

**The Effects of Different Types of Unfocused Corrective Feedback on Complexity,
Accuracy and Fluency in L2 English Academic Writing**

Laurence Craven

University of Essex

A Thesis Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Table of Contents

List of Tables	5
List of Figures	7
List of Abbreviations.....	8
Abstract.....	9
Chapter 1: Introduction	10
1.1 Aims and Significance of the Study.....	10
1.2 Overview of the Research Design and Context	12
1.3 Outline of the Thesis	13
Chapter 2: Literature Review	15
2.1 Written Corrective Feedback	15
2.1.1 Empirical Studies of Corrective Feedback – Text Revisions and New Tasks.....	17
2.1.2 The Different Types of Written Corrective Feedback	18
2.1.3 Focused and Unfocused Corrective Feedback and a Review of Studies on Unfocused Corrective Feedback	24
2.2 Complexity, Accuracy and Fluency	32
2.2.1 Definitions and Operationalisation of Complexity	33
2.2.2 Definitions and Operationalisation of Accuracy.....	35
2.2.3 Definitions and Operationalisation of Fluency	36
2.2.4 Issues with the Definitions and Operationalisation of CAF.....	37
2.2.5 Skehan’s Trade-off Hypothesis.....	37
2.2.6 Robinson’s Cognition Hypothesis	39
2.2.7 Complexity and Dynamic Systems Theory.....	40
2.2.8 Relationship of CAF with Other Variables	41
2.2.9 Future Direction of CAF Research	42
2.3 Individual Differences.....	43
2.3.1 L2 Proficiency.....	43
2.3.1.1 Definitions of Proficiency.....	44
2.3.1.2 The Main Proficiency Frameworks, Scales and Tests	44
2.3.1.3 L2 Proficiency and Written Corrective Feedback.....	45

2.3.2 Language Learning Aptitude: Definitions and Operationalisation	46
2.3.3 Attitudes and Beliefs	58
2.4 Conclusion and Gaps in Existing Research.....	64
Chapter 3: Methodology	67
3.1 Introduction	67
3.2 Research Questions	67
3.3 Ethical Approval	68
3.4 Setting	69
3.5 Participants	69
3.6 Experimental Treatment.....	70
3.7 Instruments	73
3.7.1 Essay Topic Questionnaire.....	73
3.7.2 Language Background Questionnaire.....	74
3.7.3 Attitudes Questionnaire.....	74
3.7.4 The Oxford Quick Placement Test.....	75
3.7.5 The LLAMA Aptitude Test	75
3.8 Pilot Study	78
3.8.1 Pilot Study Methodology and Experimental Set-up.....	79
3.8.2 Results of the Pilot Study that Informed Changes to the Main Study.....	83
3.8.3 Conclusions of Pilot Study.....	88
3.9 Complexity, Accuracy and Fluency Measures.....	89
3.9.1 Fluency.....	93
3.9.2 Accuracy	93
3.9.3 Complexity.....	94
3.10 Procedure.....	96
3.11 Inter-rater Reliability Analysis and Data Processing	100
3.12 Preliminary Data Checks.....	101
3.12.1 Principal Component Analysis.....	105
3.12.2 Data Transformation and Normalisation.....	109

3.13 Data Analysis	110
3.14 Summary	111
Chapter 4: Results	112
4.1 Introduction and Overview.....	112
4.2 Descriptive Results for the CAF Variables	114
4.3 Descriptive Results for the ID Variables.....	122
4.5 Effects of Different Types of Unfocused Feedback on CAF in Revised and New Texts	128
4.6 Relationships among the CAF Variables and Relationships between CAF Variables and ID Variables	144
4.7 Summary of Results	150
Chapter 5: Discussion	153
5.1 Connected Growers: The Interaction of Gains in CAF	155
5.2 The Interaction of IDs and CAF	157
5.2.1 L2 Proficiency and Gains in CAF	157
5.2.2 Aptitude and Gains in CAF	159
5.2.3 Attitudes and Gains in CAF	160
5.3 The Effects of WCF Generally on Text Revisions	162
5.4 The Effects of Different Types of WCF on Text Revisions.	165
5.5 The Effects of WCF Generally and the Effects of Different Types of WCF on New Texts.	166
5.6 General Considerations Regarding Unfocused Feedback	171
Chapter 6: Conclusions, Theoretical and Pedagogical Implications.....	174
6.1 Summary of Findings	174
6.2 Theoretical Implications	176
6.3 Pedagogical Implications	177
6.3.1 Connected Growers.....	178
6.3.2 Correlations between Attitudes and Corrective Feedback.....	178
6.3.3 The Effects of the Different Types of WCF on Text Revisions and New Texts	179
6.3.4 The Effects of WCF Generally on Text Revisions and New Texts	180
6.4 Limitations of the Study and Suggestions for Further Research.....	183

References.....	187
Appendix A: Metalinguistic Error Codes Used	211
Appendix B : Essay Topic Questionnaire	212
Appendix C: Language Background Questionnaire	213
Appendix D: Attitudes Questionnaire.....	216
Appendix E: Correlations of CAF Measures and PCA.....	220
Appendix F: Tests of Normality	227
Appendix G: Kruskal-Wallis Test.....	232
Appendix H: Friedman Test for Effect of Corrective Feedback on Complexity Measures.....	233
Appendix I: Kolmogorov-Smirnov Test.....	235
Appendix J: Tests of Normality.....	238
Appendix K: Descriptives Statistics by test for the Composite Variables.....	241
Appendix L : MANCOVA Re-test to Post-test Gains	250
Appendix M: MANCOVA Pre-test to Re-test Gains.....	254
Appendix N: MANCOVA Pre-test to Post-test Gains.....	260
Appendix O: Pearson Correlations Pre-test to Post-test	266

List of Tables

Table 1. Different Types of Written Corrective Feedback Used in Key Studies	19
Table 2. Studies Comparing Direct and Indirect Corrective Feedback and Metalinguistic Feedback	22
Table 3. Descriptive Statistics of the Oxford Quick Placement Test	70
Table 4. Experimental Set-up and Instruments.....	80
Table 5. CAF Measures, their Operationalisation and Other Studies That Have Used the Measure.....	82
Table 6. Descriptive Statistics for the LLAMA B and F.....	84
Table 7. Descriptive Statistics for Fluency.....	86
Table 8. Descriptive Statistics for Accuracy.....	86
Table 9. Descriptive Statistics for Syntactic Complexity Overall.....	87
Table 10. Descriptive Statistics for Lexical Diversity.....	87
Table 11. CAF Measures, their Operationalisation, the Way They Were Calculated and Examples from the Data Set	90
Table 12. Timeline of the Experiment	97
Table 13. Correlations: (Spearman) Pre-test CAF Measures and Proficiency	103
Table 14. Principal Component Analysis 5 Factor Solution.....	106
Table 15. Rotated Component Matrix	108
Table 16. Descriptive Statistics of the Covariates for the Whole Sample.....	122
Table 17. Kruskal-Wallis Test: Language Proficiency, Aptitude and Attitude Measures...	127
Table 18. Box's Test of Equality of Covariance Matrices	129
Table 19. Levene's Test of Equality of Error Variances.....	Error! Bookmark not defined.
Table 20. Box's Test of Equality of Covariance Matrices	Error! Bookmark not defined.
Table 21. Levene's Test of Equality of Error Variances ^a	131
Table 22. MANCOVA Pre-test to Re-test	133
Table 23. Tests of Between-Subjects Effects	133
Table 24. Pairwise Comparisons.....	135
Table 25. Summary of Significant Differences 1	138
Table 26. Summary of Significant Differences 2	139
Table 27. Multivariate Tests	141
Table 28. Between-subjects Effects.....	143
Table 29. Correlations Pre-test to Re-test Gains.....	145

Table 30. Correlations Re-test to Post-test Gains	147
Table 31. Correlations Pre-test to Post-test Gains	266
Table 32. Correlations of CAF Measures at Pre-test	220
Table 33. Principal Component Analysis 4 Factor Solution.....	222
Table 34. Rotated Component Matrix ^a	225
Table 35. Tests of Normality	227
Table 36. Kruskal-Wallis Test	232
Table 37. Friedman Test	233
Table 38. Kolmogorov-Smirnov	235
Table 39. Tests of Normality	238
Table 40. Tests of Normality	239
Table 41. Tests of Normality	240
Table 42. Box's Test of Equality of Covariance Matrices ^a	250
Table 43. Levene's Test of Equality of Error Variances.....	251
Table 44. Multivariate Tests	252

List of Figures

Figure 1. LLAMA_B User Interface	77
Figure 2. The LLAMA_F User Interface	78
Figure 3. Scree Plot.....	107
Figure 4. Box Plots of Fluency Gain Scores Pre-test to Re-test (revised task)	115
Figure 5. Box Plots of Fluency Gain Scores Pre-test to Post-test (new task).....	115
Figure 6. Box Plots of Accuracy Gain Scores for Pre-test to Re-test (revised task)	116
Figure 7. Box Plots of Accuracy Gain Scores for Pre-test to Post-test (new task).....	117
Figure 8. Box Plots for Complexity Gains Pre-test to Re-test (revised task)	118
Figure 9. Box Plots for Complexity Gains Pre-test to Post-test (new task).....	118
Figure 10. Box Plots for Complex Nominals per Clause Gains Pre-test to Re-test (revised task).....	120
Figure 11. Box Plots For Complex Nominals per Clause Gains Pre-test to Post-test (new task)	121
Figure 12. Box Plots for Lexical Diversity Gains Pre-test to Re-test (revised tasks).....	121
Figure 13. Box Plots for Lexical Diversity Gains Pre-test to Post-test (new task).....	122
Figure 14. Box Plots for the Oxford Quick Placement Test by Group.....	123
Figure 15. Box Plots for the LLAMA_F Aptitude Test by Group	124
Figure 16. Box Plots for the LLAMA_B Aptitude Test by Group.....	125
Figure 17. Box Plots for the Attitudes by Group.....	126
Figure 18. Estimated Marginal Means of Fluency Gains Pre-test to Re-test.....	137
Figure 19. Estimated Marginal Means of Lexical Diversity Gains Pre-test to Re-test.....	137
Figure 20. Scree Plot.....	224

