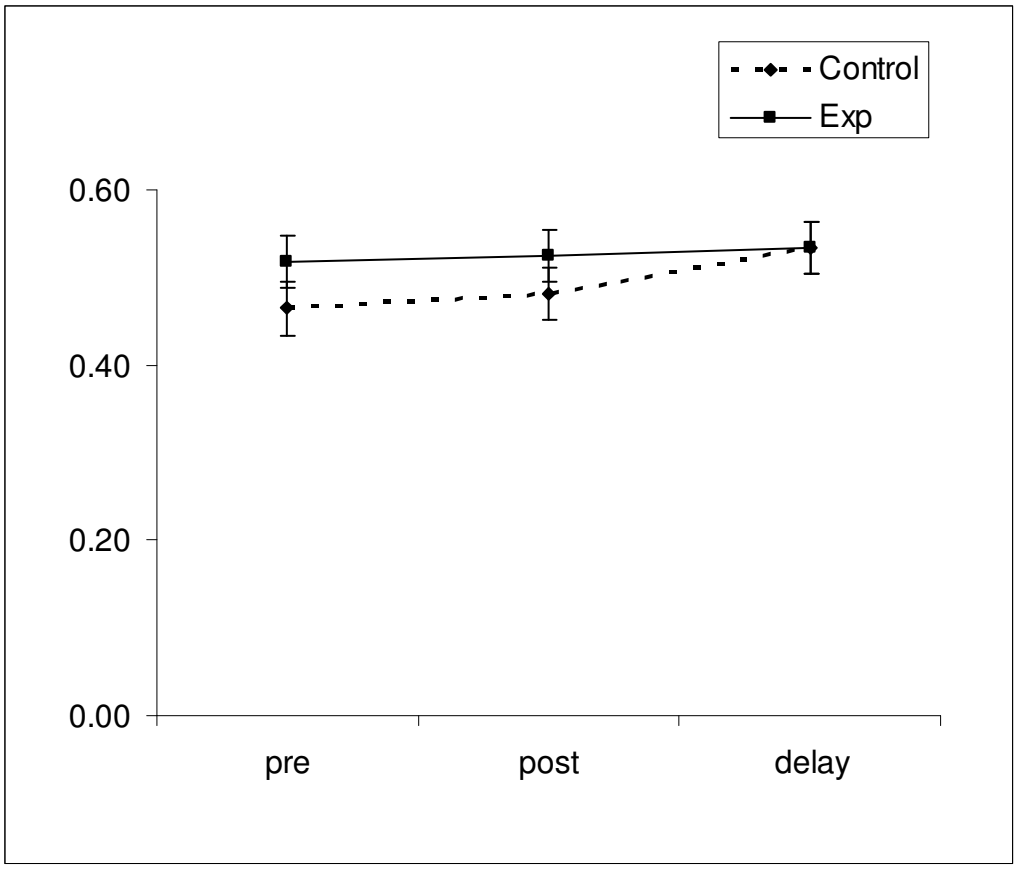

Figure 15. Mean paragraph-level LSA measure

A third measure calculated by the LSA software was the *SD* of each text's mean sentence-level LSA score. This was an incidental measure, and it was not used for between-group statistical analyses, but the *SD* does give some insight into how consistently a text's sentences related to each other: high standard deviations indicated a range of high and low sentence-pair relationships, while low *SD*s indicated that each sentence pairing was a similar level of relation. Of course, this measure would not indicate whether the degree of variability in sentence cohesion was effective or ineffective. Nevertheless, the measures provided additional insight into the patterns of lexical cohesion within texts.

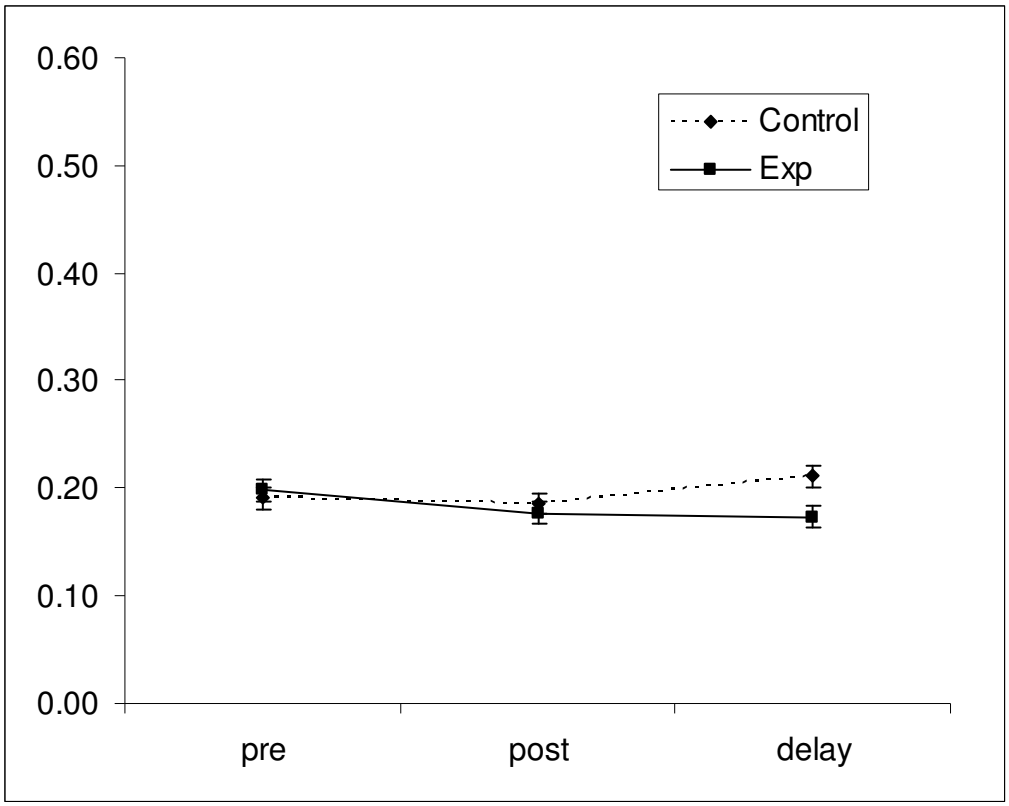

Figure 16. Mean *SD* for sentence-level LSA measures

Both groups showed a decrease in this measure from pretest to posttest, indicating that there was less variation in the amount of connections between sentences. From posttest to delayed posttest, the control group reversed the trend and increased, while the experimental group continued to decrease. Figure 18 shows the scatterplot for the LSA_Sent and the LSA_SD scores. It is clear that there is a roughly linear relationship between the mean level of lexical relatedness in a text and how much that relatedness varied between sentences.

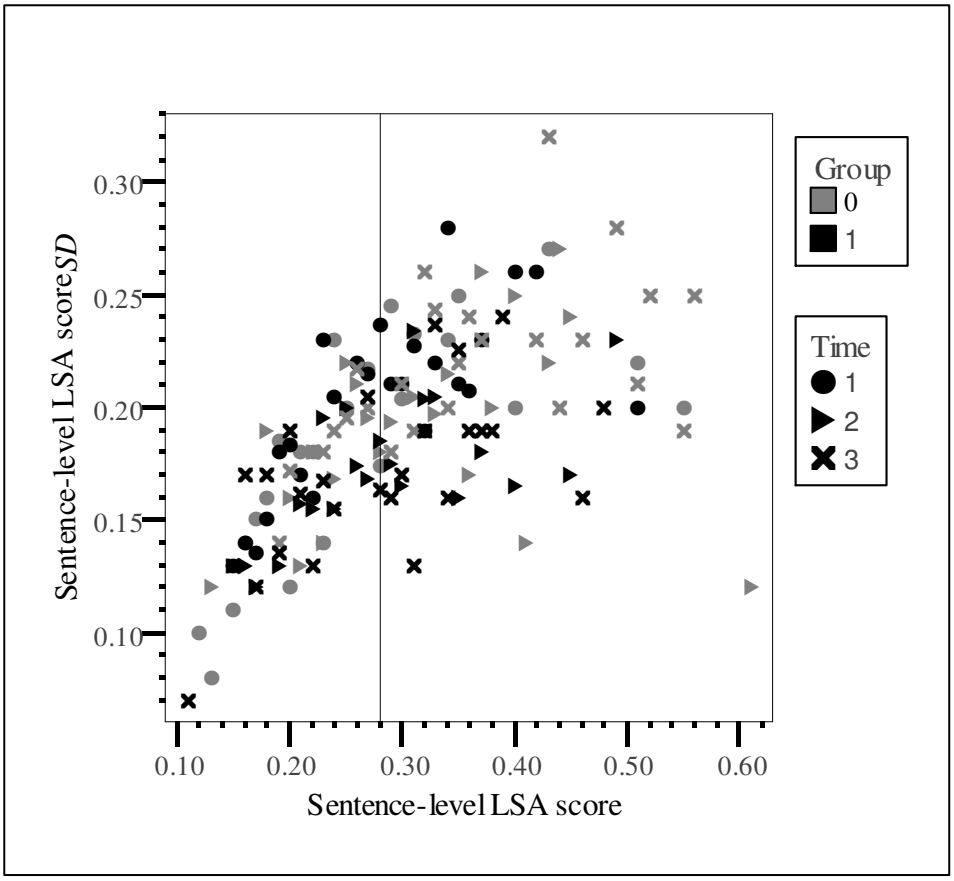

Figure 17. Scatterplot of sentence-level LSA score and standard deviations

The scatterplot also shows that, above the mean sentence-level LSA score of .28, the linear relationship is less distinct. To the right of the vertical line indicating the mean, the clustering of dots becomes more diffuse. The distribution shown in Figure 18 indicate that, as texts demonstrate a higher overall level of semantic relatedness between sentences, there is more opportunity for variation, more patterns of high and low-related sentences. For those texts which demonstrate a lower overall level of semantic relatedness, there is less variation, a larger number of sentence pairings demonstrate a similar level of connectedness.

Connector Use. Connector use is a second component of cohesion. Connectors, rather than creating relationships between textual units, instead serve as markers for relationships

created by lexical reference chains. Connector use was both a feature of cohesion as well as an explicit target of the intervention sequence in the present study. The detailed analyses of the patterns and changes in participant use are discussed in detail in the results of RQ3, focusing on the effect of the intervention sequence. For the analysis of RQ1, 3 measures are included. The first measure was the relative frequency of connector use per 100 T-units (Con/100T). This provided information of the overall frequency of connector use in the texts. The second measure, connector categories was a measure of the diversity of connector use. Each text received a score of 0-7 based on the number of different categories of connector, and thus the number of different relationship types, were signaled by the writer. The third measure was a ratio of enumerating connectors to the total number connectors used.

Summary of cohesion measures. To interpret these findings, it was necessary to determine whether there was any connection between LSA_Sent and writing quality has yet to be shown. In the present study, there were indications that LSA measures may not be the most effective means of teasing apart effective and ineffective lexical cohesion in writing, and even, as suggested by Folse (2007), that higher LSA scores had a negative, though indirect, relationship with writing quality. Indicators of this indirect relationship come out of the direct inverse relationship between LSA measures and lexical diversity, as measured by TTR, a relationship discussed in the following section.

Latent Semantic Analysis and Lexical Diversity. Table 21 presents the correlations of mean total scores, the measures of fluency and complexity and lexical diversity, and the LSA scores. As suggested by the results of the PCA, there were small-to-medium size correlations for both LSA measures, as well as the standard deviations of the sentence-level LSA score with

type-token ratio. These correlations were negative, indicating that there was an inverse relationship between TTR and LSA measures of cohesion. This is likely a result of the weight given to repeated terms in LSA calculations. Table 22 presents three sentence pairings, created by the author as examples, and the associated LSA scores.

Table 21. Spearman's ρ for rater score, LSA scores, and developmental measures

	Total Score	LSA_Sent	LSA_Pgh	LSA_ Sent*SD*	TTR	Words	T-units	W/T
Total Score	-	-.02	.05	-.04	.29**	.4**	.36**	.004
LSA_Sent		-	.48**	.54**	-.36**	.14*	-.01	.2**
LSA_Pgh			-	.24**	-.43**	.13#	.08	.07
LSA_Sent*SD*				-	-.18[&]	.06	.07	-.09
TTR					-	.36**	.3**	.03
Words						-	.78**	.1
T-units							-	-.36**
W/T								-

[#] $p = .03$
*$p = .02$
[$] $p = .002$
** $p < .0001$

As Table 22 demonstrates, the change of a single word can have a relatively large effect on the LSA measure of relatedness between two segments of text. A writer who uses a wider range of synonyms or hypernyms will almost necessarily produce a text with a lower cohesion score than a writer who engages in simple repetitions of the same word types.

Table 22. Sample sentence-level LSA scores

	Text	LSA score
Base	The old doctor opened his bag and prepared the needle.	-
Pair 1	The nurse glanced worriedly at the elderly **doctor**.	.59
Pair 2	The nurse glanced worriedly at the elderly **physician**.	.32
Pair 3	The nurse glanced worriedly at the elderly **man**.	.27

The use of synonyms, while potentially signaling a broader lexical repertoire, does not in and of itself create more effective writing. The examples in Table 18 are not intended to argue that a sentence pairing containing one token each of *doctor* and *physician* is inherently more advanced than a pairing containing two tokens of *doctor*, but simply to show how a repeated word can affect the LSA measure. The difficulty then is teasing apart the effects of lexical diversity and lexical cohesion on writing quality. A partial correlation, holding TTR constant, was run to determine if, separate from the effect of TTR, there was a relationship between LSA measures of cohesion and measures of writing quality. The results are presented in Table 23.

Table 23. Partial correlation for rater score, LSA score, and developmental measures, controlling for type-token ratio

	LSA_Sent	LSA_Pgh	LSA_Sent *SD*	Words	T-units	W/T
Rater Score	.11	.21**	.02	.35**	.27**	.01
Sentece-level LSA	-	.38**	.46**	.27**	.09	.21**
Paragraph-level LSA		-	.21**	.3**	.16*	.06
Sentence-level SLA *SD*			-	.15*	.15*	-.07
Words				-	.79**	.02
T-units					-	-.54**

*p = .01
** $p < .0001$

When the effect of TTR was controlled for, there was still no significant relationship between rater score and cohesion as measured by LSA at the sentence level. However, a relationship emerged between paragraph-level cohesion and the raters' judgments of writing quality. The fact that the more global measure of cohesion, rather than the local, correlates with writing quality lends further support to the theoretical position that effective cohesion is created by the interactions of lexical changes throughout a text, rather than simply at the local level.

Summary of LSA Results. The results of the statistical analyses of between-group and within-group differences for the sentence-level and paragraph LSA measures found no interaction effects for group and time. The only main effect was found for group on the paragraph level LSA measure, which indicated that, over the course of the semester, both groups increased the cohesion between their paragraphs of their texts.

There was also no clear relationship between the LSA measures and the mean total scores for the texts. This lack of relationship was probably driven to some extent by the negative correlation between LSA measures and the lexical diversity of a text, operationalized as TTR. When TTR was partialed out of the correlation analysis, a significant relationship was found to

exist between the paragraph-level LSA measure and raters' judgments of quality. These findings indicate that, although a growing body of research has reported on the links between LSA and other measures of language proficiency and development, both in written and spoken production, LSA analyses privilege the simpler forms of lexical cohesion over more complex lexical relationships, which prior research has suggested is more important for effective writing.

Research Question 3

RQ3: Can learner use of cohesive devices be modified through instruction, and is there a corresponding change in perceived quality?

The operationalization of cohesion was not able to identify group differences that might have accounted for the differences in rater scores. A second set of analysis analyzed the participant texts for direct evidence of the effect of the instructional sequence. The three pedagogical targets were (1) use of adverbial connectors (Con), (2) the use of *Determiner + summary noun* (DetSN) constructions, and (3) the use of definitional elements (DefEl). Unlike the language measures presented above, the targets of this set of analyses were not obligatory, and so a number of texts often contained no tokens. The absence of a particular structure is in itself possibly informative, but the relatively large number of zero values meant that inferential statistical analyses were not always appropriate. Group means were not normally distributed, and often had large standard deviations as a large number of cases were clustered at the zero value. Other indicators of central tendency, such as medians, could also be skewed given the large number of zero values.

Connector use. To investigate the participants' use of connectors, the relative frequencies and distributions of the subcorpora were first compared. Figure 19 shows the relative frequencies

of all adverbial connectors per 100 T-units. Both experimental and control subcorpora displayed an overall increase in the relative frequencies of Adverbial connectors across the three stages of data collection.

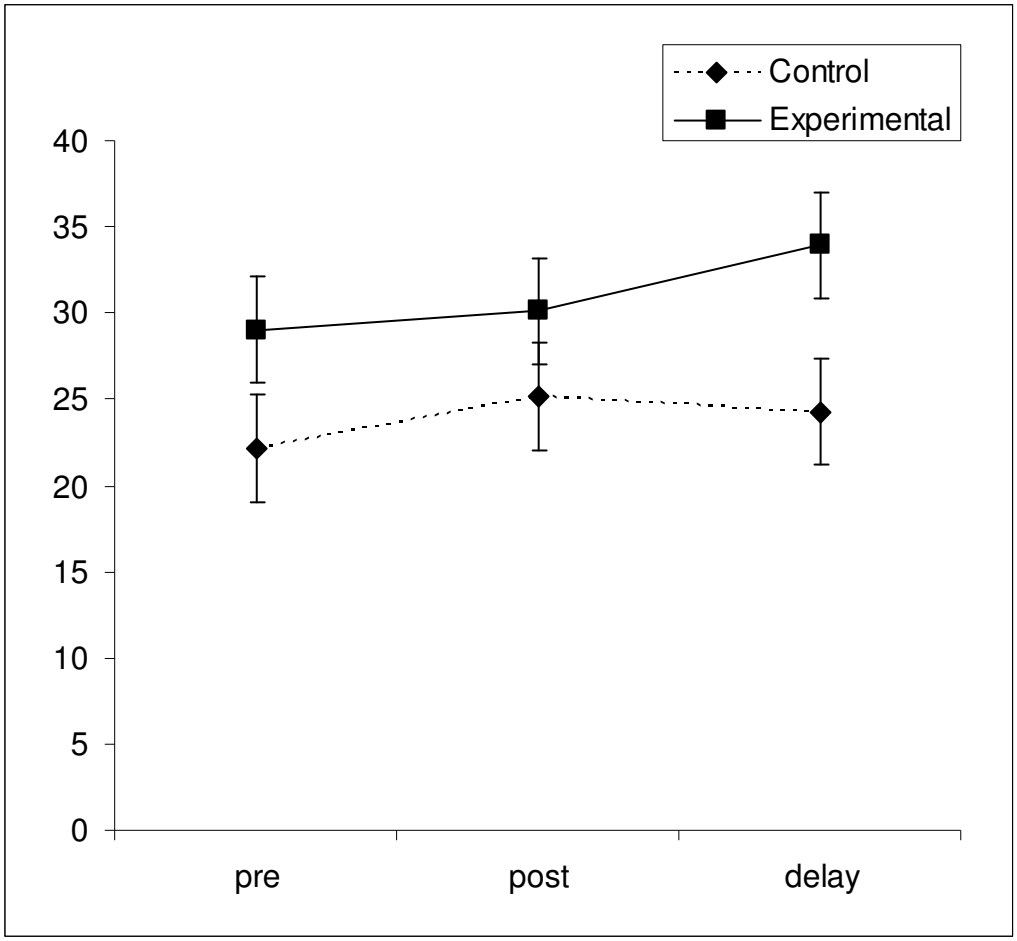

Figure 18. Adverbial connectors per 100 T-units

Using participant as the unit of analysis, the median relative frequencies were compared between groups and across time. Kolmogorov-Smirnoff tests and a visual inspection of histograms indicated that while some subcorpora did display a normal distribution, others did not. Three Mann-Whitney *U* tests confirmed that, at all stages of data collection, the experimental group produced significantly more connectors than the control group. A pair of

Friedman's ANOVAs were run to determine whether there were any significant within-group differences across time. The median frequencies are presented in Table 24 with the results of the Friedman's ANOVAs.

Table 24. Results of Friedman's ANOVA for connectors per 100 T-units

Group	Pre	Post	Delay	χ^2	*p*
Control					
Median	21.24	24.36	21.98	.08	.96
Range	50	96.66	59.09		
Experimental					
Median	29.41	28	33.33	5.89	.05
Range	84.21	69.23	66.44		

The results showed that the control group did not differ significantly across time. For the experimental group, the test did indicate a difference significant at $p = .05$. However, Wilcoxon Signed-Rank tests conducted as a *post hoc* analysis found no significant difference between any pairing of data collection stages (Table 25).

Table 25. Results of *post-hoc* Wilcxon signed-ranks test on Experimetal group connectors per 100 T-units

Time	*T*	*p*	*r*
Pre-Post	524	.86	.01
Post-Delay	394	.11	.23
Pre-Delay	423	.14	.22

Thus, an overall statistical analysis indicated that there was significant change in the experimental group's production of adverbial connectors, and while the descriptive statistics suggest the increase from posttest to delayed posttest the largest change, the significance of that change was not confirmed through statistical analysis.

Connector type. Previous research (Shea, 2009) has suggested that it is not simply the frequency, but also the type of adverbial connector used that affects raters' judgments.

Specifically, the proportion of enumerating connectors to total connectors used in a text correlated negatively with judgments of writing quality. Figures 20 and 21 present the use of each category of connector as a percentage of the total relative frequency of connector use (per 100 T-units) within the six subcorpora.

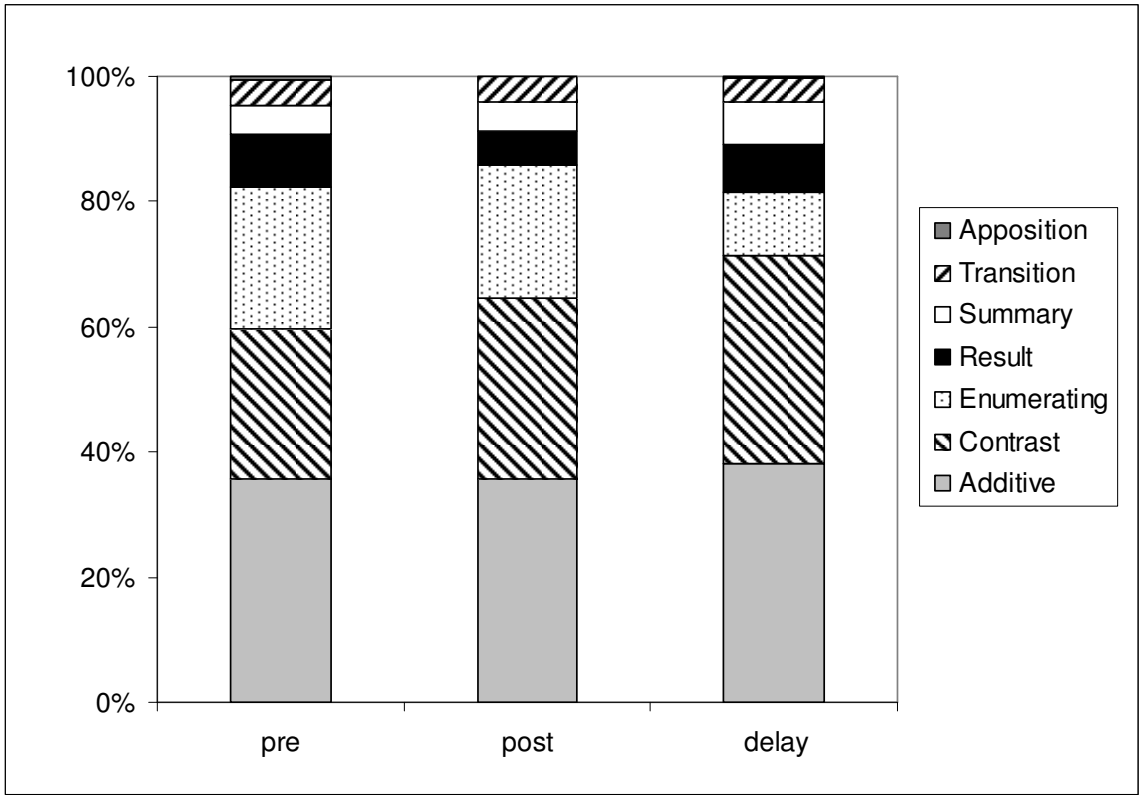

Figure 19. Percentage of connector categories per 100 T-units: Control

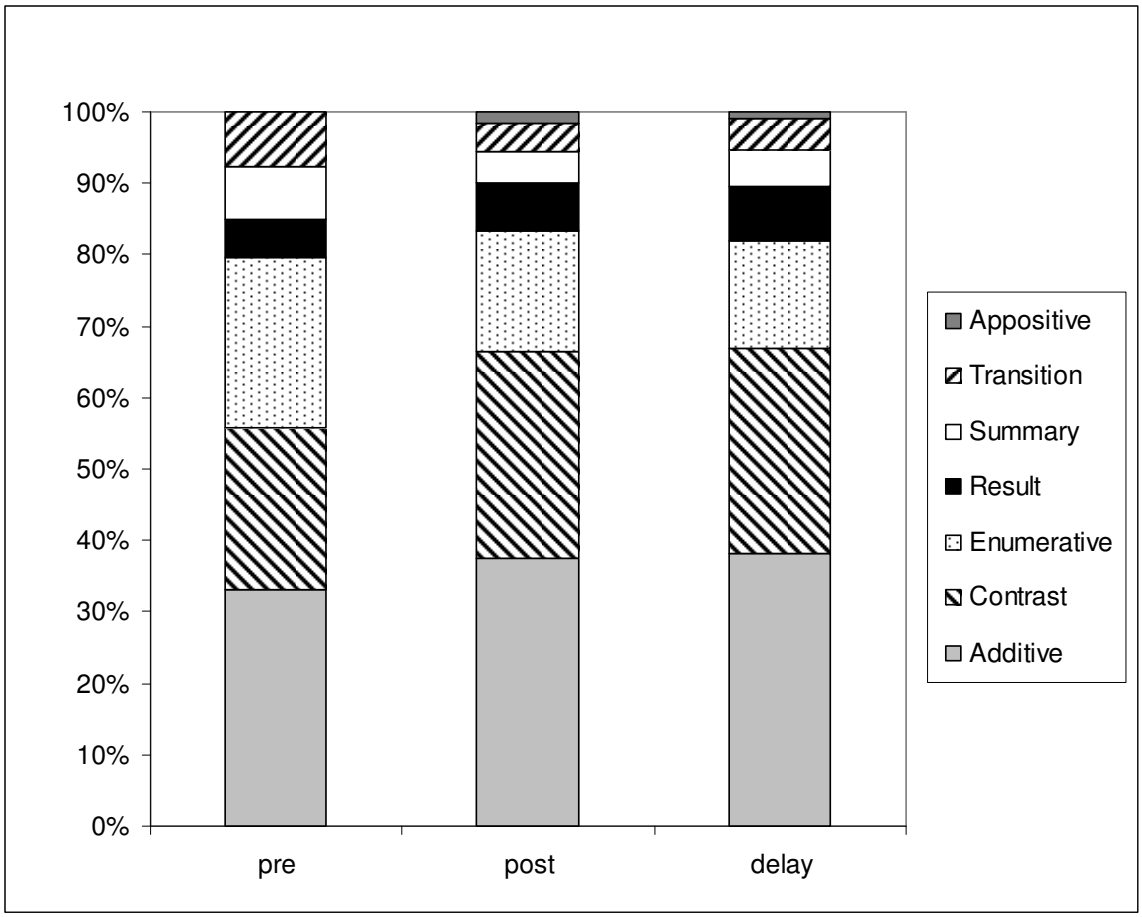

Figure 20. Percentage of connector categories per 100 T-units: Experimental

Figures 20 and 21 show that each group reduced its enumerator use taken as a percentage of the overall frequency of connector use. However, the two groups did so at different stages of data collection. A pair of Friedman's ANOVAs indicated that the control group's change across all three times was significant, $\chi^2 = 10.25$, $p = .006$ while the Experimental group's was not, $\chi^2 = 2.09$, $p = .35$. Wilcoxon signed-ranks tests conducted as *post hoc* analyses indicated that the differing patterns of change shown in Figures 20 and 21 were in fact significant. Both groups differed significantly in their pre and delayed posttest proportion of enumerating connectors. However, the control group's change occurred nearly entirely from posttest to delayed posttest,

and also differed significantly between those two scores. The experimental group exhibited a more gradual rate of change, and so did not demonstrate significant within group-differences between pre and posttest or between posttest and delayed posttest. The results of the Wilcoxon signed-ranks tests are presented in Table 26, and Figure 22 presents a graph of the two groups' means across times, which demonstrate the differing patterns.