List of Abbreviations

ACTFL	The American Council on the Teaching of Foreign Languages
BALLI	Beliefs About Language Learning Inventory
CAF	Complexity, Accuracy, Fluency
CANAL-F	Cognitive Ability for the -Novelty in Acquisition of Language – Foreign
CEFR	Common European Framework of Reference
CF	Corrective Feedback
ESL	English as a Second Language
ID	Individual Difference Variables
ILR	Interagency Language Roundtable Scale
L2	Second Language
MLAT	Modern Language Aptitude Test
MTLD	Measures of Textual Lexical Diversity
SLA	Second Language Acquisition
TTR	Type-token Ratio
UAE	United Arab Emirates
WCF	Written Corrective Feedback

Abstract

Research on written corrective feedback in second language (L2) learning has made progress, answering the unknowns regarding its effectiveness. Currently, debate focuses on the most effective way of giving feedback. Controversy, however, remains and there is a scarcity of research on unfocused feedback. The present study examines the effects of unfocused direct, indirect and metalinguistic written corrective feedback (WCF) on the complexity, accuracy and fluency (CAF) of 139 L1 Arabic or Urdu – L2 English students' writing. The study also investigates if the moderating variables of aptitude, attitudes and proficiency affect the uptake of feedback. Students in four intact groups were designated as feedback groups, plus one control group. They wrote argument essays and were given four rounds of feedback and feedback support sessions over fourteen weeks; whereas learners in the control group received no feedback or support sessions. Students wrote both text revisions and new texts. Results showed that on text revisions, the direct and metalinguistic feedback groups had losses in fluency compared to the indirect and control groups. The indirect feedback group had significantly lower lexical diversity than the direct and metalinguistic groups. On new texts, there were no significant gains or losses from the unfocused feedback. The moderating variables of proficiency and aptitude had no significant relationships with CAF gains or losses, but positive attitudes towards feedback had a negative relationship with gains in complexity and fluency on text revisions. These results reveal that on text revisions, some forms of unfocused feedback have effects on fluency and lexical diversity, but on new texts there are no effects. Future work should examine if increasing the number of treatment sessions has positive effects on CAF, and discover at what point unfocused WCF may become too cognitively demanding. The results provide useful information for practitioners who could use a more blended approach between focused and unfocused WCF and increase the treatment sessions.

Chapter 1: Introduction

During the past few decades, approaches and methods to teaching English have been constantly evolving. The teaching of academic writing to English second language students has also continued to progress. However, with all the changes made to the ways and methods of teaching writing, there is one constant: the inclusion of written feedback.

1.1 Aims and Significance of the Study

Written corrective feedback (WCF) is also known as error correction, and researchers have examined its effects on language learning and development to help students improve their writing. There are three major forms of WCF: direct WCF which involves providing students with the correct form; indirect WCF, which is when the teacher underlines or circles the error without providing the correction; and metalinguistic WCF where the teacher will give the students metalinguistic codes showing them the type of error (Ellis, 2009). Many believed that WCF helps students to improve their writing and language development, until Truscott (1996) published his critique of the practice. This led to an ongoing debate about the effectiveness of corrective feedback on English second language (L2) students' writing. Truscott mentioned several theoretical and practical reasons why WCF should be abandoned and that it could be a harmful practice. He explained how a simple information transfer through WCF could not be effective, since the language learning process is gradual and not a linear transfer from teacher to student. He further pointed out that the WCF given may not correspond well with the learners' developmental stage, and stated that any learning which came from WCF was likely due to "pseudo-learning" (Truscott, 1996, p.345). Truscott's theoretical arguments, however, have been rebutted by researchers such as Swain (1985), Schmidt, (1990), Ferris (1999), Ashwell (2000), Ellis (2010), and Bitchener and Ferris (2012). Empirical studies have also attempted to

explore the benefits of WCF, and many have also rebutted Truscott's (1996) claims. The following studies demonstrated through empirical evidence that WCF was effective by improving accuracy in revisions, and some in new writing tasks: Ferris (1999, 2004), Chandler (2003), Sachs and Polio (2007), Sheen (2007), Bitchener (2008), Ellis et al. (2008), Van Beuningen et al. (2008, 2012), Bitchener and Knoch (2009), and Ferris and Roberts (2001).

In general, WCF studies have made progress in answering questions surrounding the effectiveness of WCF, and thus the focus of recent research has shifted to how WCF can be utilised for optimal results (Bitchener, 2019). Many studies have investigated the effectiveness of focused WCF, which refers to feedback that is given only on a specific and preselected type of error; however, fewer studies have researched unfocused WCF, which is WCF that targets a wide range of errors (Van Beuningen, et al., 2012). Unfocused WCF is also known as comprehensive WCF (Falhasiri, 2021); however, the term unfocused WCF is more predominant in research and thus will be used throughout the thesis.

When the effects of corrective feedback on students' written performance are analysed, researchers sometimes look at the variation in complexity, accuracy, and fluency of student output. These measures are also known as CAF measures. CAF research is of great value to L2 researchers as they argue that the principal dimensions of L2 performance can be captured by the notions of complexity, accuracy, and fluency (Ellis & Barkhuizen, 2005).

Most WCF studies, for practical reasons, look at group averages without taking individual student profiles into account. Thus, recently there have been calls for studies to be conducted on the relationship between individual difference (ID) variables, such as proficiency, attitudes toward WCF, aptitude and how they moderate the uptake of WCF (Li, 2015); (Kang & Han, 2015). These studies, similar to the studies on unfocused WCF, are also scarce.

There have been calls in the literature to improve research on WCF and make it more applicable to what actually happens in the classroom, for example, by comparing independent written works

instead of comparing an initial text to a text revision, and thus examining if longer-term language development is taking place. However, the largest lack of research is in the area of unfocused feedback and research in this area is of importance, since unfocused feedback is the form of WCF that most writing teachers give to their students (Ellis et al. 2008) and thus this study addresses these shortcomings.

This research study examines unfocused direct, indirect and metalinguistic WCF and sees if it leads to improvements in learners' overall written complexity, accuracy, and fluency. It also examines the way in which proficiency, attitudes toward WCF and aptitude moderate the uptake of WCF. Since WCF is provided for the benefit of the learners, understanding their attitudes as well as their preferences, is important to obtain a clearer idea about the effectiveness of WCF and its effects on CAF. It is also clear there is a relationship between aptitude and learning and that different aptitude components demonstrate differential predictive validity for various aspects of learning. Explicit forms of WCF work more favourably when the learners have high language analytic ability. When WCF targets a single linguistic feature, this improves learners' accuracy, especially when the learners have high language analytic ability (Benson & DeKeyser, 2018). However, with unfocused WCF, the relationship with aptitude is not clear and this study attempts to shed light on it. The research on proficiency and WCF is still in its infancy, but the consensus is that it has a moderating effect on the efficacy of different feedback types. This study will attempt to shed more light on these issues.

1.2 Overview of the Research Design and Context

The research was conducted at a co-educational private university in the emirate of Sharjah in the United Arab Emirates (UAE). The government of the UAE sees English as important for the economy and thus although Arabic language universities exist, the majority of students attend an English medium university. Tertiary education in the UAE consists of government universities and private universities. The university where the research was conducted is a private American accredited

university, with an American liberal arts curriculum. Freshmen students upon entering the university are required to take an academic writing course. The rationale for this is they must write a research paper for the major before graduation, and the academic writing course prepares them for this.

The present study examines the effects of different types of WCF on academic writing students' performance. The participants came from the academic writing classes for freshmen. The academic writing classes were chosen, as the students had sufficient English writing proficiency to produce the length and level of writings required, but would also have errors in their compositions and thus feedback could be given on errors they made. In total, 139 English academic writing students participated voluntarily in the study. Students in four intact groups were designated as feedback groups comprising of a direct, indirect, and metalinguistic feedback group, together with a control group. Students were then instructed to write argument essays and were given four rounds of feedback and feedback support sessions over fourteen weeks, while learners in the control group received no feedback. Students wrote text revisions and new texts in order to see if WCF had effects on students' performance and to discover the way in which proficiency, attitudes toward WCF, and aptitude, moderate the uptake of WCF.

1.3 Outline of the Thesis

This thesis is structured as follows. Chapter 2 presents the review of the literature on the effectiveness of WCF in L2 writing, the literature on CAF and ID variables. The chapter then concludes with a discussion of the rationale for the current study, and then presents the research questions. The third chapter presents the methodology used in the present research study. Chapter 4 reports the results and findings for the research questions, and chapter 5 explores the meaning of the results by way of the discussion. Chapter 6 presents the conclusions, the pedagogical implications of the findings and then

ends with a discussion on the limitations of the present study while also providing directions for further research.