Table 26. Results of Wilcoxon signed-rank tests for enumerating connector ratio

Time	Mean Difference	*T*	*p*	*r*
		Control		
pre-post	.01	313.5	.98	.04
post-delay	.11	92.5	.002*	.45
pre-delay	.12	94.5	.013*	.36
		Experimental		
pre-post	.05	307	.17	.2
post-delay	.02	291.5	.51	.1
pre-delay	.07	276.5	.04*	.3

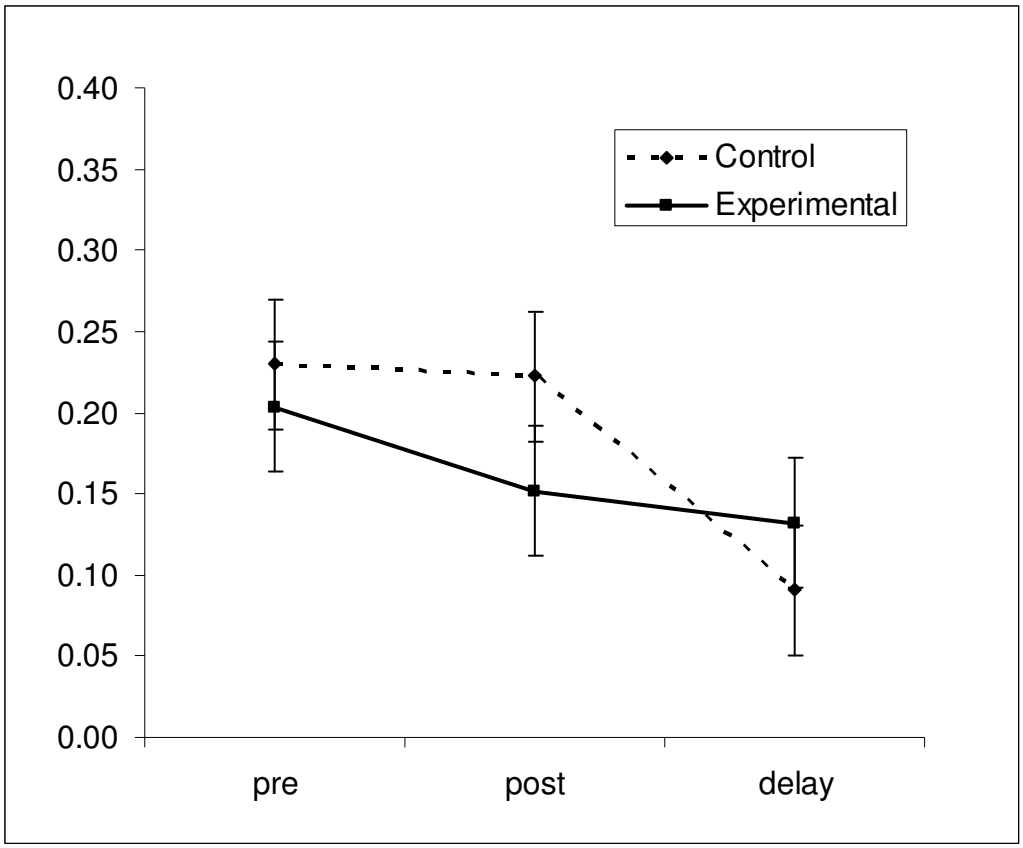

Figure 21. Ratio of enumerating connectors to all connector categories

In terms of enumerating connector ratio, there were no between group differences that received support from inferential statistical analysis. However, each group demonstrated a different pattern of change in the enumerating connector ratio. The control group showed little change from pretest to posttest, while the experimental group demonstrated a decrease that, while not itself statistically significant, did contribute to a significant decrease from pretest to delayed posttest. From posttest to delayed posttest, the control exhibited the largest decrease of the sample while the experimental group continued to decrease, but by a minimal amount.

Despite the limited findings of statistical analyses, a visual inspection of the data presents clear similarities to the significantly different patterns of the mean total scores, suggesting that

the lessening reliance on enumerating connectors was in some way a component of the broader changes in writing quality.

Variety of Adverbial Connectors. The ratio of enumerating connectors alone did not indicate any clear differences between the groups, although it did apparently correspond to the patterns of writing quality. The enumerating connector ratio focused on the use of one specific connector categories. A second analysis of the diversity of connector use was conducted using the number of categories of connector used by each group. Figures 23 and 24 show the counts of texts using a certain number of connectors.

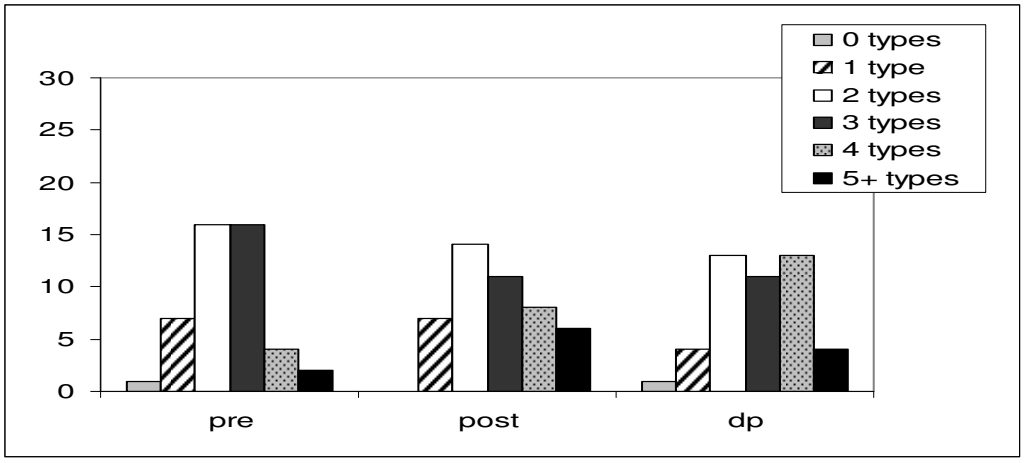

Figure 22. Control texts by number of connector categories

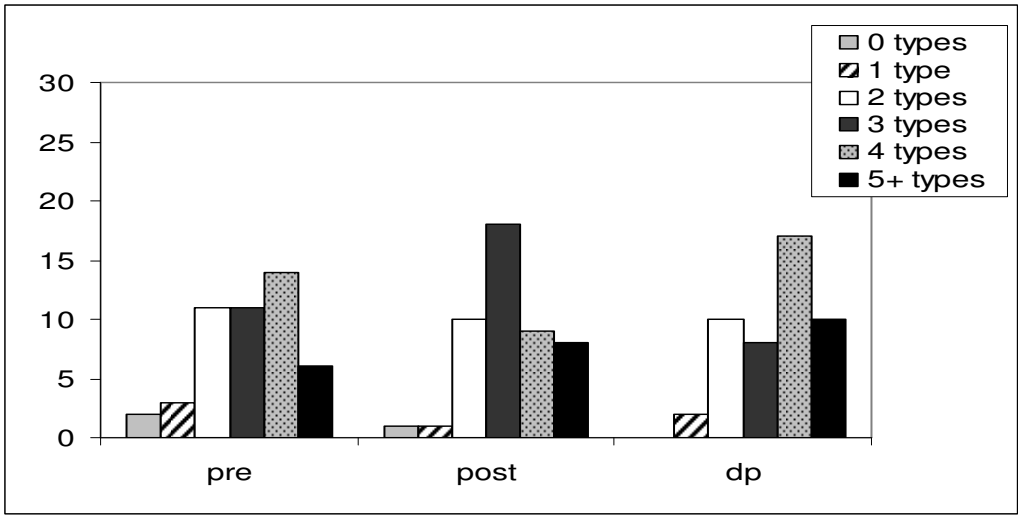

Figure 23. Experimental texts by number of connector categories

Only a limited number of texts contained five or more types, and so these were collapsed into a single category. The data presented in Figures 23 and 24 suggest several comparisons between the control and experimental groups. For both groups at all times, there were arelatively few low-category (0 and 1-2) texts. The control group consistently had more 1-category texts than the experimental group, although the number of 1-category control texts decreased consistently over the course of the sample. At the posttest, the control sample displayed a fairly broad distribution of text types, while, the experimental group contained more of both the 4-and 5+ category texts, and there were nearly twenty 3-category texts. At the delayed posttest, the control group again had a fairly even distribution of 2, 3, and 4-category texts. For the experimental group, there were fewer 2 and 3-category texts, and 4-category texts were the most frequent. There were more than double the number of 5+ category texts for the experimental group relative to the control group.

The distribution of connector categories presented a number of interacting patterns, and there was an arguable difference at posttest, the point in the data collection where the groups

differed. While the control group skewed toward the lower-type distributions, the experimental group texts were concentrated within the 3-type category. In addition, the experimental group produced more 4 and -5+ type texts than did the control group. The quantitative differences were not clear-cut however, but may point to more subtle qualitative differences, a point returned to in the discussion.

Determiner + Summary Noun Constructions

Pronominal vs. Determiner Production. Figure 25 presents the mean relative frequency 100 T-units of *determiner* and *pronominal* forms by both groups at the three stages of data collection. The pattern of *Pro* form production for both groups was similar, although the experimental group generally produced fewer forms than the control group. However, the divergent patterns of *Det* form production is of interest, as the pronounced difference at the posttest echoed the difference in raters' judgments.

The control group's production of *Pro* forms did not exhibit a great deal of change over the course of data collection, increasing by .8 tokens from prettest to posttest and decreasing by .4 from posttest to delayed posttest. The control group's production of *Det* decreased from pretest to posttest by approximately 3 tokens per 100 T-units and then increased by nearly the same amount from posttest to delayed posttest. The experimental group's production of *Pro* forms remained fairly steady throughout data collection, increasing by roughly .5 tokens from pretest to posttest and decreasing by that same amount at the delayed posttest. The experimental group's production of *Det* forms increased by 2.7 tokens from pretest to posttest, and that increase was maintained at the delayed posttest.

A pair of Friedman ANOVAs conducted on the two groups' performance found no significant variation in their performance over the three stages of data collection, although a *post-hoc* Wilcoxon signed-rank test indicated that the experimental group's increase in *Det* production from pretest to posttest was significant, $T = 324.5$, $p = .04$, $r = .29$.

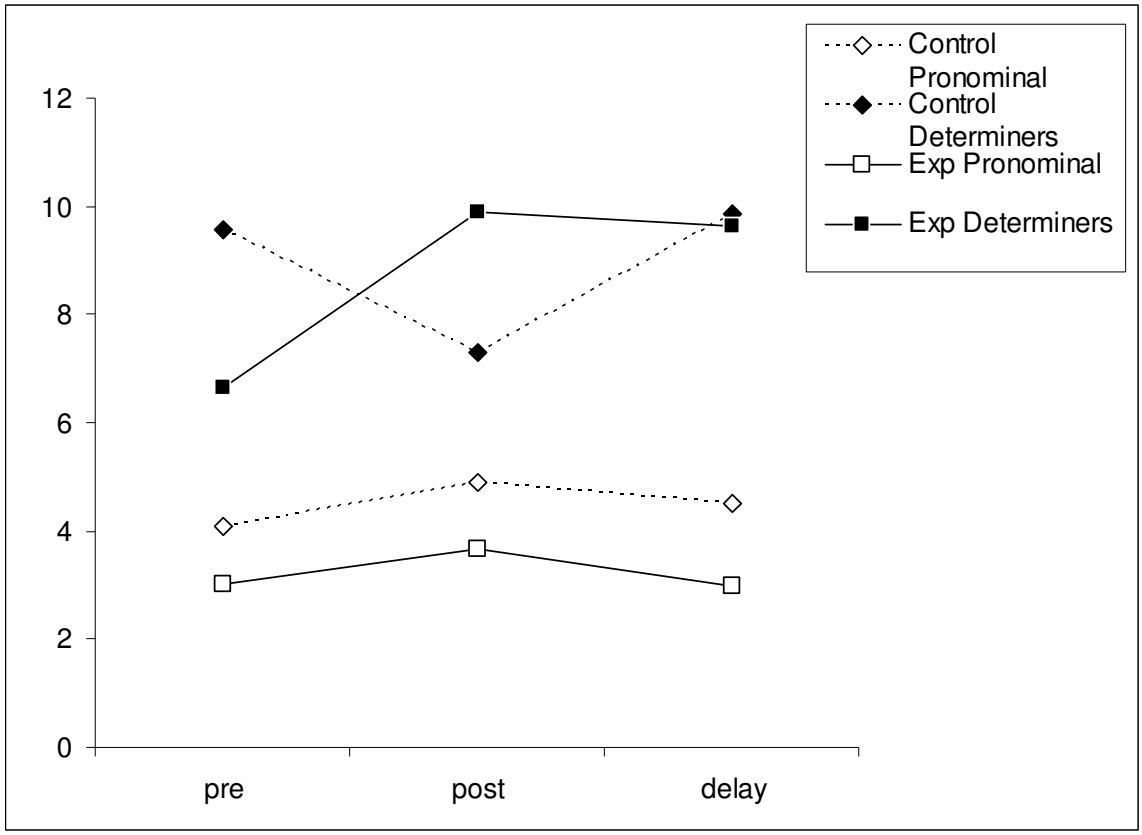

Figure 24. Production of pronominal and determiner demonstrative forms

Target Summary Nouns. In addition to the syntactic component of the determiner + SN construction, there was a lexical component. A set of summary nouns (Appendix C) was presented during the pedagogical intervention. It was of interest to determine whether participants in the experimental group had incorporated these lexical items into their writing. First, the results of the overall subcorpora are presented. Figure 26 presents the relative frequencies across the six subcorpora. It is important to emphasize that the analyses in this

section do not discuss the use of these summary nouns solely within *Det* constructions, but anywhere throughout the corpus.

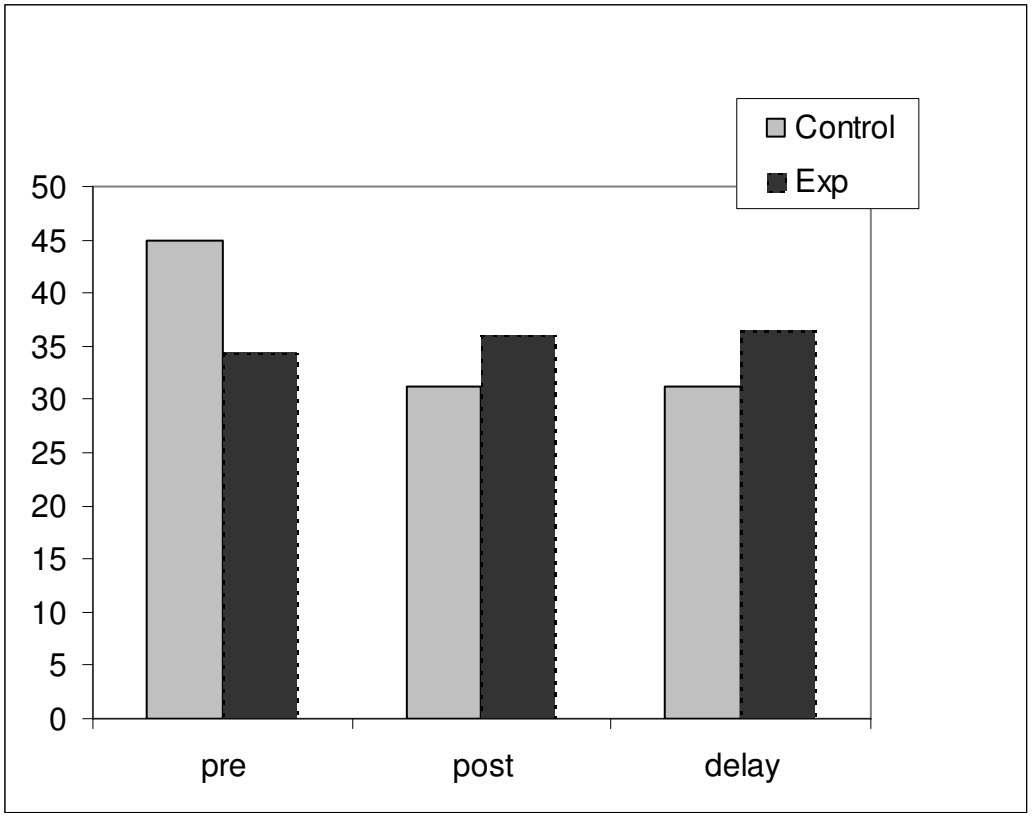

Figure 25. Production of target summary nouns per 100 T-units

As Figure 26 shows, at the prettest the control group actually produced more tokens per 1000 words. From pretest to posttest, the control group displayed a drop of approximately 12 tokens per 1000 words, and the experimental group increased by approximately 2 tokens. From posttest to delayed posttest, neither the control nor experimental group's production showed any appreciable change.

In addition to the small increase in the relative frequency of tokens, Figures 27 and 28 present the distributions of the terms. There were 49 summary noun types presented during the intervention sequence. No text contained tokens for more than 7 types. The control group's

distribution can be thought of as a baseline, as they received no instruction focused on those words as a particular set. There are few recognizable patterns in the control distribution histogram: the number of texts containing zero types remained much unchanged, but relatively low. There were slight decreases in the number of 4 and 5-type texts from pretest to posttest.

The experimental group presented a more evident pattern. The number of zero-type texts decreased from 9 texts at pretest 3 texts at posttest. The higher-type (4 and 5 types) texts also increased from pretest to posttest, and those increases were maintained at the delayed posttest.

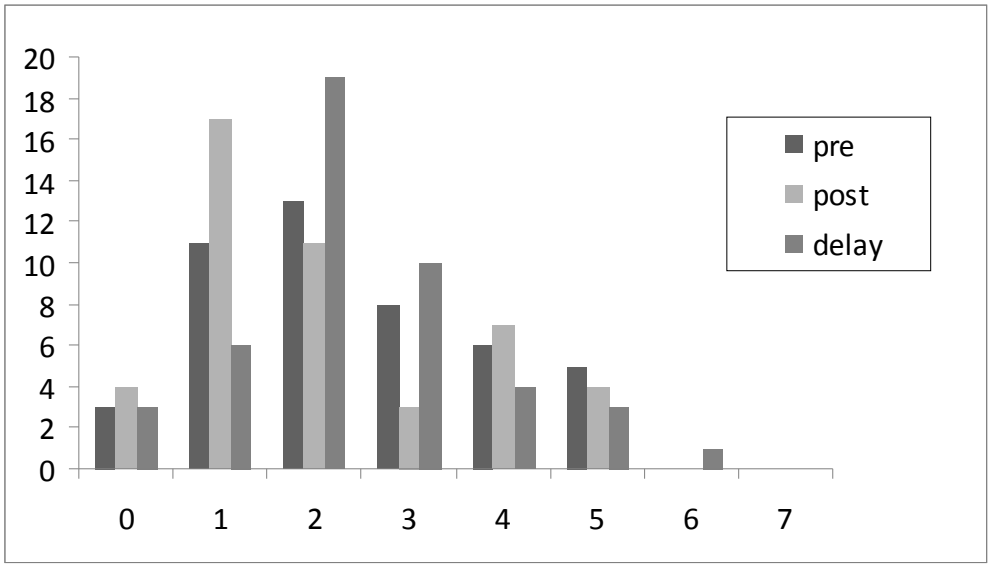

Figure 26. Control distribution of summary noun types

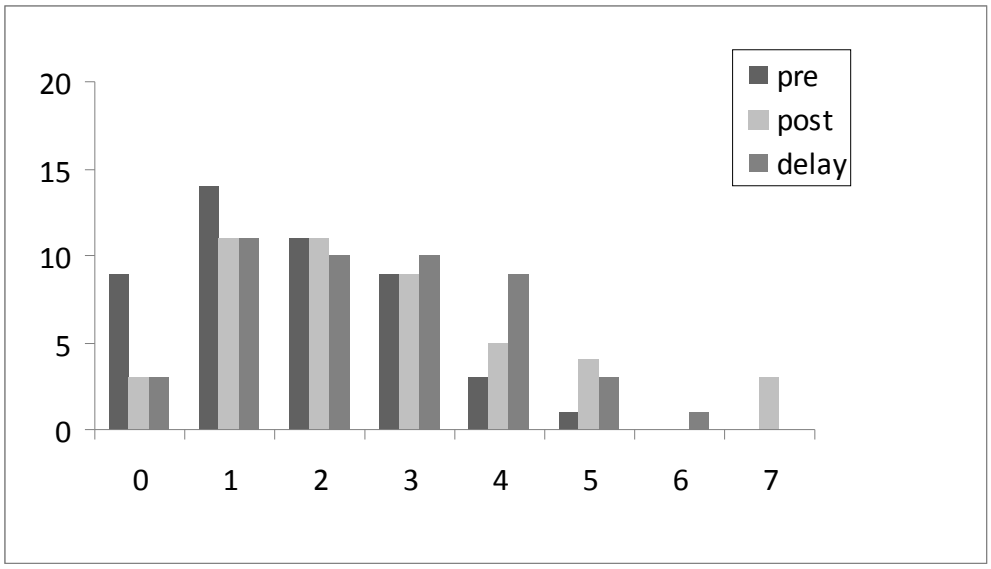

Figure 27. Experimental distribution of summary noun types

Of course, the data presented in Figures 27 and 28 only provide information on the distribution of these forms across the individual subcorpora, and do not provide insight into how individual participants were progressing. Table 27 presents the percentages of participants who registered increases, decreases or no change in the number of types at between the stages of data collection.

Table 27. Gains in production of target summary noun types

	Pre-Post	Post-Delay	Pre-Delay
		Control	
Increase	0.35	0.48	0.39
Decrease	0.50	0.33	0.39
No Change	0.15	0.20	0.22
		Experimental	
Increase	0.68	0.34	0.60
Decrease	0.26	0.40	0.23
No Change	0.06	0.26	0.17

From pretest to posttest, half of the control participants decreased the number of types of the target summary nouns produced. A little more than a third of the group increased the number of types produced. This trend reversed itself from the posttest to the delayed posttest, as nearly

half the participants increased and a third demonstrated a decrease. From pretest to posttest, an equal percentage of the control group (39%) increased and decreased the distribution of types of targeted summary words, while 22% demonstrated no change.

Relative to the control group, a larger percentage of experimental participants demonstrated an increase in the number of targeted summary noun types. A second notable difference was the low percentage of experimental participants which exhibited no change (6%). From posttest to delayed posttest, there was a relatively even distribution of participants exhibiting increases and decreases, and the number of participants exhibiting no change was 22% which was similar to the control group. Looking at the changes in distributions of types for pretest to delayed posttest, it is notable that the relatively high percentage of increases and low percentage of decreases recorded from pretest to posttest was maintained. In comparison, at the delayed posttest, an equal number of control participants had either increased or decreased their production of types of the targeted summary nouns.

Summary of preliminary analyses. Analyses at the level of the subcorpora and at the level of the participant indicated that from pretest to posttest, the control group decreased its use of both the *Det* construction and the targeted summary nouns. At the delayed posttest, the control group's use of *Det* constructions increased substantially, and its use of targeted summary nouns showed no change relative to posttest. The experimental group increased its use of both the *Det* construction and the targeted summary nouns from pretest to posttest, and maintained those increases at the delayed posttest.

The results of these initial analyses suggested that the treatment did have an effect, but it is not clear if, within participant writing, there was a connection between these syntactic and

lexical forms. In other words, at posttest, did the experimental group produce more summary nouns within *Det* constructions, or were the two phenomena unrelated?

Determiner+Summary Noun Constructions. Ultimately, the target of the pedagogical intervention was the use of *Det* constructions incorporating summary nouns. The initial analysis of the syntactic form indicated that experimental group produced fewer *Det* constructions, particularly at the posttest, but it remained to be seen what proportion of the *Det* constructions included *summary nouns,* as that was a focus of the intervention sessions. Figure 29 presents the relative frequency of *Det* constructions per 100 T-units across the six subcorpora, separated by type (summary vs. concrete). The control group displayed a drop in total constructions from pretest to posttest, which reflected a decrease in both types of constructions. The experimental group displayed an increase in both types across all three stages of data collection. The initial production of *Det+concrete noun* (*DetCN*)forms was much lower relative to the production of summary noun forms and demonstrated a larger relative increase from pretest to posttest, but both types of structure increased over the course of data collection.

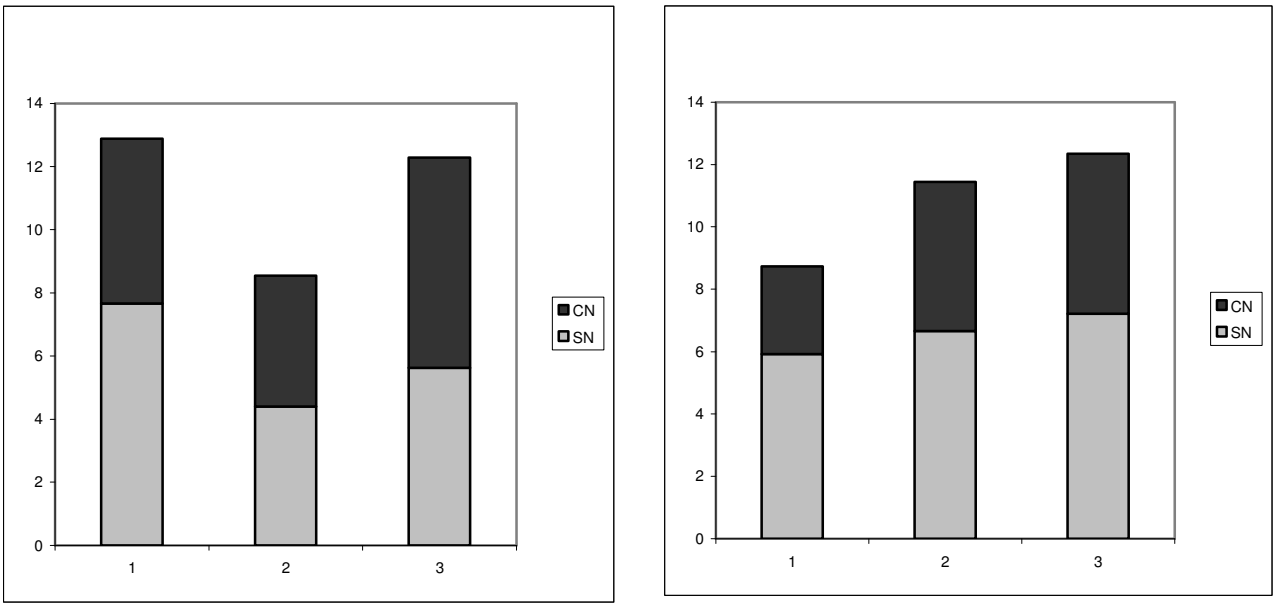

Figure 28. Determiner + Concrete Noun (CN) and Determiner + Summary Noun (SN) constructions in 6 subcorpora

Table 28 presents the same production data, in terms of the percentage of SN and CN constructions produced by each group at each stage of data collection. Both groups displayed a decrease in the percentage of SN. The control group's usage of SN decreased at each stage until, at the delayed posttest, the percentage of SN had dropped below 50%. Displaying a different pattern, the experimental group's SN percentage decreased to 58% at posttest and did not change from posttest to delayed posttest.

Table 28. Percentage of concrete and summary determiner constructions per 100 T-units

Type	Pre	Post	Delay
		Control	
Concrete	.4	.48	.54
Summary	.6	.52	.46
		Experimental	
Concrete	.32	.42	.42
Summary	.68	.58	.58

Looking at the data in Figure 29 and Table 28 together, it is clear that the decrease in the percentage of SN constructions occurred within different contexts of production for both groups.

For the control group, the decrease in the percentage of SN constructions at the posttest occurred in the context of an overall drop in the frequency of *Det* constructions. From posttest to delayed posttest, the control group increased its production of *Det* forms to a higher level than at pretest, but the increase represented in large part an increase in the use of *Det*+CN constructions. In contrast, the experimental group's decrease in the percentage of SN used occurred within the context of a consistent increase in the relative frequency of *Det* forms, and increases of both SN and CN constructions.

As for whether the experimental group's production of *Det*+*SN* constructions incorporated mainly the target summary nouns, initial analyses at the level of the subcorpora are presented in Figure 30. Figure 30 presents the production of *Det*+*SN* constructions per 100 T-units across the 6 subcorpora, categorized by whether the construction used one of the nouns targeted during the pedagogical treatment or another summary noun. It is clear from the figure that, within each group, the pattern of usage was similar whether the target nouns or other summary nouns were analyzed: the control group's production decreased from pretest to posttest and then increased from posttest to delayed posttest, while the experimental group displayed increases at both pretest and delayed pretest. The similarity between the control group's usage of targeted and untargeted summary nouns was expected, as for control participants, there was no reason to differentiate between the targeted SNs and other SNs. The experimental groups' increase for both targeted and untargeted summary nouns is of interest, as it suggests that the participants were able to generalize the strategy presented in the pedagogical intervention to other lexical items. This hypothesis is supported by the slight difference in the patterns of increase for the targeted and untargeted nouns. From pretest to posttest, the slope of the targeted

SN line was slightly steeper than that of the untargeted SN line. From posttest to delayed posttest, the pattern was reversed. This could be interpreted as a focus on targeted forms immediately following the intervention sequence, followed by greater attention to a wider variety of forms in subsequent writing.

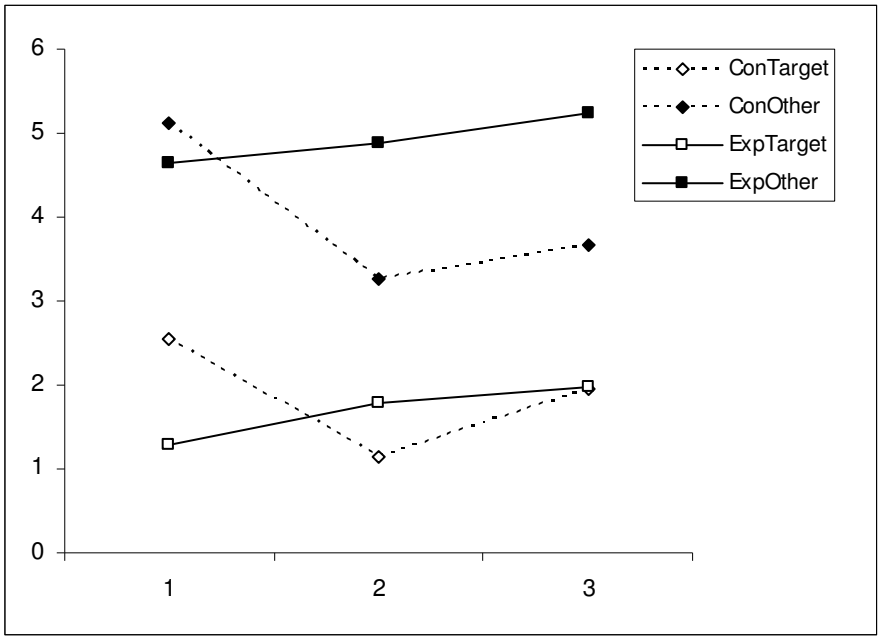

Figure 29. Production of Determiner + target summary noun and Determiner + other summary noun constructions

Definitional Elements

To investigate patterns of production of definitional elements, The subcorpora were first analyzed as units. Table 30 presents the relative frequency of definitional elements for each of the six subcorpora.

Table 29. Relative frequency of definitional elements per 100 T-units (by subcorpora)

Group	Time		
	Pre	Post	Delay
Control	6.78	6.88	5.62
Experimental	6.40	6.75	9.39

Unlike the patterns for the Det+SN constructions, the major between group difference for the production of definitional elements occurred at the delayed posttest. Both groups maintained the level of definitional elements at pretest and posttest. At the delayed posttest, the control group's production dropped by slightly more than one token per 100 T-units, while the experimental group's production rose by more than 2.5 token per 100 T-units.

It was also of interest to consider how the experimental increase manifested in terms of their distributions across texts. Figures 31 and 32 display the distribution of definitional elements across the texts in each subcorpora. For the control group, no clear pattern was immediately apparent. Noteworthy features of the distributions include the rise in the number of texts containing no definitional elements from posttest to delayed posttest. For the experimental group, there was a general pattern of decreasing low definitional element texts and an increase in high definitional-element texts. The number of texts with no definitional elements fell from 16 at pretest to 13 at posttest and then to 4 at delayed posttest. The number of 2 definitional element texts remained steady from pretest to posttest, then rose from 8 to 13 at delayed posttest. The number of texts containing 3 or more definitional elements rose from 7 at pretest to 14 at posttest, and then to 18 at delayed posttest.