Chapter 2: Literature Review

2.1 Written Corrective Feedback

To aid practitioners with guidance for optimum ways to teach L2 writing, to contribute to our understanding of how L2 learning works and to develop theories of L2 learning, many researchers over the past few decades have studied the effects of error correction for both oral and written language. Error correction on written language is also referred to as written corrective feedback (WCF). Many believed that WCF helps students to improve their writing and their language development, until Truscott (1996) published his critique of the practice. Since then, there was an ongoing debate about the effectiveness of corrective feedback on L2 students' writing. Truscott mentioned several reasons why WCF should be abandoned and that it could be a harmful practice. First, Truscott (1996) argued that a simple information transfer through WCF could not possibly be effective in the acquisition of knowledge since the language learning process is a gradual and complex process and not a linear transfer from teacher to student. Truscott's second theoretical argument against the use of WCF on L2 acquisition was based on whether the WCF given would correspond well with the learners' developmental stage. He argued that the natural order of learning must be followed and that learners would be unable to correct the errors that they are not developmentally ready for. The third theoretical argument Truscott (1996, p. 345) provided was that any learning which came from WCF was likely due to "pseudo-learning". He explains this by showing that the improvements the students may make initially, could disappear within a few months. Although at first, students may seem to have acquired the target language, it could in fact be superficial learning. Truscott (1996) also presented some practical arguments for the abandonment of WCF relating to the ability of teachers to find the errors on student papers, and the inability of learners to understand the feedback.

Truscott's theoretical arguments, however, have been rebutted by some researchers. Truscott (1996) argued that a simple information transfer through WCF could not be effective in the acquisition of knowledge according to the Natural Order Hypothesis (Krashen, 1985), and Pienemann's (1989) Teachability or Learnability hypothesis. This is because they believe that the language learning process is a gradual and complex process and not just a linear transfer. However, according to skill acquisition theories, WCF can help due to the declarative knowledge provided by WCF being automatised to the point where it becomes procedural knowledge (Bitchener & Ferris, 2012). Another theoretical argument Truscott (1996, p. 345) provided was that any learning which came from WCF was due to what he called "pseudo-learning", whereby explicit knowledge, in this case WCF, will not become implicit. However, many second language acquisition (SLA) researchers argue that there is an interface connecting implicit and explicit knowledge bases (Schmidt, 1990), and they argue that the difference between explicit knowledge and language use can gradually be connected by output practice (DeKeyser, 2003). Furthermore, Ellis (2010) argues that WCF can further assist this proceduralisation of declarative L2 knowledge.

Ferris (1999) responded to Truscott's claims and questioned them, because the evidence Truscott proposed had methodological problems both in design and analysis. The first issue with Truscott's argument is that he defines error correction vaguely and Ferris notes that when "discussing whether or not grammar correction is effective, it is crucially important to know what sort of error correction we are discussing" (1999, p.4). The second issue Ferris notes is regarding the review section of Truscott's paper. She notes that the L2 correction studies Truscott cites used participants who are not comparable and vary widely, a point that was also acknowledged by Truscott himself (1996). Furthermore, the research paradigms and teaching strategies differ widely across the studies, for example some studies covered a semester while some were a 'one shot' experimental treatment. Some studies lacked control groups and the mechanisms used for giving feedback varied (Ferris, 1999). Ferris also states that Truscott overstates research findings that support his thesis, but

dismisses studies that contradict him. In 1999, Truscott admitted that further research should investigate which approaches to error correction may have value. Truscott (1999) and Ferris (1999) both noted, however, that research should now be focused on investigating the long term effects of WCF on new texts, rather than initial so called pseudo-learning (Truscott, 1996). This recommendation from both researchers resulted in a large number of studies attempting to explore the benefits of WCF and many rebutted Truscott's (1996) claims. These studies include: Ferris (1999, 2004), Ashwell (2000), Ferris & Roberts (2001), Chandler (2003), Sachs & Polio (2007), Sheen (2007), Bitchener (2008), Bitchener and Knoch (2008, 2009) Ellis et al. (2008), Van Beuningen et al. (2008, 2012). These studies demonstrated with empirical evidence that WCF was effective by improving accuracy in revisions, and some in new writing tasks.

2.1.1 Empirical Studies of Corrective Feedback – Text Revisions and New Tasks

Among the multitude of studies on WCF, there are two main different types. The first examines the effects of CF as an editing tool, and the second examines the effects of corrective feedback on new pieces of writing. The studies on text revisions, in general, demonstrate that language students are able to improve the accuracy of a particular piece of writing, based on the feedback provided. Studies on text revisions include Semke (1984), Robb et al. (1996), Polio et al. (1998), Ashwell (2000), and Ferris & Roberts (2001). All studies found an improvement in accuracy except for that by Polio et al. (1998) where there was no improvement, and the study of Semke (1984) where the results were unclear.

Truscott and Hsu (2008) argued that the results of studies on text revisions should not be considered to be proof of genuine learning. They point out that students' improved performance on text revisions does not demonstrate that they have internalised the forms, or if they would be able to apply the corrected target forms in a new piece of writing. In their study, Truscott and Hsu (2008) revealed that the CF group significantly outperformed the control group one week later on a text

revision task, but when the two groups wrote a new piece of writing, both groups performed with the same results. Van Beuningen et al. (2012) note that more interesting research consists of studies that have investigated the effects of WCF on new pieces of writing, as they will show if genuine language development is taking place, rather than immediate effects that do not transfer to new tasks.

A significant addition to the literature was Kang and Han's (2015) meta-analysis that analysed whether or not WCF is effective in the development of L2 learners' written accuracy. Their meta-analysis yielded an overall effect size of $g = .54$ and according to Cohen (1988), this is a moderate effect. This shows that WCF does have an effect on L2 written accuracy on new tasks.

Most recent studies of WCF that attempt to find if language development is taking place, have avoided the research design flaws of the early studies. One of the recurring problems of the early studies was the lack of a control group. They examined the improvement of learners receiving feedback, but did not compare the results with students who did not receive any feedback. Recent studies have examined not only the effectiveness of feedback in general, but also whether these effects differ across different types of feedback; the most common being direct, indirect, and metalinguistic feedback.

2.1.2 The Different Types of Written Corrective Feedback

The effectiveness of WCF has been debated in the literature due to the results yielded from studies being varied, and the debate surrounding whether WCF is effective or not has mostly moved on to analysing which type is the most effective. Table 1 has been adapted from Ellis (2009) and shows the different types of WCF used in several key studies. The table includes a short description of the type of feedback, some key studies as well as whether or not the feedback was effective. The studies in the table include those that focused on text revisions and those that focused on new texts. The contents of the table will be elaborated on in the following paragraphs.

Table 1. Different Types of Written Corrective Feedback Used in Key Studies

Type of CF	Description	Studies	Effective?	Type of Text
Direct CF	The teacher provides the student with the correct form.	Lalande (1982) Robb et al. (1986) Polio et al. (1998)	Yes No No	Text re-write Text re-write Text re-write
Indirect CF a. Indicating + locating the error b. Indicating only	The teacher provides an indication that an error exists, but does not provide the correction.	Chandler (2003)	Yes	New task
Metalinguistic CF (use of error code, or brief grammatical descriptions)	The teacher provides a metalinguistic clue of the nature of the error.	Chandler (2003) Bitchener (2005) Sheen (2007)	Yes	New task New task Both

With direct WCF, the teacher provides the student with the correct form. This feedback can take a number of different forms, for example correcting a phrase or a word by writing the correct form near the erroneous form (Ellis, 2009), and advising on how to correct the error, writing the missing word that a student had forgotten to include and changing words to more appropriate options. Many argue that direct WCF requires minimal processing on the part of the learners and thus, it may not contribute to long-term learning. The benefit of direct feedback is that students do not become confused and know exactly what error they made and how to correct it. Alimohammadi and Nejadansari (2014) state that direct WCF is more immediate and more explicit.

Unlike direct feedback, indirect WCF involves making the learner aware of an error, but not providing the corrected form. Within the indirect feedback category, it is important to note that there are several subcategories that depend on how explicitly the error type and location are indicated to the learner. Indirect WCF can be provided by underlining errors or using marks in the margins to show that there are errors in the student's text. Alternative forms include circling the error or placing a cross in the margin next to the line containing the error. If errors are marked in the margins, the person giving the WCF can decide to show where the exact location of the error is, or just to indicate that the line has an error. Some language acquisition theorists and researchers argue that indirect feedback is preferable for most student writers, as it engages them in problem solving which causes them to think deeply about the error (Lalande 1982; Storch & Wigglesworth, 2010; Ferris et al., 2013).

Another form of WCF available to practitioners is metalinguistic WCF, which is when learners are provided with an explicit comment regarding the type of the error they have made that explains the error. This can take two forms: error codes which are an abbreviated label for different kinds of errors, or metalinguistic explanations of errors. There are strengths and weaknesses of this approach to feedback. First, Ellis (2009) notes that possibly due to them being time consuming, metalinguistic explanations of errors are used less often than error codes. They can also cause other issues for students, such as confusion over the meaning of the labels and writing in the correct form of the error.

The strengths of metalinguistic feedback are that it trains students to become independent learners (Ellis, 2009), and Bitchener (2012) argues that metalinguistic feedback is noticeable to L2 learners since it explicitly provides them with the opportunity to diagnose the errors they made. This saliency then enhances the strength of the corrective function of metalinguistic feedback (Bitchener, 2012).

2.1.2.1 The Effectiveness of the Different Types of Corrective Feedback

Despite the evidence in favour of providing written corrective feedback in order to improve students' accuracy, there also remains doubt about which type of feedback may be the most effective (Storch & Wigglesworth, 2010; Bitchener, 2012). Table 2 represents a collection of some of the most important studies often cited in the literature on WCF that compares the effectiveness of direct, indirect, and metalinguistic forms of WCF.