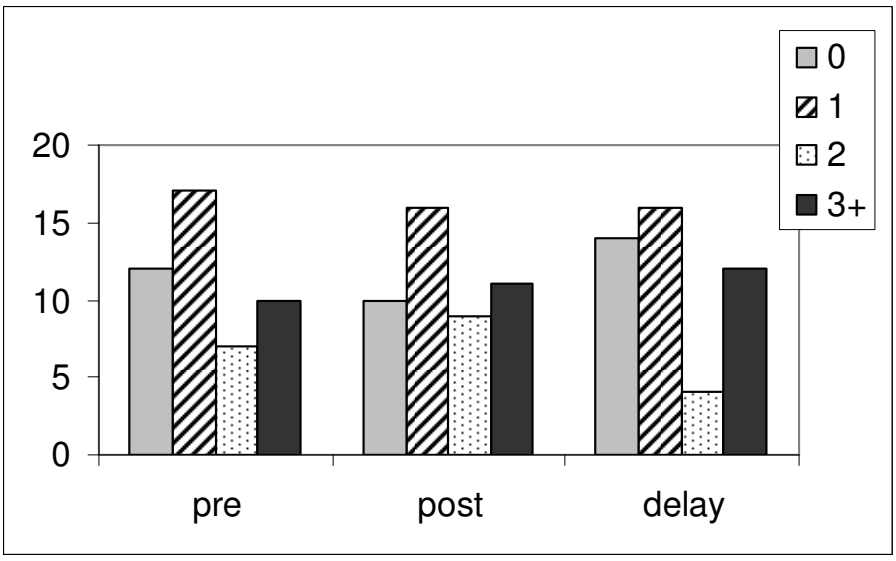

Figure 30. Definition of definitional elements across control texts

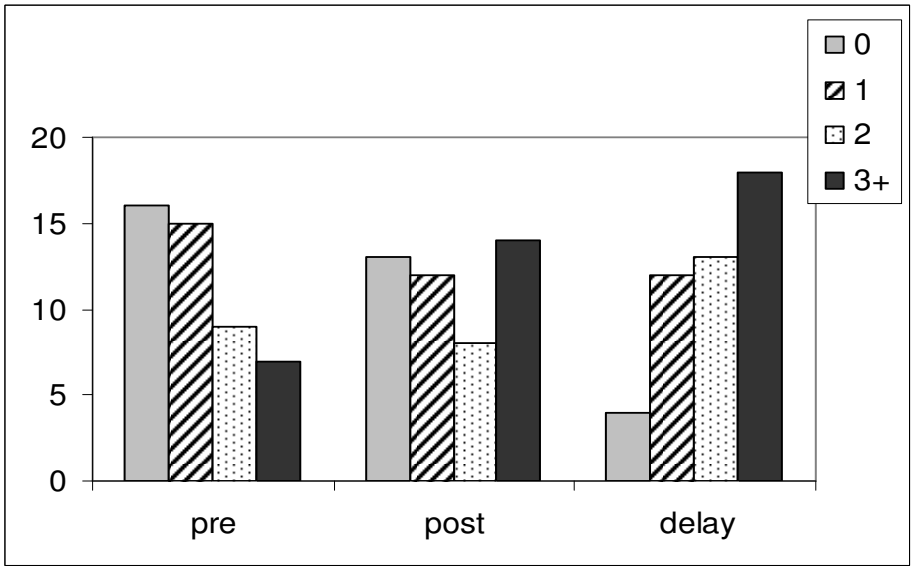

Figure 31. Distribution of definitional elements across experimental texts

Table 30. Percentage distribution of definitional element texts

tokens	Pre	Post	Delay
		Control	
0	.26	.22	.3
1-2	.52	.54	.43
3+	.22	.24	.26
		Experimental	
0	.34	.28	.09
1-2	.51	.43	.53
3+	.15	.3	.38

Table 31 presents these data in terms of percentages of the sample, collapsing the 1 and 2 definitional element texts for easier interpretation. Looking at the data for the control group, it is clear that there was relatively little change over time. The 1-2 band comprised roughly half the distribution at each time, and the remainder was split fairly evenly between the 0 and 3+ bands. The pattern for the experimental group was similar to that of the control group in one respect: the 1-2 band comprised roughly half the distribution at all three stages of data collection. However, the 0 band decreased from 34% at pretest to 9% at posttest, with the majority of the change coming between posttest and delayed posttest. The 3+ band doubled from pretest to posttest, and increased a further 8% at delayed posttest.

Overall, the control group did not display a clear pattern of change over time. There was very little change in the number of texts contributing 0 definitional element tokens: a 4% drop from prettest to posttest was followed by an 8% increase from posttest to delayed posttest. The other categories also showed little change: the largest change pretest to posttest was a 5% increase in the number of 2 definitional element texts, and the largest change posttest to delayed posttest was a 9& decrease in 2 definitional element texts and an 9% increase in the 3+ texts.

In contrast, the experimental group (Figure 32) distributions showed a clear drop in the number of texts containing 0 tokens of a definitional element. At pretest, 34% of the texts

contained 0 tokens. That percentage decreased to 28% at posttest and to 9 % at delayed posttest. There was a concurrent rise in the number of texts containing 3+ tokens. The percentage of 3+ texts doubled, from 15% to 30% at posttest, and increased a further 8% at posttest.

The frequency distributions provide an overall picture of the distribution of definitional elements but do not indicate how individual participants performed. Table 32 displays the percentage of participants in each group who increased or decreased between pretest and posttest and between posttest and delayed posttest

Table 31. Percentage of participants increasing, decreasing, or no change in definitional element production

Tokens	Pre-Post	Post-Delay	Pre-Delay
		Control	
No change	.22	.2	.28
Decrease	.39	.48	.41
Increase	.39	.33	.3
		Experimental	
No change	.28	.19	.15
Decrease	.32	.26	.21
Increase	.4	.55	.64

From pretest to posttest, the differences between the groups were not pronounced. However, between the posttest and delayed posttest, 48% of the control group decreased the use of definitional elements compared to a 33% increase for the experimental group. In the experimental group, 26% decreased from posttest to delayed posttest while 55% increased. From pretest to delayed posttest, 41% of control participants demonstrated a decrease in the number of definitional elements produced compared to a 30% increase, while 21% of the experimental group demonstrated a decrease compared to a 64% increase. Because the data were not normally distributed, a series of three Mann-Whitney U tests were conducted on the gain scores from pretest to posttest, posttest to delay, and pretest to posttest. The results are presented in Table 33.

Table 32. Mann-Whitney U for definitional element gain scores

Time		*Mdn*	*Min*	*Max*	*U*	*p*	*r*
Pre-Post							
	Control	0	-6	4	1000	.53	-.07
	Experimental	0	-5	4			
Post-Delay							
	Control	0	-4	5	816.5	.04	-.21
	Experimental	1	-6	4			
Pre-Delay							
	Control	0	-5	4	725.5	.006*	-.28
	Experimental	1	-4	4			

*significant at adjusted alpha level of $p = .016$

Taken together, the number of experimental participants showing an increase in the use of definitional elements (Table 32) combined with the significant difference in the gains made by the experimental group as compared to the control group (Table 33) suggests that there was an effect for the treatment.

Summary of effect of treatment.

Three intervention targets were analyzed: the use of adverbial connectors , the use of determiner + summary noun constructions, and the use of definitional elements. Based on the rater scores, potential group differences at posttest were of particular interest.

The experimental group produced more determiner constructions than the control group at posttest. Of those determiner constructions, a greater percentage were determiner + summary noun constructions. From pretest to posttest, the experimental group's increase in determiner+summary noun production made particular use of the summary nouns presented in the intervention sequence. However, from posttest to delayed posttest, the increase was driven more by the use of summary nouns that had not been targeted in the interventions.

For definitional element measures, the experimental group did appear to increase its production of definitional elements more than the control group did, but these between-group differences manifested themselves most clearly at the delayed posttest. While these results indicate the intervention sequence did have an effect, they do not account for the difference in scores at posttest. Because the analyses reported in the present study focus on frequency of occurrence, the possibility remains open that there was a change in the type or effectiveness of the Experimental groups DefEL production at posttest, while the quantifiable change only manifested itself at the delayed posttest.

Overall, the experimental group produced more adverbial connectors at every stage of data collection. However, while there was not unequivocal support provided by statistical analysis, the proportion of enumerating connectors to all connectors and the number of connector types used by participants suggest that the experimental group developed a more varied and sophisticated understanding and use of adverbial connectors. The groups differed on these measures most clearly at posttest, the same stage of data collection which yielded differing scores of writing quality.

Treatment targets and writing quality.

Table 34 presents the correlations between mean total score, developmental measures, LSA measures, and connector measures. The relative frequency of connectors per 100 T-units only correlated with other T-unit based measures. The number of categories of connectors correlated with the total score, fluency, and TTR ratios. All three connector measures correlated with each other, although the relative frequency and number of connector categories correlated

with each other between two and three times as highly as did either with the enumerator percentage.

Table 33. Spearman ρ for writing quality, developmental measures, and connector measures

	Words	T-units	Words per T-unit	Type-Token	Sentence-level LSA	Paragraph-level LSA	Connectors per 100 T-units	Connector Categories	Enumerating connector ratio
Rater Score	**0.40**	**0.36**	0.00	**0.29**	-0.02	0.05	-0.01	0.19**	-0.09
Words	-	**0.78**	0.10	**0.36**	**0.14**	**0.13**	-0.03	0.18$	0.10
T-Units		-	**-0.49**	**0.30**	-0.01	0.08	0.21**	0.11	0.07
Words per T-unit			-	0.03	**0.20**	0.07	0.28**	0.06	-0.01
Type-Token				-	**-0.36**	**-0.43**	0.02	0.16*	0.03
Sentence-level LSA					-	**0.48**	0.04	0.02	-0.03
Paragraph-level LSA						-	0.05	0.09	0.02
Connectors per 100 T-units							-	0.65**	0.20**
Connector categories								-	0.27**

#$p = .03$ $\quad$ $\$p = .002$

*$p = .01$ $\quad$ ** $p < .0001$

Table 35 presents the correlation matrix for the relationships between the measures of determiner + summary noun production discussed above, rater scores, and the LSA measures. In terms of the target form's relationship to writing quality, proportion of determiner constructions to all demonstrative construction, regardless of whether they included a concrete noun or summary noun, was the only measure to correlate with mean total score. In addition, the determiner/demonstrative ratio did not correlate with broader developmental measures, suggesting that the use of determiner constructions play a role in readers' perceptions of writing quality unconnected to more general features such as fluency or overall lexical diversity.

Additional significant correlations indicated that the number of summary noun types correlated with fluency and type-token ratio measures. This measure also correlated negatively with the LSA measures. Given he findings discussed above, which indicate that type-token ratio and LSA do vary inversely with each other, it was perhaps unsurprising that the summary noun-type measure, which is essentially a measure of lexical diversity within a very specific domain, also demonstrated a negative relationship with the LSA measures.

An additional problematic finding for the use of LSA as a measure of cohesion was the negative correlation of both sentence and paragraph-level LSA scores with the relative frequency of summary nouns per 100 T-units. Unlike the summary word type measure, the summary noun per 100 T-units measure had no correlation to type-token ratio, suggesting that this negative relationship was not the result of a more general lexical diversity. A key feature of LSA analyses are the weighting functions, which emphasize words that appear frequently in a particular text and infrequently in other types of text. One of the core elements of a summary noun is the fact that it can appear across a number of semantic contexts and with an array of referents. Summary

nouns' flexibility, a feature which makes them highly desirable from a pedagogical and rhetorical standpoint, may actually decrease the measured cohesion of a text.

Table 34. Spearman ρ for rater scores, developmental measures, and connector measures

	Words	T-Units	Words per T-unit	Type Token	Sentence LSA	Paragraph LSA	Determiner Ratio	Det per 100 T-units	Summary per 100 T-units	Summary per Det	Summary Type
Rater Score	0.40	0.36	0.00	0.29	-0.02	0.05	.19 $	.01	0.07	-0.06	.19**
Words	-	0.78	0.10	0.36	0.14	0.13	0.08	.02	-.09	-.02	.14 *
T-Units		-	-0.49	0.30	-0.01	0.08	0.00	-0.11	-.15*	-0.05	.16 #
Words per T-unit			-	0.03	0.20	0.07	0.12	0.26**	0.21**	0.04	0.00
Type Token				-	-0.36	-0.43	-0.01	-0.07	0.03	-0.03	0.24**
Sentence LSA					-	0.48	0.03	0.06	-0.13 #	0.00	-0.23**
Paragraph LSA						-	-0.05	0.02	-0.12 #	0.08	-0.21**
Determiner Ratio							-	0.55**	0.03	-0.10	0.02
Det per 100 T-units								-	0.08	-0.06	-0.04
Summary per 100 T-units									-	0.05	.79**
Summary per Det.										-	0.09

$p = .03$ $ $p = .002$
* $p = .01$ ** $p < .001$

Table 35. Spearman ρ for rater scores, developmental measures, and definitional element measures

	Words	T-Units	Words per T-unit	Type Token	Sentence LSA	Paragraph LSA	Definition Elements per 100 T-units	Definitional Elements
Rater Score	0.40	0.36	0.00	0.29	-0.02	0.05	0.01	0.10
Words	-	0.78	0.10	0.36	0.14	0.13	-0.11	0.11
T-Units		-	-0.49	0.30	-0.01	0.08	-0.21**	0.07
Words per T-unit			-	0.03	0.20	0.07	0.21**	0.07
Type Token				-	-0.36	-0.43	0.04	0.14
Sentence LSA					-	0.48	-0.01	-0.02
Paragraph LSA						-	-0.04	-0.03
Definition Elements per 100 T-units							-	0.94**

***$p < .001$

Table 36 presents the correlations between mean total score, developmental measures, LSA measures, and definitional element measures. The relative frequency of definitional elements per 100 T-units correlated with other measures calculated according to T-unit production and these relationships likely reflected a mathematical artifact than a theoretically relevant relationship. The raw frequency of definitional elements, which interestingly did not correlate with fluency measures, did correlate positively with type-token ratio. As type-token ratiocorrelated positively with mean total score, there is an indirect relationship between the use of defintional elements and writing quality. Keeping in mind the inverse relationship between type-token ratio and LSA cohesion measures, it seems that any cohesion created through defining language could not have been appropriately measured using LSA.

Overall, the relationships between the intervention targets and writing quality appeared limited. The total number of determiner constructions was found to relate to the mean total score. The distribution of types of summary words used in the determiner + summary noun constructions correlated positively with fluency and with type-token ratio, but correlated negatively with the LSA measures. For definitional elements, The raw frequency of definitional elements correlated with type-token ratio and did not seem to do so as a function of fluency. For connectors, the variety of connector categories used did correlate positively with mean total score, as well as with fluency and type-token ratio measures.

The correlation of a number of theses measures with type-token ratio indicates that, in as much as one of the goals the goal of the intervention sequence was to provide participant's with additional resources for the creation of complex lexical chains and thus, the creation of more effective cohesive links within their writing, the intervention sequence targeted appropriate

elements of written English. While determiner + summary noun and definitional element measures did not themselves correlate with the mean total score, their correlation with type token ratio and its correlation with the mean total score suggest an indirect relationship between these constructions and writing quality.

Interpretation of Results

To aid in the interpretation of theses results, a single participant's three essays were selected, chosen by the simple criteria of selecting a participant from he experimental group whose pattern of rater's scores followed the overall pattern shown by the group mean of a large increase between pretest and posttest, followed by a relatively small change from posttest to delayed posttest.

Tables 36 and 37 presents some of the descriptive data for these specific texts, along with the experimental group means for comparison.

Table 36. Mean rater scores for sample participant and experimental group

	Content	Organization	Vocabulary	Language	Mechanics	Total
			Participant			
Pretest	12.33	11.67	9.67	10.33	4.8	48.83
Posttest	15	12	14	12.67	6.5	60.17
Delayed Posttest	14.67	15.33	15.33	14.67	6.16	66.67
			Group			
Pretest	11.04	10.84	10.87	11.17	6.41	50.33
Posttest	12.61	12.53	12.45	12.28	6.66	56.48
Delayed Posttest	12.64	12.22	12.55	12.26	6.78	56.46

While the scale subscores were not included in the statistical analyses, they are provided here for additional context. The participant, *Jason* (pseudonym), began the study performing below the mean for the experimental group. While his content and organization subscores were slightly higher than the group mean, his vocabulary and language skills were lower. From pretest

to posttest, there was a dramatic jump in Jason's scores, with the largest increases coming in the content and vocabulary subscores, in addition to the mechanics subscore. These increases were maintained at the delayed posttest, while the organization and language subscores increased to a similar level. At both posttest and delayed posttest, Jason performed better than the experimental group mean.

Table 37. Developmental measures for sample participant and experimental group

	words	W/T	Ty/Tok	LSA_Sent	LSA_Par
			Participant		
pre	279	11.47	6.35	.27	.39
post	251	12.05	6.47	.22	.41
delayed	281	12.77	6.66	.11	.39
			Group		
pre	258.26	13.14	5.36	.27	.52
post	307.77	13.15	5.55	.28	.52
delayed	313.28	13.87	5.51	.27	.53

Looking at the broader developmental measures presented in Table 38, there does not seem to be an obvious change in fluency, accuracy, or lexical diversity that might account for Jason's increase in score. The number of words showed no pattern of increase, and actually dropped from above the man at pretest to below the mean at posttest and delayed posttest. Both the word per T-unit and Type-Token ratios showed steady improvement, but the complexity measure was consistently below the group mean and the lexical diversity measure was consistently above it. Neither seems to offer an explanation for the dramatic jump in Jason's scores from pretest to posttest.

The rightmost two columns present the individual and group LSA scores. Again, there is very little here that would indicate that Jason's essays were being judged as higher quality with

time. The sentence-level LSA measure decreased with time, but the paragraph-level measure remained practically unchanged.

To discuss the potential effects of the intervention sequence, the three texts are presented in Figures 33-35. For clarity, the spell-checked versions of the texts are provided. In Figures 33-35 adverbial connectors counted for the study are italicized, *Det/Pro* constructions are bolded, and definitional elements are underlined.

Table 38. Occurrence of intervention targets in example texts

	pr_TotalCON	pr_ConEN	Pro	Det	DetSN	TarSum	DefEl
pre	4	2	2	1	0	4	2
post	7	0	2	2	1	11	1
delayed	2	0	1	3	1	6	3

Table 39 summarizes the occurrences of the highlighted structures in the three texts. With regard to the use of adverbial connectors, it is notable that after using two enumerating connectors at pretest, Jason used none in his posttest or delayed posttest texts. The larger import of this seemingly minor change can be seen by looking at the three texts (Figures 33-35). In the pretest essay, the two enumerating connectors each begin a paragraph, and are indicative of the fact that the two paragraphs do not relate to each other in any particularly cohesive way; the first addresses governments' reactions to possibly criminal rich people, while the second discusses issues related to paper versus real wealth. The connector phrase signaling an opinion, which in the present study was coded as an additive connector, begins a one-sentence paragraph in the pretest essay that may be functioning as the essay's thesis. In contrast, in the posttest and delayed posttest essays, the connectors are embedded within paragraphs, and are used to signal local cohesive relations, rather than paragraph level shifts in topic.

There is a slight increase in the number of demonstrative constructions used, and that increase is the result of a more specific increase in the number of determiner constructions. Both the posttest and the delayed posttest contain an example of a *Det+SN* construction, which did not appear in the pretest sample.

Two definitional elements were identified in the pretest essay, both elaborating on Jason's discussion of real versus paper wealth. In the posttest essay, there is only one definitional element, but it occurs in an interesting context. The definitional element identified in the posttest essay. The identified definitional element provides elaboration on what Jason means by the phrase *stop their steps*. It is signaled by an appositive adverbial connector, one of the few appearing in the corpus. It is followed by a *Det*+SN construction using one of the summary nouns, *phenomenon*, introduced in the pedagogical intervention. Jason integrates the three techniques introduced in the intervention sequence in order to create an extended discussion of a fairly sophisticated idea: the slow waning of ambition in the face of difficult competition and unavoidable setbacks.

This segment of the text highlights two important points regarding the results of the present study. The first is that it is not my intent to argue that the particular segment is problem-free, or that it would not cause confusion for a reader unused to the writing of L2 learners. The cohesion strategies introduced in the intervention sequence were presented as serving a communicative function. Writing was conceptualized and discussed as a communicative act, in which a writer must try and anticipate the needs of the absent reader and provide additional support and elaboration when the writer feels the reader may have trouble grasping the ideas contained in the writing. The fact that Jason chose to incorporate all three of the cohesive

strategies at a point in the text which he was clearly struggling to gain control over the language to express his idea, and that the idea was central to his argumentation, suggests that this particular participant had developed the awareness of his writing necessary to apply the strategies in an appropriate place. The fact that the strategies were applied together indicates that he understood them as mutually supportive constructions for the building of a communicative message.

With the growing development of economy, people who catch the chance and opportunity in the booming-age become richer than the decade before in China. They gain tons of money just in short time. When the new rich men come out, there is a lot of problem coming with. The huge gap between the rich and the poor is the main trouble. So, some people provide the question that is it possible for someone to have too much money.

From my personally opinion, I do not believe that someone has too much money is impossible. There are serious reasons about my ideal.

Firstly, the citizens who obtain too much money are the troubles to the Country. We know that some billionaires in Russian were arrested in five years ago. They are rich people, but the government think they may do harm for the unity of country and become some Local power or authorities to against the national policies. So the government will control the balance of treasure.

Secondly, **this** is the age which everyone gaining the money equals to his or her work. No one can authorize a company which has great future. Because in nowadays, if you want to be rich as quick as possible, you have to let your company in to the stock market. You may have a lot of stocks in your company, but **this** is a paper currency which is only on the computer. **That paper** is not real money, and also it relates the stock market very tightly. Maybe, only one night, you lose your hole money.

All in all, I maintain that there is impossible to someone have too much money.

Figure 32. Jason's pretest essay

To be a successful man or woman should be most people's goals not only in **this day and age** but *also* in the past times. *However*, how to achieve their goals or make dream come-true becomes an issue to every individuals. Some of them hold the preference or success is the consequence of hard working, *on the other side*, people believe success also need luck. From my personal view, I obtain idea that success is not the result of hard working, but *also* it needs the lucky factor.

To be frank, Success is a good to everyone, so it means only few people can achieve their goals and satisfied themselves. Therefore, most goal-achievers stop their steps, because of the really crucial competition and gradually satisfied their work situation. *Namely*, they lose their ambitions; when they pursuit success. Why **this phenomenon** happens? Some failure may tell you that his boss does not like him, or the main manager is jealous his talent and worries he will replace manager position. *Indeed*, they work very hard, why the unfair things come to them and become a barrier to their career? We can say, **those people** need some luck in working positions. If the boss and manager are fair to every, they have chance of promoting.

So, *all in all*, whatever how hardworking you are, you need a person who are enough talent and obtains the eyes to discover your promotional abilities. I think **this** is the lucky factor in becoming success process.

Figure 33. Jason's posttest essay

One outstanding government should shoulder much more responsibilities and fulfill tons of various applications for their citizens. A question should never be underestimates that more attentions on providing excellent services for people is more important on supporting arts. *From my opinion*, I claim that a good government should pay more attention on its arts.

As you know, American is the only one powerful country in the world. The welfare system is quite advanced. The all kinds services which government provides to citizens are almost satisfied. *However*, depending on the short history of the American, and the boosting development in civilization, two out of three American cities looks like a same model. It is hard to the foreigners to distinguish the difference between Lansing and Grand Rapids. So the most American cities lose their icons or souls. They do not have many special names, because of the Lack of culture. **This situation** is impossible for most people from European or Asian. Let my own experiences as an example. I come from China. In my hometown, nearly every streets has its special name. Maybe in **this tiny shady streets** had five famous writer in History of China. Or, **that corner** was the most important historic building.

In my country, we have many Arts or historic features.

So, because of the limitation of American History. Government should pay much more attention and financial support to the few art records. **These** are the really worth to citizens. Good service just for the physic comfort. It can be improved by the development of the whole society. So protecting and supporting Arts which are the soul of one city even more one country should be never underestimated.

Figure 34. Jason's delayed posttest essay

CHAPTER 4: DISCUSSION

The construct of cohesion

The results of the principal component analysis conducted on measures of lexical diversity, adverbial connector usage, and LSA scores, indicated that there is likely not a unified construct of cohesion. Rather, cohesion appears to be made up of at least two separate elements, one being lexical cohesion and the other being the use of connectors to signal relationships between propositions.

This result is not necessarily unexpected. Lexical cohesion is created through a wide variety of interacting words and operates between both adjacent sentence pairs and long-range, across intervening sentences and paragraph boundaries. Adverbial connectors, on the other hand, tend to occur locally, whether between sentences or as organizing signals at the start of paragraphs. In addition, lexical cohesion is created through the use of a wide variety of open-class words, at varying levels of sophistication. Adverbial connector measures reflect the knowledge and production of a closed, specialized set of lexical items.

A second finding from the PCA, supported by the results of the Spearman's correlation analysis, suggested that Type-token ratio and lexical cohesion, at least when measured using LSA, have an inverse relationship within a text. That is, due to the fact that LSA scores are heavily affected by direct repetitions of lexical items, a text that uses a smaller variety of lexical items will likely receive a lower LSA score. It would be interesting to see if this same result obtained using other measures of lexical cohesion, for example, manual coding of lexical reference chains. However, the result raises questions regarding the ability to use automated methods to measure the cohesion of learner writing.

Cohesion and writing quality

Framed in the context of researchers' differing opinions on the use of repetition in learner writing, LSA measures appear to reflect a quality of text that would be valued by those researchers (e.g. Hinkel, 2003) who argue that the benefit of repletion to clarity and unity outweigh the potential negative effects of overly repetitious writing focused on by McGee (2009). At the same time, type-token ratio, as a measure representing the lexical development of a learner's interlanguage, is itself a desirable quality, as evidenced by its medium-effect ($\rho = .29$) correlation with rater scores.

A significant correlation is not of course a license to interpret causation, but based on the nature of the type-token ratio and LSA measures, it seems likely that type-token ratio, and lexical development, is a construct closer to the core of a learner's language, while LSA is a measure that is driven more by the language used in a given production task. Assuming this distinction to be true, then learners with a higher level of lexical development are more likely to produce texts with a lower level of lexical cohesion as measured by LSA scores. At the same time, a partial correlation analysis showed that when type-token ratio was held constant, paragraph-level LSA measures did correlate significantly with rater scores , though with a relatively low effect size ($r = .2$).

Effect of instruction

The measures of the effect of the intervention sequence provided the clearest positive results of the study. For this particular population, namely, college-level learners of English, familiar to some extent with academic learning and classroom writing, the analyses of target structure use showed changes from pretest to posttest or delayed posttest.

Connector use. At all three stages of data collection, including the prestest, the experimental group produced more adverbial connectors than did the control group. This rendered straightforward between-group comparisons unhelpful. However, within groups, the experimental participants demonstrated a significant increase in their use of adverbial connectors that the control group did not.

With regards to enumerating connectors, both groups demonstrated a significant decrease in the proportion of enumerating connectors to total number of adverbial connectors used over the course of the study. While the control group displayed the majority of that change from posttest to delayed posttest, the experimental group demonstrated a more gradual pattern of decrease.

As the example of Jason's essays showed, the change from the use of enumerating connectors to a more varied range of connector categories can signal a change in the methods of textual organization that a writer is employing. In Jason's first essay, produced at pretest, he used a organization form that relied on the listing of separate, unconnected arguments supporting his main thesis. The fact that the two main supporting ideas in his essay were unrelated had implications for the effectiveness of his conclusion and introductory paragraphs as well; essentially, they could say very little because there was very little in terms of a coherent main idea to discuss.

In the subsequent essays, Jason's decision to create a more unified text, exploring a single idea over a variety of paragraphs, naturally resulted in the use, indeed the absence, of enumerating connectors. This change occurred concurrently with the dramatic rise in rater scores that Jason's writing received.

It is certainly not my intention to argue that enumerating connectors are inherently less sophisticated than other types of propositional relationships. Nor is it my intention to argue that an enumerated text, moving through a series of separate causes, arguments, or other types of content is inherently less appropriate or less advanced than an elaborated text which addresses a single idea at length. There are certainly any number of tasks, both academic and outside the classroom, for which an enumerated or sequential listing of points is the most appropriate, and perhaps the only appropriate, organizational pattern for a writer to select. But for the prompts used in the in the present study, and for any number of other writing tasks used as language learning or content learning activities, it is not necessarily the case that an enumerated organization is better than an elaborated one.

Based on anecdotal evidence and my own experience as a writing instructor, and supported to some extent by the patterns of enumerating connector use in the present study, I would argue that while producing enumerated texts is not in itself a characteristic of a lower level of language development, often, students use it as a fall-back strategy: an easily constructed, relatively simple organizational style in which it is possible to write what is essentially a series of separate paragraphs connected by a general theme, rather than a coherent text which builds a discussion of a single idea.

Identifying organizational patterns in texts can be quite time consuming. Identifying the use of enumerating connectors is relatively simple. The ration of enumerating connectors to all connector categories appeared to decrease over the course of data collection for all participants, at the same time as their rater scores were increasing. While enumerating connector ratios would not likely be an effective means of assessing language development, as evidenced by the lack of

correlation with rater scores, it may serve as an indicator of the breadth of organizational patterns learner writers have in their repertoire.

Determiner + Summary Noun Constructions. At all stages of data collection, the control group produced approximately 1 *Pro* construction per 100-T-units more than the experimental group. While maintaining that difference, both groups displayed a similar pattern of *Pro* production, with the posttest production slightly higher than pretest of posttest, but no significant within-group differences. The production of *Det* constructions displayed a very different pattern both between groups and over time. The control group displayed a V-shaped pattern of production, lowest at the posttest, although a Friedman's ANOVA found no significant differences in production over time. The experimental group's production of *Det* forms demonstrated a pattern very similar to that of its rater scores, increasing from pretest to posttest and maintaining that increase at delayed posttest.