Table 2. Studies Comparing Direct and Indirect Corrective Feedback and Metalinguistic Feedback

Study	Participants	Operalisation and Definition	Findings
Semke (1984) (text revision)	141 college-level students, studying German as a foreign language at an American university	<i>4 groups</i> 1.Direct error correction 2. Content comments 3.Direct error correction and content comments 4.Indirect coding	No difference
Robb et al. (1986) (text revision)	134 Japanese EFL learners	1.Direct error correction 2.Indirect feedback 3.Highlighting 4.Marginal error totals	No difference
Van Beuningen et al. (2008)	62 Dutch multilingual secondary school learners	1.Direct error correction 2.Indirect feedback 3. Control group (writing practice) 4.Control group (self-correction revision)	Long-term effect of direct error correction is stronger than the other types. In the short-term, direct and indirect feedback are both effective.
Van Beuningen et al. (2012)	62 Dutch multilingual secondary school learners	1.Direct error correction 2.Indirect feedback 3.Control group (writing practice) 4.Control group (self-correction revision)	Direct corrective feedback is effective for improved grammatical accuracy and indirect feedback is better for non-grammatical accuracy
Eslami (2014)	60 EFL students in Iran	1. Direct error correction 2. Indirect feedback	Indirect was more effective
Gholaminia et al. (2014)	60 Iranian ESL students	1.Direct error correction 2.Metalinguistic error codes	Metalinguistic was more effective
Suzuki et al. (2019)	88 Japanese ESL students	Four groups: direct corrective feedback with metalinguistic explanation; direct corrective	All types of WCF had a positive effect

Study	Participants	Operalisation and Definition	Findings
		feedback only; indirect corrective feedback with metalinguistic explanation; and indirect corrective feedback only. Target structures: English indefinite article and the past perfect tense	

Overall, some studies have concluded that there is no difference between direct and indirect feedback, some researchers found direct WCF more effective in their studies, and others have found indirect WCF the most effective and others metalinguistic. Kim and Bowles (2019) most recently stated that there may not be a clear answer for the most effective type of feedback, and it can also depend on type of error, for example sentential and paragraph-level errors such as sentence structure and organisation or surface-level errors such as punctuation.

Elsami (2014) also notes that the results of studies investigating the difference between direct and indirect WCF are mixed. Kang & Han's (2015) conclusion from their meta-analysis on the effects of WCF on L2 written accuracy did not find a significant difference between direct and indirect feedback. They note that one possible reason could be due to the multiple types of feedback used and the dissimilarities among the students, making it difficult to conclude on the effectiveness of the feedback. Hyland and Hyland (2006) point out that intervening factors, such metalinguistic awareness or a learner's level of L2 proficiency, can affect the effectiveness of WCF. They also perceive that other individual difference variables - such as aptitude and attitudes towards WCF - may also moderate the uptake of WCF (Bitchener & Storch, 2016). As well as the distinction between direct and indirect WCF, another difference in which WCF can be given is whether it is given on all errors or only selected errors.

2.1.3 Focused and Unfocused Corrective Feedback and a Review of Studies on Unfocused Corrective Feedback

A question many would like to have answered is if WCF should be selective, or address different types of errors at the same time. This distinction is called 'unfocused' and 'focused' WCF. Focused feedback refers to feedback that is given only on a specific and preselected type of error. An example would be CF provided only on errors displaying incorrect use of the past tense. Unfocused feedback refers to feedback that is given on all or a range of error types. Van Beuningen et al. (2012) note that there are downfalls to using unfocused WCF and it is possible that students will receive a large

amount of correction on a long piece of writing and thus may not be able to check all their errors. Ellis et al. (2008) and Sheen et al. (2009) also argue that learners might be able to notice and acquire the form when they receive WCF on only one targeted feature, since they have limited processing capacity. Sheen et al. (2009) point out that a focused approach may enhance learning due to noticing of errors and the monitoring of the accuracy of writing by students - by tapping into their existing explicit grammatical knowledge.

Ellis et al. (2008) note that many existing studies on the effectiveness of WCF have investigated the effects of focused WCF as opposed to unfocused corrective feedback. Most of the studies on WCF produced over the past ten years have transitioned away from unfocused WCF and turned their attention to focused WCF, possibly because of its clarity. Ferris (2010) claims that researchers have conducted more studies of focused feedback, since it is easy to control and not because it is more effective than unfocused corrective feedback. Following this, Xu (2009) and Van Beuningen (2010) have called for more studies to investigate unfocused WCF. Van Beuningen et al., (2012) further argue that improving students' written accuracy in general, and not only one or two grammatical features, should be one of the goals of error correction. Xu (2009) also argues that focused feedback studies yield results in quasi-experimental research designs due to the narrowing of students' attention to a specific grammatical issue, and that giving feedback using focused WCF is limited. Its findings also reflect a limited aspect of L2 writing ability.

Little research has been conducted on the effects of unfocused WCF. In general, findings suggest that WCF works well when it is focused and concentrated on a specific linguistic error. However, unfocused WCF has the advantage of addressing a range of errors, so although it may not be as effective in helping learners to acquire specific features in the short term, in the long term it might be more beneficial to second language students' writing development. Furthermore, unfocused feedback has greater ecological validity (Kang & Han, 2015), since it is typically the type of feedback given to ESL students in writing classes the world over.

The studies of Kepner (1991), Robb et al., (1986), Semke (1984) and Sheppard (1992) have investigated unfocused WCF, and found no statistically significant benefits of unfocused WCF on accuracy, compared to the control content group. These studies, however, have been criticised because the studies of Kepner (1991) and Robb et al. (1986) did not include a non-feedback control group, and Semke's (1984) study used students' journals to provide WCF. Ferris (2003) among others, has argued that journals are unlikely to motivate students to pay attention to grammatical accuracy as the purpose of writing a journal is usually to encourage fluency. Another study on unfocused feedback by Gee (1972) looked at L1 English-speaking students who were divided into three groups. The groups either received praise; no comments; or criticism on grammatical errors, content and style. Gee's results showed that the negative criticism and no-comments groups both wrote less than the group receiving praise; however, whether negative criticism can be regarded as WCF is debatable. In Sheppard's (1992) study, the students' writing was conducted at home and thus the time spent on the task and whether any outside help was available is impossible to determine. The following unfocused WCF studies, to the researchers' knowledge, are the only ones without the design and execution errors of the earlier studies.

Truscott and Hsu's (2008) study of unfocused WCF examines forty-seven graduate students' compositions from a university in Taiwan. The data was collected from an in-class writing assignment conducted during weeks twelve to fourteen of the study. The researchers split the students into two groups; the treatment group received unfocused WCF in the form of indirect feedback using underlining of errors, and the other group was a control group with no feedback. Both groups were given a guided narrative to write, based on pictures given to them, and 30 minutes to write it. In week 13, students' narratives were returned with indirect feedback for the treatment group, and no feedback for the control group. They were given 30 minutes to revise their narratives, and in week 14, students again wrote a new second guided narrative. After the errors were marked, each piece of writing was assigned an error rate, which was measured as the total number of errors divided by the total number of words written. On the revision, the students who received indirect WCF performed better than the

control group; however, the two groups were virtually identical regarding the error rate on the post-test (the new second narrative). The results of the study indicated that unfocused corrective feedback was only effective for the revision, but not on the new guided narrative. Truscott and Hsu (2008) argue that this indicates the feedback did not have a significant effect on students' writing development. They note that with the students who received corrections on their drafts, it did not seem to influence their writing performance on the next assignment. They further elaborate that correction does help students reduce their errors on the editing of their writing, and that the effect on improving the editing was substantial, but that the benefits of feedback on the revision task were not found on the new writing task that was completed a week later. Furthermore, they state that no relation was found between success on the revision task and learning, if measured by performance on a new text. They conclude that error reduction during revision is not a predictor of learning, and the gains on the revision do not transfer to learning (2008). The limitation of the Truscott and Hsu (2008) study was that only one form of feedback was investigated.

Van Beuningen et al. (2008) carried out a three-week study that consisted of three classes of students who were randomly assigned to four different treatment groups: one that received direct WCF, one with indirect WCF using error codes. The researchers also included two control treatments called practicing writing, and revision without feedback. The participants were 62 fourteen-year old Dutch multilingual secondary school students. Students in the practice control group did not receive any feedback or revision, but they completed two new tasks to practice their writing skills. Van Beuningen et al. (2008, p.283) note that that a treatment on the control groups was included "to be able to unambiguously distinguish between effects of error correction and time-on-task effects", and in that way students in the practice group thus spent as much time writing as the students in the error correction group. Students in the self-correction control group revised their own texts, but they did not receive any feedback.

In the first session, the researchers gave the participants a vocabulary test to establish the students' overall language proficiency. The second session consisted of students then writing the first

writing task and a week later, students received feedback and revised their texts. The control groups practiced their writing skills once more, or self-corrected their errors without feedback. In the third session, the students were presented with two new writing tasks. The results of the study were that the groups that revised their writing, including the self-correct group, produced fewer errors in their revisions than in their initial texts; however, only the groups receiving WCF had significant accuracy gains on their revisions. Van Beuningen et al. (2008) note that the results also show that direct WCF has a long-term effect on students' accuracy on new tasks. Direct error correction was deemed to be superior to indirect WCF in the long-term in the study, and the researchers acknowledge that this contradicts the work of Ferris (1995) and Lalande (1982), who have argued that learners benefit more from indirect corrective feedback. Van Beuningen et al. (2008) note that the explanation given by Chandler (2003) for her study's results also make sense for their results, in that students who received direct WCF were able to instantly internalise the correct form. However, students who revised their texts based on indirect feedback could not do this because they did not know whether their own correction was accurate or not. From the results, the researchers concluded that the long-term effect of direct error correction is greater than the other types; and on text revisions, direct and indirect feedback are both effective.

In Ruegg's (2010) study, first-year L2 English majors from a 12-week semester writing class at a Japanese university wrote journal entries where feedback on content alone and grammar was given. Ruegg (2010) looked at repetitions of the same errors to measure accuracy. One group of participants was given feedback on content, while another was given indirect feedback on grammatical and spelling errors in addition to feedback on content. The mistakes made by both groups were analysed, and repetitions of the same errors on subsequent instances of writing were counted. The results showed that the group receiving the indirect feedback had significantly less repetition of the same errors in subsequent journal entries in comparison with the control group. Unlike other studies of WCF, Ruegg (2010) used a unique method for measuring accuracy by looking at repetitions of the same errors.