In terms of the target summary nouns introduced in the intervention sequence, the control group produced an initially high number of tokens per 1000 words, which decreased at the posttest. The experimental group displayed a very slight rise from pretest to posttest and no further change from posttest to delayed posttest. When the variety of summary noun types, rather than the frequency of tokens, was examined, there were very different patterns for the control and experimental groups. From the pretest to the posttest, 68 percent of the experimental group increased the number of types of summary nouns they produced, while only 35 percent of the control group did so. From pretest to delayed posttest, 60 percent of the experimental group displayed an increase in summary noun types, compared to 39 percent of the control group.

This difference in patterns of production reflected the patterns in rater scores. It also reflected the significant between-group difference found in type-token ratio at posttest. This is not necessarily surprising, as the count of SN types was in some sense a more focused version of a type token ratio. However, it provides some insight into what particular changes in lexical production were driving the changes in overall TTR measure. This point is expanded on further in the following section, but it raises the interesting possibility of connecting more focused, fine-grained measures of instructional effectiveness to broader, more commonly understood measures of language development that may not be as responsive to changes over shorter periods of time.

Taking the syntactic and lexical elements together, the experimental group demonstrated a clear pattern of increasing determiner construction use both over time and relative to the control group. For both groups, *Det+SN* constructions made up the majority of Det construction at pretest and posttest, although for the control group, the distribution at posttest was nearly equal. At the delayed posttest, the Det+CN constructions represented more of the overall Det production for the control group; the experimental group produced 16% more Det+SN constructions than Det+CN constructions at both posttest and delayed posttest.

Of the various measures used to represent the development and production of Det+*SN* constructions, one, the ration of Det constructions to Pro constructions correlated positively with rater scores ($\rho = .2$, a small effect). This result was expected inasmuch as the intervention sequence was designed to improve student writing, but it was also surprising, as the production of *Det* constructions might seem a relatively minor facet in the complicated array of lexicosemantic and discourse-level factors that comprise a piece of writing. However, just as the SN type measure correlated with the measure of TTR, likely representing a subcategory of the

overall language element measured by TTR, it is possible that this change on the part of the participants was a tangible feature of a larger understanding of cohesive relations and of reader expectations that underlay the intervention sequence.

Definitional Elements. In terms of relative frequency, the two groups did not appear to differ either within or between groups until the delayed posttest, at which point the experimental group increased its production by more than three tokens per 100 T-units. That this change in production occurred at delayed posttest, rather than posttest, is difficult to interpret in terms of its effect on rater judgments, as the two groups differed significantly in terms of rater scores at posttest only. However, the frequency of the definitional element may not tell the whole story. The fact that the experimental group did demonstrate a dramatic increase init s production at the delayed posttest is a strong indicator that the intervention sequence did have an effect.

In the example posttest essay, Jason only produced one identified definitional element, but it was deployed in conjunction with a number of other cohesive resources to create a cohesive sequence of discourse steps in which he makes an assertion, elaborates on that assertion to provide additional opportunities for his reader to understand his idea, and then uses a *Det+SN* within a rhetorical question to move his discussion forward. This sequence would not manifest itself in a frequency count of definitional elements, but the sophisticated use of multiple lexicosemantic and discourse constructions may be present in limited numbers throughout the experimental group's posttest texts, but with an increase in quality that contributed to the rise in rater scores.

Methodological Implications

One of the main difficulties in using the standard CAF (complexity-accuracy-fluency) developmental measures in writing research, or L2 research in general, is the fact that they are broad, and not as effective at distinguishing small changes over shorter periods of time, or differentiating within single proficiency groups or between adjacent proficiency levels. Often, the response to these difficulties is a call for more longitudinal research. Longitudinal research into the development of second language ability is of course desirable and necessary, but the logistic difficulties with such research designs are well known. There are also a number of benefits to shorter-term or cross-sectional studies, and a real value to knowing, at the level of a semester, what types of instructional practices and foci are benefit the development of L2 writing. In some sense, the CAF and lexical constructs might be thought of as highly resistant to instruction, and as representing aspects of language that may develop at very individualized paces, regardless of a particular course of instruction (assuming of course, a general equality in the quality of that instruction).

CAF measures then, are certainly necessary for researchers aiming to investigate the development of L2 writing. However, for research attempting to evaluate the effectiveness of experimental treatments, perhaps these CAF measures, though of obvious benefit due to their use in comparing language development across different populations, may be too broad to detect effects of particular interventions.

One take away from the present study was the fact that it was possible to detect changes in the higher-order targets of the intervention sequence, and these changes appeared to co-occur with short-term differences in rater's judgments of writing quality. In addition, it was possible to

connect some of the intervention-specific measures to more generalizable measures such as syntactic complexity (e.g., the correlation between the use of *DetSN* constructions and W/T).

After all, in the case of, for example, syntactic complexity, that complexity has to be built on something, presumably independent clause and phrases that the learner was not previously capable of producing or expanding upon. In this case, while an overall gain in syntactic complexity will get lost in the noise, the increase in *Det* constructions, which ultimately will contribute to an overall level of syntactic complexity, can be clearly measured.

This can perhaps serve as a model for researchers looking to conduct studies on the effects of particular pedagogical interventions, strategy instruction, or other short term, more explicit instruction. General measures of linguistic development should be calculates, but rather than using those measures as a dependent variable for the study, a measure specific to the intervention used should be selected. As a secondary analysis, and preferably a step carried out during pilot testing for study, these measures should be connected to one of the more general measures such as the CAF construct. This approach will have the benefit of providing researchers with the opportunity to use a measure that has some chance of detecting the effect they are looking for, but allows the use of language in discussing the results that can tie specific research findings to more widely recognized and understood measures of language development.

Limitations

The chief limitation of the present study stemmed from the fact that, contrary to indications from pilot testing, when applied to the larger corpus, the LSA measures proved to be a less than effective operationalization of lexical cohesion. There is of course a second possibility: that the LSA measures did indeed accurately measure lexical cohesion, and, as with

fluency, the experimental and control groups simply did not differ over the course of the study. However, the case of the LSA measures' negative correlation with the production of the use of summary nouns per 100 T-units measure (Table 29) is, I think, indicative of the disconnect between the LSA measures and the goal of the present study. The present study aimed to increase lexical cohesion while avoiding encouraging students to engage in overly mechanical repetitions of lexical items from sentence to sentence. Recognizing that alternatives such as synonyms relied on acquiring large amounts of domain-specific vocabulary, the pedagogical materials focused on constructions such as summary words and extended elaboration through the use definitional elements to create lexical cohesion by encouraging students to write in a more elaborated style in which they expanded and developed their ideas.

But due to the fact that these techniques were designed to be topic independent, they often resulted in segments of text that, although clearly recognizable as part of a cohesive chain by a human reader, were weighted by the LSA algorithm as providing poor differentiation between segments of text. By using summary nouns to create clear connections between propositions, the experimental participants were actually reducing the LSA score of their text.

One unanticipated, though certainly beneficial, effect of the treatments seemed to be an increase in lexical diversity, as measured by type-token ratio. This increase, combined with the relatively strong negative relationship between LSA measures and lexical diversity, may have rendered the use of LSA measures to track changes in learner writing unfeasible. Claims made by LSA researchers working in writing assessment and analysis, as well as pilot analyses for the present study, suggested that the targeted constructions, while not themselves direct sources of cohesive chains, would provide the textual environment for effective lexical chains.

Unfortunately, the complexity required for effective lexical cohesion seemed to defeat the ability of the automatic analysis to detect. This is not to say that LSA is not in many ways an effective tool for analyzing various samples of language, however, it was not an effective tool for assessing the effect of participants' abilities to create lexical relationships that (1) occurred throughout texts and (2) were manifested in a variety of lexicogrammatical relations. While it was not expected that LSA would accurately capture all the relationships signaled by, for example, pronoun reference, it was hoped that the overall relatedness of sentences and paragraphs would be represented.

This turned out not to be the case. Or, put another way, the textual similarities captured by LSA at the paragraph and sentence level were not those wither emphasized n the pedagogical interventions or those privileged by the essay raters. As previous research has found, it is complex and sophisticated lexical chains, in other words, reference chains that include a variety of lexical forms, which provide effectively cohesive texts. However, the results indicated that cohesion scores calculated by LSA are affected by the level of lexical diversity in a text, and thus do not accurately reflect the two dimensions of repetition and variety of form considered necessary for effective cohesion.

A second point, which is not necessarily a limitation but should be discussed, is the fact that in correlation analyses, even variables which demonstrated a significant correlation with rater scores or other developmental measures did so with a relatively low Spearman's ρ, typically between .2 and .4. Assuming that ρ can be interpreted similarly to Pearson's *r* (Ferguson, 2009), these should be considered low to moderate effect sizes. However, the measures used in this study to measure the use of connectors, of *Det+SN* constructions, and of definitional elements,

are looking at very fine-grained features of a participant's written production. Further, these features are likely influenced by a number of more basic variables, such core language proficiency and content knowledge. It may be that, within the context of the noisy data stream that is L2 written production, correlation coefficients should indeed be considered highly meaningful, particularly if they can be replicated across other data sets.

Inasmuch as there are not reference points by which to evaluate these correlations, their low size is a limitation of the present study. But they do provide an initial starting point from which to evaluate more fine-grained measures of writing development and treatment effect.

Future Research

The present study looked at cohesion in L2 writing using a wide variety of measures. The quality of the writing was assessed, as were features of general language development, automatically generated LSA measures, and specific measures of treatment effects. These measures were collected and analyzed in an attempt to develop a quantitative model of lexical cohesion that could function as a research instrument, and aid in curriculum and materials design, and contribute to the theory of textual composition. With so much data in so many different forms, there are a number of unresolved questions that remain to be addressed, as well as directions suggested by the current findings that might be fruitful avenues for future research.

One of the most salient features of the study results was the difficulty in teasing apart the effects of lexical diversity, operationalized as type-token ratio, and lexical cohesion, operationalized as LSA scores at the sentence and paragraph level. The TTR of a text correlated directly with the scores assigned by raters. When the effects of lexical diversity were partialed out of the analysis, LSA measures at the level of the paragraph also correlated positively with the

mean total scores. Ambiguity can be found in the literature on cohesion and writing instruction, namely, whether it is better to encourage student writers to repeat key terms in order to create cohesion, or whether such repetitions actually decrease the quality of the writing. That same ambiguity expressed by researchers and educators was found in the quantitative analyses of the texts collected for the present study. Future research should seek to examine the interrelations between lexical development and the choice of cohesive elements that learner writers employ, and relate these two features of learner writing to how it is perceived by a reader.

A second direction is to incorporate fine-grained assessments of the quality of the constructions that served as the operationalization of effect of treatment. In this extensive but initial analysis, the focus was on a quantitative measure of the frequency of particular language features. The criteria for selecting these features was largely formal, the framework for identifying DefEls was a good example of this. Another example would be the use of less effective summary words, such as *thing* or *stuff*, which in the analysis for the present study were not treated as different from more advanced or academic language.

However, it may be that these types of constructions do develop only, or even mainly, in terms of frequency. When discussion cohesion and coherence, many of the elements need only occur once in a text, if indeed they are required by the language system and by the relevant communicative conventions to occur at all. This raises the possibility that looking for an increase in frequency may not be the most effective way to tease apart how cohesion develops in L2 writers. To focus on the quality of particular types of constructions, rather than the quantity, may find more clear distinctions between two experimental groups, However, with the adoption of a more quantitative measure, the researcher gives up objectivity and reliability in their measures.

This is clearly a tension that researchers hoping to pursue the roots of textual cohesion in learner writing should be aware of.

Conclusion

The results of the present study were mixed. While the chosen operationalization of lexical cohesion proved ineffective, there were clear effects for a number of the pedagogical interventions provided to the experimental group. Experimental participants appeared to adopt a number of the techniques presented during the intervention sessions, and the increases in their use coincided with increases in rater scores.

Due to the non-obligatory nature of a number of these elements, there were often many zero cells: cases in which no tokens of a particular for were produced. This rendered some of the data unanalyzable through inferential statistics. However, in many cases, there were unmistakably congruent patterns of change in rater scores and the use of treatment targets.

The results reported in the present study focused on the relative frequency and the variety of forms of the intervention targets. Based on these objective criteria alone, suggestive connections between their development and the increases in rater scores could be drawn. Using the results of the present study as a guideline for search criteria, future research can identify these elements and begin to analyze how differences in the quality of their use, in addition to their frequency, might affect rater judgments.

The most disappointing finding was the failure of LSA measures to adequately represent the textual relationships formed by complex repetitions and paraphrases. This returns to the question about where the distinction theoretical construct of cohesion and coherence should be drawn. It is all very well to say that cohesion is that which resides in texts and coherence is that

which resides in the created understanding of the reader, but the results of the present study highlight how difficult it is to draw that line.

The case of summary nouns is most instructive to this point. As was seen in the results, the use of summary nouns in many correlated negatively with the LSA scores, as summary nouns, by virtue of their non-specificity, do not differentiate well between texts. At the same time, they are no doubt desirable components of academic language and should be added to students' linguistic repertoires.

The results of the present study drive home the point, raised by others, (e.g., Folse, 2007) that more cohesive texts are not necessarily better texts. The present study was conceived and designed with that thought in mind. The treatment targets and activities were designed according to best practices following the theoretical literature, writing pedagogy and classroom experience. To a large extent, the evidence collected did indicate that the objectives of the intervention sequence were successful. Members of the experimental group incorporated more and more varied forms of the targeted constructions into their writing at posttest and delayed posttest.

APPENDICES

Appendix A: Participant Language Background Questionnaire

1. Please list the languages that you speak (including your first language) in the order that you first learned/used them. Please indicate how proficient you think you are in those languages by circling the appropriate number from 1-5.

Language		**Beginner**		**Intermediate**		**Fluent**
	Speaking/Listening	1	2	3	4	5
	Writing/Reading	1	2	3	4	5
	Speaking/Listening	1	2	3	4	5
	Writing/Reading	1	2	3	4	5
	Speaking/Listening	1	2	3	4	5
	Writing/Reading	1	2	3	4	5
	Speaking/Listening	1	2	3	4	5
	Writing/Reading	1	2	3	4	5

2. Do you consider your first language to be your strongest language in terms of fluency? If not, which language(s) do you consider to be your strongest language?

yes ____ no ____ strongest language:_____

3. Were you born in the United States? If not, at what age did you arrive in the United States to live or study?

_____ years old

4. What is your nationality? _____

5. How many years have you studied English (in total)? _____ years

6. How may semesters have you studied at the ELC or in another American University English Program?

_____ semesters

7. How many semesters (in any country/school) have you taken a class that focused on *writing* in English?

_____ semesters

Appendix B: Individual Teacher Training and Experience

Group	Section	Master's Degree?	Self-Identified as Native-like proficiency	Years of Language Teaching	Semesters of College-level teaching	Semesters of Writing Instruction	Participants in study
Experimental	1	Yes	Yes	31	70	50	12
	2	Yes	Yes	7	15	15	9
	3	Yes	Yes	6	14	5	15
	4	Yes	Yes	4	3	5	11
Control	5	Yes	Yes	10	11	9	14
	6	Yes	Yes	10	11	9	13
	7	Yes	Yes	10	4	2	10
	8	(In progress)	Yes	2	1	1	9

Appendix C: Summary Nouns Introduced In Intervention Sessions (N=49)

Advance	Event	Problem
Approach	Fact	Procedure
Argument	Factor	Process
Case	Fall	Question
Change	Finding	Reason
Concept	Goal	Reduction
Conclusion	Idea	Relationship
Context	Improvement	Result
Decline	Increase	Rise
Decrease	Information	Rise
Development	Invasion*	Situation
Difference	Jump	Subject
Difficulty	Method	System
Disruption	Pattern	Technique
Diversity	Period	View
Drop	Perspective	
Estimation	Point	

**Invasion* would not likely be considered a true summary noun, as it represents a semantic concept identifiable without context, but it was introduced as part of an exercise highlighting the way that choice of noun can provide additional comment on a topic (i.e., *an increase in students*

vs. *an invasion of students*). It is included here for completeness. No tokens of *invasion* appear in the corpus.