Van Beuningen et al. (2012) compared the effects of both indirect and direct written unfocused feedback on the writing of 268 secondary school L2 learners of Dutch. Their research questions looked at whether unfocused feedback improved accuracy during revision and new texts. They also wanted to discover whether unfocused error correction leads to avoidance of complex structures and if there was an influence of pupils' educational level on WCF efficacy. The researchers included two experimental treatment groups as well as two control groups: one control group undertook self-editing and the other completed writing practice without revisions. The first experimental group received unfocused direct WCF where the researcher identified all existing linguistic errors and provided the pupils with the correct target forms. The indirect group received indirect feedback that was an indication, and the category of each error, and thus could be called metalinguistic feedback. All the groups wrote texts for the pretest, followed by a treatment session. They then wrote a post-test text and a delayed post-test text. The results from the pre-test showed that the groups started out with similar proficiency levels, but the treatment groups receiving feedback had a higher proficiency level and fewer errors in the following tests. Van Beuningen et al.'s (2012) results showed that both direct and indirect unfocused WCF led to improved accuracy in two new pieces of writing, when compared to the two control groups. They conclude that unfocused WCF is of use and helps students to increase their accuracy and proficiency level, both in the text revisions and in new texts (2012). They further refuted Truscott's (1996) claim that unfocused WCF causes students to avoid more complex structures. Furthermore, to be able to guarantee that any improvements made by the experimental group were not due to the extra time spent on self-editing or revision, Van Beuningen et al. (2012) introduced a self-editing group that would be able to control for time-on-task. They also found that what they refer to as the higher educational level pupils, outperformed the lower educational level learners on the different linguistic measures used in the study. However, they never found a significant interaction between WCF effectiveness and learners' educational levels.

Fazilatfar et al.'s (2014) study on unfocused WCF looked at the effect of WCF on students' accuracy, and syntactic and lexical complexity development. Thirty advanced learners from an

English course aged between 16- 23 participated in the study. All the participants passed previous upper intermediate conversational levels or were enrolled by the Oxford Quick Placement Test. The learners were divided into experimental and control groups. Both groups received exactly the same instruction and wrote ten compositions throughout the three-month course. The experimental group received unfocused WCF for each composition. The results showed a significant gain for both syntactic and lexical complexity in the unfocused WCF group. Thus, they concluded that unfocused WCF was beneficial and that unfocused WCF would not prevent learners from making attempts at more complex features in their new compositions. Furthermore, they noted that it also may lead to improving complexity in their interlanguage acquisition.

Bonilla López et al. (2018) investigated the potential of unfocused WCF as editing and learning tools. The participants were 139 low-intermediate second language writers at a university in Costa Rica. Their native language was Spanish and their majors were either in English teaching or English. Their average English proficiency level was lower intermediate on the Oxford Quick Placement Test. They were randomly divided into four groups: direct corrections of grammatical errors, direct corrections of grammatical and nongrammatical errors, metalinguistic codes for grammatical errors, or metalinguistic codes for grammatical and nongrammatical errors – together with a control group. Students wrote a 175-word opinion essay about topics related to chapters in their course textbook. When revising their original texts, they were only given their original text without the WCF as the researchers noted that they wanted to ensure the revision task would not become a copying exercise. Participants also wrote a new piece, and four weeks later, another new text. Results showed that direct corrections and codes effectively enhanced learners' grammatical and nongrammatical accuracy on text revisions, but four weeks later, the long-term effect was only seen in the direct grammar groups. This study of unfocused feedback is thus one of the studies that supports the case for direct WCF.

Karim and Nassaji's (2018) study attempted to investigate the short and long term effect of unfocused WCF. Here, 53 adult intermediate students of English as a second language (ESL) studying at two ESL schools in Canada, were randomly divided into four groups: direct; underline only;

underline and metalinguistic; and a control group. The participants produced four narrative pieces of writing from different picture prompts and revised them over a three-week period. In the sixth week, all WCF groups wrote a new text from a new picture prompt. The results showed that on the revision tasks the WCF groups had significant gains in accuracy compared to the control group, but on new texts all accuracy gains were non-significant. The authors concluded that although studies of WCF can provide insights into the effectiveness of different types of WCF in certain contexts, they cannot establish which form of WCF is superior in all contexts.

Another recent study by Nicolás–Conesa et al. (2019), also found that the unfocused WCF groups outperformed the control group in accuracy on text re-writes, but also in the long-term on new texts. Their study was made up of 46 English majors with Oxford Placement Test intermediate level scores, enrolled in a semester-long composition course at a Spanish university, with two treatment groups receiving direct or indirect WCF. The participants were asked to process the feedback via written *linguaging*, “the process of making meaning and shaping knowledge and experience through language” (Swain, 2006, p.98). The researchers also included a control group who wrote and rewrote their texts, but also engaged in linguaging. The participants wrote narrative tasks and had access to their original texts without corrections, during the rewriting of the pretest. The results showed that there were positive significant gains in the short-term and also in the long-term from the combined effect of WCF and written linguaging. The authors note that the combined effect of WCF and linguaging could be an issue, as they did not include a group that received WCF, and did not engage in linguaging.

In conclusion, focused feedback refers to feedback that is given only on a specific and preselected type of error, and unfocused feedback refers to feedback that is given on all or a range of error types. A pattern that emerges from these studies of unfocused WCF is that in the short-term on text revisions, unfocused WCF can improve students’ performance on accuracy measures, but in the long-term on new texts, the results of these studies still show a mixed picture.

2.2 Complexity, Accuracy and Fluency

When the effects of corrective feedback on students' written performance are analysed, researchers **sometimes** look at the variation in complexity, accuracy, and fluency of student output. These measures have become common in the literature and are also known as CAF measures, which usually take the form of ratios, frequencies, or formulas. CAF measures represent three dimensions of L2 performance, and have been a popular area of research since the 1990s. Research on CAF measures began in the 1980s when researchers started to point out a distinction between fluent and accurate language use, but complexity appeared later in the literature on CAF measures, starting in the 1990s (Housen & Kuiken, 2009). Skehan (1989) proposed a model with CAF measures as the three main dimensions of proficiency. Pallotti (2009) suggested that CAF measures are not a theory or a research programme in themselves, but that research into CAF measures is of great value to L2 researchers as they believe that the principal dimensions of L2 performance can be captured by the notions of complexity, accuracy and fluency (Ellis & Barkhuizen, 2005). CAF measures also have additional value for L2 teachers because they can utilise the research findings to improve their practice and their students' language performance. CAF measures are complex and multidimensional, and researchers of L2 acquisition have different views on how these components can be defined and operationalised (Housen & Kuiken, 2007). Comparing and generalising results becomes difficult, as researchers often use different measurements (Ellis & Barkhuizen, 2005). Thus it is important to gain a clear understanding of what the different approaches to operationalise CAF measures are. Recently, CAF measures have become commonplace in the literature, and are a growing research area, appearing mostly as dependent variables in SLA research (Housen & Kuiken, 2009). Studies using CAF measures as dependent variables have focused on the effects of L2 acquisition influenced by such things as instruction, individual learner differences, task type, learning contexts and corrective feedback (Freed, 1995; Bygate, 1999; Skehan & Foster, 1999; Derwing & Rossiter, 2003; Yuan & Ellis, 2003; Muñoz, 2006; Kuiken & Vedder, 2007; Truscott & Hsu, 2008; Van Beuningen et al. 2012).

2.2.1 Definitions and Operationalisation of Complexity

Out of the three proficiency measures, complexity is usually said to be the most controversial and has been defined in various ways. Since various definitions of complexity coexist, choosing which one to use is problematic for researchers. Ellis and Barkhuizen (2005, p.139) define complexity “as the extent to which learners produce elaborated language”, whereas Wolfe-Quintero et al. (1998, p.69) define complexity as “a wide variety of both basic and sophisticated structures and words that are available to the learner”. Wolfe-Quintero et al. (1998, p.4) also defined complexity as “the scope of expanding or restructured second language knowledge”. Skehan (1998) defines complexity as challenging language, and Ellis (2008, p.475) defines complexity as “the capacity to use more advanced language”. According to Bulte and Housen (2014) the more components a feature or system consists of, and the denser the relationships between its components are, then the more complex the feature or system is. Although at first these definitions appear to be worded differently, on closer examination of the language used they are similar in meaning. In summary, the literature shows there are many definitions, which are rather vague and overlapping using similar language, for example: challenging, sophisticated and advanced.

Researchers have used a variety of ways to measure complexity. According to Norris and Ortega (2009), at least three grammatical complexity measures (global complexity, phrasal complexity, and complexity by subordination) should be measured, because language can be elaborated at three different levels. Morphological complexity is a relatively new construct in second language (L2) studies (Pallotti, 2015) (Brezina & Pallotti, 2016); however, SLA researchers most often focus on syntactic or grammatical complexity (Ellis & Barkhuizen, 2005). Syntactic complexity has been measured as clauses per T-unit, also known as minimally terminable units. This is defined as the shortest grammatically allowable sentences into which writing can be split (Hunt, 1965), the number of dependent clauses per total clauses, or number of dependent clauses per T-unit. Another further method to measure complexity is by looking at lexical diversity, which can be

measured by the type-token ratio (TTR), which is the number of word types divided by all word tokens. TTR, however, is flawed because long works depress TTR (Ellis & Barkhuizen, 2005).