Appendix D: Scaffolded Writing Sheet. Session 3, Section 4

In the past two decades, communication technology such as email, Skype, and social-networking websites, has developed at a very high rate. **<u>These advances in Internet-based systems</u>** have started to make our world feel much closer by improving the way people communicate, that is, they way they connect by interacting, expressing feelings and ideas, and exchanging information. *These new systems allow us to make these connections stronger and more meaningful than ever before.*

Appendix E: Sample Review Cloze Activity

8

Body Pgh. 2

For many people, the idea of getting news about friends and loved ones as if they were celebrities may seem strange, even cold and impersonal. **(1)**_____**,** I believe that **this** _____actually creates more meaningful interactions between people. **(2)**_____ following celebrities, following loved ones involves people we have real relationships with. When we read news updates about the people we know, it is not just a one-way communication. It is _____an invitation to reply or to comment on the events in their lives. **(3)**An update about a new job, _____, might generate messages of congratulations, advice, and encouragement. **(4)**Some friends might be too busy to check online at the moment, but the news will be there waiting when they have time. (5) _____ **these** _____, loved ones maintain contact even when they don't have time to speak directly, and foundations are built for meaningful conversations.

Sample Answers

For many people, the idea of getting news about friends and loved ones as if they were celebrities may seem strange, even cold and impersonal. **(1)<u>However,</u>** I believe that **this <u>process</u>** actually creates more meaningful interactions between people. **(2)<u>While/Unlike</u>** following celebrities, following loved ones involves people we have real relationships with. When we read news updates about the people we know, it is not just a one-way communication. It is **<u>actually</u>** an invitation to reply or to comment on the events in their lives. **(3)**An update about a new job, **<u>for example</u>**, might generate messages of congratulations, advice, and encouragement. **(4)**Some friends might be too busy to check online at the moment, but the news will be there waiting when they have time. (5) **<u>Because of/as a result of/Due to</u> these <u>interactions</u>**, loved ones maintain contact even when they don't have time to speak directly, and foundations are built for meaningful conversations.

Appendix F: Timed Writing Prompts

A: Governments are responsible for providing a variety of services for their citizens. Some governments choose to give some support to artists, including musicians, poets, authors, and painters. Do you think government money should be used to support the arts?

B: Many people in the world lack money, and many people have had a lot of financial success. As some members of society become richer and richer, some argue that they are **too** rich: they are so rich that it is harmful to society. Do you think it is possible for a person to have **too much** money?

C: Is success the result of hard work alone, or is luck also a factor?

D: In many cultures, men and women have often not received equal treatment or opportunity. In many parts of this world, this situation has changed over recent decades, or during the past century, and men and women have been treated more equally. Some people feel that there is still inequality, especially in high-level positions. Do you think that governments should require a percentage of high-level positions be reserved for women?

F: Sometimes, historical events can be described as "turning points"—they represent a major change for a country or a society or people. Choose one such turning point for your country or for another country (for example, the USA): explain how you think it changed that country's history.

Appendix G: Essay Grading Rubric

	Content		**Organization**		**Vocabulary**		**Language Use**	Score /2	**Mechanics**
20 16	Thorough and logical development of thesis Substantive and detailed No irrelevant information Interesting A substantial number of words for amount of time given	20 16	Excellent overall organization Clear thesis statement Substantive introduction and conclusion Excellent use of transition word Excellent connections between paragraphs Unity within every paragraph	20 16	Very sophisticated vocabulary Excellent choice of words with no errors Excellent range of vocabulary Idiomatic and near native-like vocabulary Academic register	20 16	No major errors in word order or complex structures No errors that interfere with comprehension Only occasional errors in morphology Frequent use of complex sentences Excellent sentence variety	20 16	Appropriate layout with indented paragraphs No spelling errors No punctuation errors
15 11	Good and logical development of thesis Fairly substantive and detailed Almost no irrelevant information Somewhat interesting An adequate number of words for the amount of time given	15 11	Good overall organization Clear thesis statement Good introduction and conclusion Good use of transition wordsGood connections between paragraphs Unity within most paragraphs	15 11	Somewhat sophisticated vocabulary Attempts, even if not completely successful, at sophisticated vocabulary Good choice of words with some errors that don't obscure meaning Adequate range of vocabulary but some repetition Approaching academic register	15 11	Occasional errors in awkward order or complex structures Almost no errors that interfere with comprehension Attempts, even if not completely successful, at a variety of complex structures Some errors in morphology Frequent use of complex sentences Good sentence variety	15 11	Appropriate layout with indented paragraphs No more than a few spelling errors in less frequent vocabulary No more than a few punctuation errors

10 6	Some development of thesis Not much substance or detail Some irrelevant information Somewhat uninteresting Limited number of words for the amount of time given	10 6	Some general coherent organization Minimal thesis statement or main idea Minimal introduction and conclusion Occasional use of transitions words Some disjointed connections between paragraphs Some paragraphs may lack unity	10 6	Unsophisticated vocabulary Limited word choice with some errors obscuring meaning Repetitive choice of words No resemblance to academic register	10 6	Errors in word order or complex structures Some errors that interfere with comprehension Frequent errors in morphology Minimal use of complex sentences Little sentence variety	10 6	Appropriate layout with most paragraphs indented Some spelling errors in less frequent and more frequent vocabulary Several punctuation errors
5 0	No development of thesis No substance or details Substantial amount of irrelevant information Completely uninteresting Very few words for the amount of time given	5 0	No coherent organization No thesis statement or main idea No introduction and conclusion No use of transition words Disjointed connections between paragraphs	5 0	Very simple vocabulary Severe errors in word choice that often obscure meaning No variety in word choice No resemblance to academic register	5 0	Serious errors in word order or complex structures Frequent errors that interfere with comprehension Many error in morphology Almost no attempt at complex sentences	5 0	No attempt to arrange essay into paragraphs Several spelling errors even in frequent vocabulary Many punctuation errors

Appendix H: Connectors Included in Corpus Search

according to this
actually
additionally
after all
all in all
also
anyhow
anyway
as a consequence
as a result
at any rate
at first (meaning first)
at least
at the same time
at the same time (temporal)
besides
by contrast
consequently
conversely
despite this
especially
fifth
finally
first
first (temporal)
first of all
first of all (temporal)
firstly
for example
for instance
for that reason
fourth
further
furthermore
hence
however
in addition
in any case
in any event
in brief
in conclusion
in consequence
in contrast
in fact
in fact
In my opinion
in other words
in short
in sum
in summary
in that case
in the meantime
in turn
in turn
initially
instead
last
last (temporal)
lastly
later
like (for example)
likewise
meanwhile

moreover

nevertheless

next

nonetheless

on the contrary

on the other hand

otherwise

overall

rather

second

second (temporal)

secondly

secondly (temporal)

similarly

that is

that is to say

then

then (temporal)

thereby

therefore

third

thirdly

thus

to conclude

to sum up

to summarize

WORKS CITED

WORKS CITED

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on Computer* (pp. 80-93). Harlow: Addison Wesley Longman Limited.

Anthony, L. (2011). AntConc3.2.1w.

Bae, J. (2001). Cohesion and Coherence in Children's Written English: Immersion and English-only Classes. *Issues in Applied Linguistics, 12*(1), 51-88.

Bardovi-Harlig, K. (1990). Pragmatic word-order in English Composition. In U. Connor & A. M. Johns (Eds.), *Coherence in Writing: reserach and pedagogical perspectives* (pp. 43-66). Alexandria, Virginia: Teachers of English to Speakers of other languages, Inc.

Bestgen, Y., Lories, G., & Thewissen, J. (2010). *Using latent semantic analysis to measure coherence in essays by foreign language learners?* Paper presented at the JADT 2010: International Conference on Statistical Analysis of Textual Data.

Biesenbach-Lucas, S., Meloni, C., & Weasenforth, D. (2000). Use of cohesive features in ESL students' e-mail and word-processed texts: A comparative study. *Computer Assisted Language Learning, 13*, 221-237.

Bolton, K., Nelson, G., & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics, 7*(2), 165-182.

Brown, J. D. (2005). *Testing in Language Programs*. New York: McGraw Hill.

Castro, C. D. (2004). Cohesion and the social construction of meaning in the esays of Filipino college students writing in L2 English. *Asia Pacific Education Review, 5*, 215-225.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The Grammar Book: An ESL/EFL Teacher's Course*. United States: Heinle & Heinle Publishers.

Cheng, A. (2011). Language features as the pathways to genre: Student's attention to non-prototypical features and its implications. *Journal of Second Language Writing, 20*(1), 69-82.

Chiang, S. (2003). The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System, 31*(4), 471-484.

Cobb, T. (2011). Web Vocabprofile: an adaptation of Heatley & Nation's (1994) Range Retrieved Feb.-Mar., 2011, from http://www.lextutor.ca/vp/

Connor, U. (1985). A study of cohesion and coherence in English as a second language students' writing. *Papers in Linguistics, 17*(3).

Crossley, S. A., & McNamara, D. M. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*, 119-135.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. M. (2008). *LSA as a measure of second language natural discourse.* . Paper presented at the Proceedings of the 30th Annual Conference of the Coginitive Science Society, Washington, D.C.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 410+ million words, 1990-present [Electronic Version], from Available online at http://www.americancorpus.org

Dennis, S. (2007). How to use the LSA website. In T. K. Landauer, D. M. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 57-69). London: Lawrence Erlbaum Associates Inc.

Enkvist, N. E. (1990). Seven Problems in the study of coherence and interpretability. In U. Connor & A. M. Johns (Eds.), *Coherence in writing: research and pedagogical perspectives* (pp. 9-28). Alexandria, VA: Teachers of English to SPeakers of Other Languages, Inc.

Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28*, 414-420.

Field, A. (2006) *Discovering Statistics Using SPSS*. London: Sage Publications

Flowerdew, J. (2006). Use of signalling nouns in a learner corpus. *International Journal of Corpus Linguistics, 11*, 345-362.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaces to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes, 24*, 321-332.

Foltz, P. W. (2007). Discourse Coherence and LSA. In T. K. Landauer, D. M. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 167-183). London: Lawrence Erlbaum Associates Inc.

Granger, S., & Tyson, S. (1996). Connector Usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes, 15*(1), 17-27.

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing, 9*(2), 123-145.

Green, C. F., Christopher, E. R., & Mei, J. L. K. (2000). The incidence and effects on coherence of marked themes in interlanguage texts: a corpus-based inquiry. *English for specific purposes, 19*(2), 99-113.

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. New York: Longman.

Hasan, R. (1984). Coherence and cohesive harmony. In J. Flood (Ed.), *Understanding Reading Comprehension* (pp. 181-219). Newark: International Reading Association.

Heatley, A., & Nation, P. (1994). Range. from http://www.vuw.ac.nz/lals/

Hinkel, E. (2001). Matters of Cohesion in L2 Texts. *Applied Language Learning, 12*(2), 111-132.

Hinkel, E. (2002). *Second Language Writers' Text: Linquistic and Rhetorical Features*. Mahwah, New Jersey: Lawrence Erlabaum Associates, Inc.

Hinkel, E. (2004). *Teaching Acdemic ESL Writing: Practical Techniques in Vocabulary and Grammar*. Mawah, New Jersey: Lawrence Erlbaum Associates, Inc.

Hooey, M. (1991). *Patterns of Lexis in Text*. New York: Oxford University Press.

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating composition. *System, 19*(4), 459-465.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highl rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Jiminez Catalan, R., & Moreno Espinosa, S. (2003). Lexical cohesion in English L2 students' compositions. In P. Salazar, M. J. Esteve & V. Codina (Eds.), *Teaching and Learning the English Language from a Discourse Perspective* (pp. 73-90). Castello: Universitat Jaime I.

Johns, A. M. (2008). Genre awareness for the novice academic student: An ongoing quest. *Language Teaching, 41*, 237-252.

Khalil, A. (1989). A study of cohesion and coherence. *System, 17*(3), 359-371.

Landauer, T. K., & Dumais, S. T. (1997a). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Landauer, T. K., & Dumais, S. T. (1997b). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T. K., McNamara, D. M., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. London: Lawrence Erlbaum Associates, Inc.

Lee, I. (2002). Teaching coherence to ESL students: a classroom inquiry. *Journal of Second Language Writing, 11*(2), 135-159.

Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chines undergraduates. *System, 33*(4), 623-636.

Lores Sanz, R. (2003). The translation of tourist literature: The case of connectors. *Multilingua, 22*(3), 291-308.

Mahlberg, M. (2006). Lexical cohesion: Corpus linguistic theory and its application in English language teaching. *International Journal of Corpus Linguistics, 11*, 363-383.

Martin, D. I., & Berry, M. W. (2007). Mathmatical foundations behind Latent Semantic Analysis. In T. K. Landauer, D. M. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 35-55). London: Lawrence Erlbaum Associates Inc.

McEnry, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. New York: Taylor & Francis.

McGee, I. (2009). Traversing the lexical cohesion minefield. *ELT Journal, 63*, 212-220.

Milton, J., & Tsang, E. S. C. (1993). A corpus-based study of logical connectors in EFL students' writing: Directions for futrue research. In R. Pemberton & E. S. C. Tsang (Eds.), *Lexis in Studies* (pp. 215-246). Hong Kong: Hong Kong University Press.

Morris, J. (2004). *Readers' interpretations of lexical cohesion in text.* Paper presented at the Conference of the Canadian Association for Information Science, Winnepeg, Manitoba.

Morris, J., & Hirst, G. (2006). The subjectivity of lexical cohesion in text. In J. Shanahan, Y. Qu & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (Vol. 20, pp. 41-47). Netherlands: Springer.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English, 17*, 215-229.

Reynolds, D. (2001). Language in the balance: lexical repetition as a function of topic, cultural background, and writing development. *Language Learning, 51*(3), 437-436.

Reynolds, D. W. (2002). Learning to make things happen in different ways: Causality in the writing of middle-grade English language learners. *Journal of Second Language Writing, 11*(4), 311-328.

Richardson, I. M. (1989). Discourse Structure and comprehension. *System, 17*(3), 229-245.

Shea, M. (2009). A corpus-based study of adverbial connectors in learner texts. *MSU Working Papers in SLS.* http://sls.msu.edu/soslap/journal/index.php/sls/article/view/4

Shea, M. (2011) Syntactic complexity: Clause or phrase? Paper presented at *AAAL 2011*: Chicago

Salkie, R. (1995). *Text and Discourse Analysis*. New York: Routledge.

Watson Todd, R., Khongput, S., & Darasawang, P. (2007). Coherence, Cohesion and comments on students' academic essays. *Assessing Writing, 12*(1), 10-25.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teching assistants. *Tesol Quarterly, 26*, 693-711.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and development* (Technical Report #17). Manoa: University of Hawai'i at Manoa, Second Language Teaching and Curriculum Center

Yoon, H., & Hirvala, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing, 13*(4), 257-283.