Some studies have used a measure called G: the index of Guiraud (Guiraud, 1959). This is the transformation of the standard TTR that controls for text length effects in the calculation of the TTR. It is measured using the number of types/the square root of the number of tokens, as an index of lexical diversity (Bulte & Housen, 2014). Another newer alternative is to use the diversity index D (Malvern et al., 2004), a mathematical transformation of the standard TTR. Unlike the index of Guiraud, D reduces the effects of text length and provides an indication of the degree of word repetition in a text. The less repetition and the more varied words are used in a text, the higher the score for D. D can be computed using *vocd* in the Coh-Metrix programme (www.cohmetrix.com) (McNamara et al., 2010). *Vocd* is a method for measuring the diversity of text units, and takes a number of subsamples of tokens at random from the data. It computes the average type-token ratio for each of these lengths, and then finds the curve that best fits the type-token ratio curve which has just been produced (McKee et al., 2000).

Other researchers count the ratio of functional words to lexical words, or the number of different word families (Ellis & Barkhuizen, 2005). More recently the measures of textual lexical diversity (MTLD) (McCarthy & Jarvis, 2010) have been used. D and MTLD are similar, but D tends to be based on TTR using random selection and curve fitting to reduce the impact of text length. MTLD, however, uses TTR as a cut-off point to examine the text length, for which a writer can maintain a certain level of lexical diversity (McCarthy, 2005). McCarthy (2005, p.88) states that “maintaining the text structure rather than sampling the text provides a more authentic measure of diversity”.

In summary, there are many definitions of complexity, which are rather vague and overlapping, and researchers have used a variety of ways to measure it that include TTR, and more recent methods such as MTLD.

2.2.2 Definitions and Operationalisation of Accuracy

Accuracy is usually considered as the most straightforward construct of CAF, and refers to the degree of conformity to language usage norms. Skehan (1996, p.23) defines accuracy as “how well the target language is produced in relation to the rule system of the target language”. Wolfe-Quintero et al. (1998, p. 4) define accuracy as “the conformity of second language knowledge to target language norms”. Another definition of accuracy is given by Wolfe-Quintero et al. (1998, p.33): “the ability to be free from errors while using language to communicate in either writing or speech”. Housen and Kuiken (2009) also note that accuracy is the most straightforward and internally consistent of the three CAF measures.

The operationalisation of accuracy has been conducted in various ways. Accuracy has been measured by the number of error-free T-units, and errors per T-unit (Skehan & Foster, 1997) (Larsen-Freeman & Long, 1991). Ellis and Barkhuizen (2005) have used a general measure of accuracy, such as percentage of error-free clauses or number of errors per 100 words. Accuracy based on specific measures of accuracy is most often employed for research on a targeted structure, such as the focused CF studies as reviewed above. Regarding the focused WCF studies, the issue is that operationalising accuracy by performance on specific forms would not show a representative picture of a learner’s general use of the language, because it may not reliably be a correct representation of the students’ general accuracy.

In summary, accuracy is usually considered as the most straightforward and internally consistent of the three CAF measures. Accuracy has been measured by the number of error-free T-units, and errors per T-unit or percentage of error-free clauses or number of errors per 100 words.

2.2.3 Definitions and Operationalisation of Fluency

The construct of fluency has also been defined in different ways. Skehan (2009, p. 511) has called it “the capacity to produce speech at normal rate and without interruption”. It has also been defined as “the production of language in real time without undue pausing or hesitation” by Ellis and Barkhuizen (2005, p.139). Fluency can be further narrowed down into the following: speed fluency, breakdown fluency and repair fluency (Tavakoli & Skehan, 2005). Most of the data that has been analysed by SLA researchers regarding fluency, has been oral data (Lennon, 1990; Towell et al., 1996; Kormos & Dénes, 2004). Chambers (1997) defined fluency as the ease, eloquence, smoothness and native-like of speech or writing; and Polio (2001) defined fluency as examining how native-like the writing appears. Pallotti (2020) explains that fluency is the extent to which linguistic production is fast or smooth. On closer examination, these definitions are rather similar, but certain aspects of the definitions, such as “eloquence” and “native-like” could even be considered to overlap with some aspects of proficiency.

Similar to the other CAF measures, fluency has been operationalised in different ways. As the current focus is written data, examining the way fluency has been operationalised for written data is of importance. Larsen-Freeman (1978) used the average number of words per composition of EFL students (Polio, 2001), and recommends using the number of words, clauses, and T-units, and clauses per text, to analyse production in a writing sample. Wolfe-Quintero et al. (1998) argue that the frequency measures, for instance, the number of words, are not as valid a measure of fluency, recommending that fluency ratios for written data, such as words per minute, words per clause, words per sentence are better. Polio (2001) has also questioned the relationship between words per minute and the quality of writing. The number of correctly spelled words written and number of sentences written have also been used (Rosenthal, 2007), although this causes an overlap with accuracy.

In conclusion, fluency has also been defined and operationalised in different ways, and the operationalisation of fluency is still controversial.

2.2.4 Issues with the Definitions and Operationalisation of CAF

One of the many controversial aspects of CAF research relates to the varying definitions and operationalisation of the three CAF constructs. Some studies that investigate CAF do not explicitly define the terms complexity, accuracy, and fluency - nor how they have been operationalised. There are also problems concerning their operationalisation, because CAF can be measured in different ways. Definitions and explanations of the way they are measured in the study is also very important when explaining the methodology. CAFs are often measured differently across studies, and this limits the interpretation, and also the comparability of CAF findings. Some researchers point out that this could be one of the reasons for inconsistent findings in CAF studies (Housen & Kuiken, 2009; Ellis, 2008; Wolfe- Quintero et al., 1998). Whether general or more specific measures of CAF are more appropriate is another important issue (Norris & Ortega, 2009) and there have recently been calls for having several measures for each of the CAF constructs. Several critical surveys of measurement practices in CAF research have been conducted (Polio, 1997; Wolfe- Quintero et al., 1998; Ortega, 2003; Norris & Ortega 2009; Polio, 2001). However, despite there being challenges, CAFs are concepts that are still widely used to evaluate L2 learners and there is an absence of alternatives (Housen & Kuiken, 2009). As well as their definitions and operationalisation, other CAF research is concerned with their interaction.

2.2.5 Skehan's Trade-off Hypothesis

Studies researching CAF have found trade-off effects between the three CAF measurements (Skehan & Foster, 1997; Bygate, 2001; Michel, Kuiken & Vedder, 2007). It has been argued that learners cannot attend to all areas of CAF performance, especially when a task is demanding, due to the processing demands being greater than learners' capacity. Due to this, learners must prioritise their language performance, and according to Ellis & Barkhuizen (2005) this can result in trade-off effects.

Ellis (1994) noted that learners must attend consciously to the input and make efforts to monitor output; however, this can interfere with fluent reception and production (1994) meaning that focusing on one component of language performance may reduce performance in another component. Skehan (2009) also argues that there is a competitive relationship between CAF because of limited mental resources, specifically working memory and limited attentional capacity which is known as the limited attentional model. Skehan (1996, 2009) found that CAF components do not develop simultaneously and that students tend to overlook one area while concentrating on another.

Skehan's (2009) trade-off hypothesis states that the dimensions are interdependent such that increased performance in one area might occur at the expense of performance in the other areas. This means that a higher performance in one component, for example accuracy, corresponds to lower performance in another, for example fluency, and a competitive relationship between CAF may therefore exist. Skehan (2009) also notes that adult learners prioritise meaning over form, which may hinder further language development. Therefore, there can be trade-offs between accuracy and complexity. When there is improvement in two CAF areas, Skehan (2009) notes that the information manipulation during the task requires the students to use subordination, or co-ordination, which can help improve complexity, although the task structure may aid accuracy. The second rationale Skehan (2009) proposes is that when analysing group data, some individuals may prioritise one area of CAF, while others prioritise another area. Thus, the aggregated data may give the appearance of connected growers.

There is much focus on complexity and accuracy as a trade off, because as Skehan (1998) notes, increasing complexity reflects the increased risk-taking of learner languages, while accuracy measures the ability to avoid errors and control existing resources. When there are no trade-off effects but both areas show improvement, this could be due to the result of group aggregated data giving the appearance of two areas of CAF showing improvement – although this may not be the case in reality (Skehan, 1998). Skehan (1998) points out that there are different kinds of learners and they may emphasise a different area of CAF, for example fluency or accuracy. In a study looking at the effect

of planning during three oral tasks, a trade-off between accuracy and complexity was found by Skehan and Foster (1997). Other research suggests that students at different proficiency levels may decrease performance in one CAF area while improving in another. Yuan and Ellis (2003), in a study that looked at the effect of planning on oral language performance, found trade-off effects between accuracy and fluency when looking at group score comparisons. Other researchers who believe that human attention and processing capacity are limited include Skehan (1996); Skehan and Foster (1997, 1999); and Bygate (1999). Yuan and Ellis (2003); Michel et al. (2007); Skehan (2009); Ahmadian and Tavakoli, (2011); and Ferrari (2012) all found a trade-off between complexity and accuracy in oral tasks.

2.2.6 Robinson's Cognition Hypothesis

A different point of view is presented by other researchers, such as Robinson (2003), who holds that language learners can simultaneously access multiple and non-competitive attentional pools. This is known as Robinson's Cognition Hypothesis (Robinson & Gilabert, 2007) and claims that increased accuracy and complexity can be caused by more cognitively-demanding tasks. They state that this can be done by altering task complexity and this increased cognitive demand of a task can lead to simultaneous improvement of complexity and accuracy. Studies which have provided some support to Robinson's Cognition Hypothesis include Kuiken and Vedder (2007, 2008), and Ishikawa (2006). Ishikawa's (2006) study looked at the effects of manipulating task complexity and found that by increasing task complexity, this would increase the accuracy, fluency, and complexity of written language production. Ishikawa's (2006) results also indicated that increasing task complexity increased the accuracy, fluency, and complexity of written language production without any trade-offs appearing. Mizera (2006) found accuracy and fluency to be connected growers, and their results showed that the speed fluency measure was negatively correlated with the number of errors. However, the number of errors and the number of pauses measure was positively correlated. A longitudinal oral study by Vercellotti (2012), looked at data from 66 English L2 students from Korean, Chinese, and

Arabic language backgrounds, over three to nine months. She found that higher grammatical complexity scores were correlated with higher fluency, and thus found connected growers. She concluded that students did not focus their development on one CAF construct at the expense of another.

2.2.7 Complexity and Dynamic Systems Theory

According to some researchers, the empirical evidence available so far does not lead to support for either Robinson's (2003) cognitions hypothesis or Skehan's (2009) trade-off hypothesis, but another hypothesis that allows for both trade-offs and connected growth is the Dynamic Systems Theory. This theory assumes that cognitive resources are limited; however, they are connected and may be compensatory. Furthermore, since all variables in the system are interrelated, changes will affect all the other parts of the system. Trade-offs and connected growers are both possible in Complexity Theory (Larsen-Freeman, 2012) where trade-off effects may be found, but do not have a mutually exclusive, causal, or linear relationship (de Bot et al., 2007). Similar to Robinson's cognition hypothesis, this also assumes a more complex task will encourage learners to accomplish more. Van Geert and Steenbeek (2005) note that in the long term, language proficiency interacts in ways that are not only competitive, but also supportive. Some argue that all subcomponents of CAFs are connected growers and, thus improvements in any one area do not imply a trade-off in another measure. Spoelman and Verspoor (2010) investigated the relationship between accuracy and complexity, in a Dutch student learning Finnish, for three years. They found that accuracy varied in the early stages but then stabilised. They note that the way the interaction of accuracy and complexity changed over time and was not stable, indicate a dynamic system that neither supports Skehan's trade-off hypothesis nor the cognition hypothesis, but points at a dynamic system. Gunnarson (2012) also found neither competition between complexity and accuracy and did not find any significant relationship between syntactic complexity and fluency. Rosmawati (2013) investigated the development of

accuracy and complexity in an advanced L2 learner's academic writing. Rosmawati (2013) argues that the results may suggest that complexity and accuracy measures demonstrate the characteristics of a dynamic system as their development was variable and non-linear. He notes that the interactions among complexity, accuracy and fluency changed back and forth from a competitive relationship to a positive one. The magnitude also fluctuated, ranging from a weak association to a very strong one.

In conclusion, the trade-off effects that have been the focus of the relationships between CAF measures in the literature are much more prevalent in oral tasks than in writing tasks and the more recent studies of the interaction between CAF measures point to a more complex interaction than the linear models predict. Although there are interactions between CAF measures, there is also interesting research that examines the relationship between CAF and other variables.

2.2.8 Relationship of CAF with Other Variables

CAF may also be affected by various factors. These may include task factors such as genre of the tasks, task type, task structure, task condition, planning time, familiarity with the topic, and the degree of cognitive complexity of the tasks that learners are trying to perform (Rahimpour, 1999, 2008). The type of pedagogical intervention is also an external factor that may affect CAF, for example different types of feedback (indirect, direct, focused or unfocused) or implicit or explicit instruction and characteristics of the input (Housen & Kuiken, 2009). Learner orientation - meaning whether the learner prioritises complexity, accuracy, or fluency during language performance - has been suggested to influence CAF (Larsen-Freeman, 2006). Other factors that influence CAF can be individual difference (ID) variables, such as learners' proficiency level, anxiety of the L2 learners, motivation, or aptitude (Rahimpour, 1999, 2008). These ID variables are reviewed below.

2.2.9 Future Direction of CAF Research

CAF is still a relatively new research area and although CAFs are useful ways of measuring language performance, many controversies remain. Norris and Ortega (2009) note that since there are no precise agreed standards in the field, reported research may not contribute to the accumulated knowledge because the findings cannot be compared. Larsen-Freeman (1997) suggest looking at CAF research from a dynamic or complex systems or chaos/complexity theory. Larsen-Freeman (2009) also called for more longitudinal CAF research. Most researchers agree that using different ways of measuring and using different definitions are an area that researchers need to focus on to come up with standardised definitions and standard ways to measure CAF. Pallotti (2020) notes that it is first necessary to clearly define underlying constructs, so that each measure or group of measures refers to a well-identifiable construct, and that when this has been completed, only then will practitioners be more likely to be able to apply the research findings to their practice. In this way, they will be able to improve their students' language performance, as the results of CAF research will be more generalisable.

Although the operationalisations of CAF are controversial, to conduct a study, a balanced approach drawing from the most up-to-date research, and taking into consideration the differing points of view, is necessary. Studies using measures of CAF tend to regard the learners as homogenous; however, this is often far from the reality of the modern day classroom. New research in L2 classrooms should try to include individual differences to reflect the heterogeneity of the students and thus capture factors that potentially interact with CAF.

Dynamic systems and complexity theory also postulate that cognitive resources are limited but connected and possibly compensatory, and all variables in the system are interrelated, so changes will affect all the other parts of the system. Thus, if all variables are interconnected, CAF measures must be influenced by ID variables that learners exhibit.

2.3 Individual Differences

A number of important learner individual characteristics can affect the efficacy of a particular type of corrective feedback (Ellis, 2010). These learner individual characteristics have been recognised as important variables in the process of language learning. Research has shown that individual factors may influence the speed at which languages are being learnt, and also the level of L2 attainment (Carroll, 1962; Gardner, 1985; Ehrman & Oxford, 1995; Ellis, 2004). Individual differences (IDs) are characteristics in respect of which individuals differ from each other, and Gardner subdivided IDs into affective, cognitive and personality-related individual differences (1985). Dörnyei (2005) defines IDs in the following way: “They concern anything that marks a person as a distinct and a unique human being” (p.3) and further notes that many studies in the area of IDs have concluded that they are the most important predictors of achievement in a second language. Since individual differences, such as proficiency, aptitude and attitudes have been found to be important in the way they affect second language acquisition, the way they play a role in learners’ responses to written corrective feedback, should also be explored (Bakri, 2016). Many researchers note that language learning aptitude is one of the most important of the ID variables (Dörnyei & Skehan, 2003; Ellis, 2004; Li, 2015). There has been a large amount of research on the impact of ID variables on the effects of oral feedback, but studies on the relationship between ID variables and WCF are scarcer (Flahive, 2010).

2.3.1 L2 Proficiency

L2 Proficiency is usually not listed as an ID variable, unlike aptitude and attitudes, and it is mostly assumed that proficiency is the goal of language learning and teaching (Harsch, 2017). Tremblay (2011) points out that there is a clear lack of consensus in the methods that researchers have employed to evaluate proficiency.

2.3.1.1 Definitions of Proficiency

Proficiency has been defined, but no consensus has yet been arrived at, due to the fact that setting the basic criteria to assess it is difficult. Some argue that proficiency is an index of the production and comprehension abilities that L2 learners develop across linguistic domains and modalities to communicate (Bachman, 1990). According to Canale (1983), language proficiency encompasses a language learner's or user's knowledge systems, skills and communicative abilities. Bachman (1990, p. 16) defined language proficiency as "knowledge, competence, or ability in the use of a language". Peregoy and Boyle (2005) note that language proficiency can be defined as the ability to use a language appropriately and effectively throughout the range of personal, school, and work and social situations, required for daily living in society.

2.3.1.2 The Main Proficiency Frameworks, Scales and Tests

Although there are many proficiency frameworks and scales used across the world, the three long-established ones used to describe language proficiency are the most popular. They are the Interagency Language Roundtable (ILR) scale (the standard grading scale for language proficiency in the United States' Federal-level service); the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines, and the Common European Framework of Reference (CEFR). Harsch (2017) argues that the widespread use of the CEFR shows a relative degree of consensus about how proficiency is currently conceived, but it has also been criticised by some researchers. Cummins (2008), for example argues that proficiency is a concept which can be looked at from different angles, even though there is a dominant conception of it which is portrayed in the CEFR.

As well as proficiency frameworks and scales, there exist a large amount of L2 English proficiency tests, for example the IELTS, TOEFL, TOEIC and the Oxford Quick Placement test. These tests, intended to measure global language abilities, are mostly used to measure students' L2 proficiency and to divide participants into proficiency groups. The Oxford Quick Placement test was

developed by Oxford University Press to provide institutions with a quick, reliable way to place English language students into the correct level of English class. It assesses reading, vocabulary and grammar, and is quick and easy to administer. The Oxford Quick Placement test is available in both paper-and-pen and computer-based versions. The computer-based version, however, also includes a listening component.

2.3.1.3 L2 Proficiency and Written Corrective Feedback

There have been calls in the literature for research that examines learners at different proficiency levels to investigate if high proficiency learners benefit more from unfocused WCF, and if low proficiency level learners benefit more from a focused approach (Bitchner & Storch, 2016). In general, the literature suggests that focused and direct WCF for low proficiency learners may be preferable as it reduces their cognitive load, whereas higher proficiency level learners can cope with, process, and use a wider range of input at the same time (Kang & Han, 2015). Indirect and unfocused WCF is beneficial for advanced learners, and perhaps even preferred to allow learners to make use of their linguistic repertoire and discern their own errors (Ellis, 2009; Kang & Han, 2015). Kang and Han (2015) proposed that learner proficiency has a moderating effect on the efficacy of WCF, with an increase in effect size as the proficiency level increased, and a negative effect size for feedback when given to beginners (Kang & Han, 2015). Some other studies have shown that there is a significant relationship between WCF and proficiency level. Bitchener et al., (2005) and Chandler (2003) discovered that direct WCF might be beneficial for low proficiency learners with specific categories of errors. Other studies argue that WCF's effectiveness depends on the cognitive developmental readiness and students' writing proficiency level (Gu nette, 2007).

In summary, it is assumed that proficiency is the goal of language learning and teaching, but no consensus on its definition has been arrived at due to the fact that setting the basic criteria to assess it is difficult. There are, however, three major proficiency frameworks and many well-known proficiency tests in usage. The research on proficiency and WCF is still in its infancy, but the

consensus is that it has a moderating effect on the efficacy of different feedback types. As well as proficiency, another ID variable that can affect the efficacy of WCF is language learning aptitude.

2.3.2 Language Learning Aptitude: Definitions and Operationalisation

Carroll stated that “specialised abilities beyond general intelligence play an important role in learning a foreign language” (1981, p. 27), and with this statement, was referring to the concept of language learning aptitude. The conceptualisation of aptitude developed by John B. Carroll in the 1950s is still a benchmark for researchers today (Roehr-Brackin, 2020). Language learning aptitude assumes “there is a specific talent for learning foreign languages which exhibits considerable variation between individual learners” (Dörnyei & Skehan, 2003, p.590). Language learning aptitude has been defined as a group of abilities which enables some learners to acquire new language material faster and easier than others (Dörnyei, 2005). Another definition is, relative to other individuals, how well an individual can learn a foreign language under given conditions and in a given amount of time (Stansfield, 1989). Carroll and Sapon (2002, p.23) call language aptitude a set of cognitive abilities that are “predictive of how well relative to other individuals, an individual can learn a foreign language in a given amount of time and under given conditions”. Robinson (2005, p.46) considers language aptitude as “cognitive abilities’ information processing drawn on during L2 learning and performance in various contexts and at different stages”. Many researchers note that language aptitude is considered to be one of the most important of the ID factors influencing the rate and success of second language learning (Carroll, 1981; Skehan, 1989; Ehrman & Oxford, 1995; Sawyer & Ranta, 2001; Ellis, 2004; and Dörnyei, 2005). The definitions of language learning aptitude appear to be rather similar, lack controversy and in general imply that there are certain abilities that allow individuals to learn a foreign language, which vary between individuals. The definition of aptitude by Carroll and Sapon (2002) will be used for this study.

Carroll (1981) explains that language aptitude is distinct from intelligence and achievement, and is comprised of four components: phonetic coding ability (the ability to identify and store in long-term memory new language sounds or strings of sounds) (Carroll, 1971, p. 4); grammatical sensitivity (the ability to recognise the grammatical functions of words or other linguistic entities within sentences (Robinson, 2001, p.324); inductive language learning ability (a process where the learner discovers rules by observing examples), and associative memory (the recall of items based on their association with other items and environmental cues). Carroll (1981) also claims that these are stable traits. Skehan (1998, p.204) has further combined grammatical sensitivity and inductive language learning ability into a component he calls language analytic ability, defined as “the capacity to infer rules of language and make linguistic generalisation or extrapolations”.

Language aptitude is generally considered as a cognitive variable in second language acquisition, but the ways in which aptitude is conceptualised and operationalised in the research is far from homogeneous. This can be seen by the fact that a variety of aptitude tests have been developed for various purposes (Li, 2015). Kormos (2013) also argues that the lack of a clear definition for language learning aptitude is due to most developers of language aptitude tests using an empirically based psychometric approach to test development. Language aptitude is a complex, multi-faceted factor, and there is no single foreign language aptitude, but instead a whole range of foreign language aptitudes, which are included in the domain of cognitive IDs (Granena & Long, 2013).

The concept of language aptitude has changed over the last 15 years, from being seen as a stable and a fixed trait to being considered as a more dynamic one (Larsen-Freeman, 2001). The assumptions behind the concept of language aptitude were that it was a relatively stable characteristic, and could not be changed by training, or be affected by previous experience. However, language aptitude is now seen as a mix of different abilities that can assist in the different stages and processes of language learning (Wen et al., 2017; Kormos, 2013). Dörnyei (2010) also notes that the language-learning aptitude is of a dynamic nature. Grigorenko et al. (2000, p.401) argue that “language aptitude

is a form of developing expertise rather than an entity fixed at birth". Most studies on language aptitude have focused on instructed settings and the rate of L2 learning. Therefore, when looking at the studies on the rate of L2 learning instructed settings it is important to note that they may not generalise to high levels of L2 proficiency, and long-term achievement (Dörnyei, 2010; Linck et al., 2013).

More recently, new ideas of language learning aptitude include phonological short-term memory and working memory as relevant subcomponents of the construct (Miyake & Friedman, 1998; Sawyer & Ranta, 2001; Wen et al., 2017). Miyake & Friedman (1998) state that working memory is a cognitive mechanism that performs the dual function of information storage and processing, and argue that working memory is where the components of aptitude converge. Robinson (2001) explores a theoretically motivated model of aptitude; whereby primary abilities include speed of processing in phonological working memory, pattern recognition, and grammatical sensitivity. Robinson (2002) also looked at the correlation between working memory and language learning aptitude, and discovered that working memory had a moderately strong correlation with language aptitude scores. De Keyser and Koeth (2011) also support Robinson's (2001) model in which aptitude is a complex construct made up of several cognitive characteristics. Robinson (2005) further argues that phonological short-term memory and working memory are also components of aptitude complexes. Other researchers also believe that working memory capacity could be as important as the traditional concept of foreign language aptitude.

2.3.2.1 Aptitude Tests

Research into language aptitude first started during the late 1950s and early 1960s (Spolsky, 1995). During the 1950s and 1960s, Carroll (1962) carried out studies on second language learning and found out that learning a second language appeared to be a particular talent, independent of performance on general intelligence tests. The most interesting achievement in this period was the

Modern Language Aptitude Test (MLAT) (Carroll & Sapon, 1959; Carroll (1962). Other aptitude tests include CANAL-F (Cognitive Ability for the Novelty in Acquisition of Language – Foreign) (Grigorienco et al., 2000), the Hi-LAB (Linck et al., 2013), and the PLAB (Pimsleur, 1966).

The PLAB and MLAT are the two most frequently used aptitude tests in research; however, they have different target audiences: the MLAT is used with adults, and the PLAB is used for high school learners. Presently, the most widely used and accepted aptitude test is the Modern Language Aptitude Test (MLAT) developed by Carroll and Sapon (1959). The MLAT is an aptitude test that measures an individual's ability to learn a foreign language (Carroll et al., 2010) and is for English-speaking adults. The test was first published by the Psychological Corporation in 1959, and has only one form, which has not changed since its inception. The MLAT has mostly been used for adults in government language programmes, and is now closed to researchers. Carroll (1962) identified four components of language learning aptitude: phonetic coding ability, grammatical sensitivity, rote learning ability, and inductive language learning ability.

The MLAT consists of five sections, each one testing separate abilities: Number Learning, which measures memory as well as auditory comprehension of a foreign language; Phonetic Script that measures the ability to learn correlations between a speech sound and written symbols; Spelling Clues/Hidden Words that measures the vocabulary knowledge of English as sound-symbol association ability; Words in Sentences that measure sensitivity to grammatical structure; and Paired Associates that measure rote memorisation ability (Carroll & Sapon, 2002).

The aforementioned aptitude tests however, have been criticised. Li (2015), states that the weakness of the MLAT and similar aptitude tests are due to the test being validated empirically, but not theoretically. Furthermore, Li (2015) notes that the five subtests do not correspond directly with the four hypothesised aptitude components. Many researchers have criticised the MLAT (Robinson, 2005; Sawyer & Ranta, 2001; Skehan, 2002) based on the fact that the constructs underlying the MLAT do not represent a complete definition of L2 aptitude.

The LLAMA, an aptitude test battery (Meara, 2005), is a recent development that is freely available and is computer based. The LLAMA, however, has not been extensively standardised (Meara, 2005) for use in high-stakes situations. The LLAMA test has been used to measure aptitude in several studies such as those conducted by Abrahamsson and Hyltenstam (2008), Forsberg-Lundell & Sandgren (2013); Granena (2013); Granena & Long, (2013). The test includes four sections: LLAMA B, (vocabulary learning), LLAMA D, (sound recognition) LLAMA E (sound-symbol associations), and LLAMA F, (grammatical inferencing). The structure of the test allows test takers to focus on language forms, rehearse materials and employ problem-solving strategies (Granena, 2013). Granena argues that the LLAMA B, E and F create learning conditions that encourage explicit induction (for example consciously reasoning relationships and rules), rather than implicit induction (non-intentional learning). Granena (2013) states that LLAMA B, E, and F tapped a dimension of aptitude characterised by explicit cognitive processes, explicit associative learning, and rote learning ability. Thus, they labeled this dimension explicit language aptitude, and further argued that explicit language aptitude is likely associated with explicit language learning. Overall, Granena (2013) and Rogers et al. (2017) have shown that the LLAMA aptitude tests are robust and are not subject to external individual differences. A more critical stance toward the validity of the LLAMA test was adopted by Bokander and Bylund (2020), who suggest that researchers using the LLAMA battery must treat their results with caution, after their results showed that only the LLAMA B produced scores that fit a latent trait model with sufficient accuracy, and thus the LLAMA could be refined further. Most recently (May 2021), a new version of the LLAMA known as version 3.0, is now available; however, at present it is a beta version. The removal of many of the two-way multiple-choice options have improved the reliability of the test (Meara & Rogers, 2020).

2.3.2.2 Studies on Aptitude and the Relationship between Aptitude and L2 Development/Achievement

There has been a great deal of SLA research conducted since the MLAT and PLAB were first used (Robinson & Ellis, 2008). During the 1970s and 1980s, aptitude was hardly researched and the reason for this lack of interest was due to criticisms against the concept (Dörnyei & Skehan, 2003). Until the 1990s, there were very few attempts to research aptitude, due to the development of communicative approaches to language teaching. Individuals started to question the value of testing and labeling learners with an aptitude score, as they believed this to be anti-egalitarian. Furthermore, the lack of interest also stemmed from the focus on rote learning and grammatical patterns in the MLAT being associated with audio-lingual language learning methodologies that were becoming outdated and being replaced with the communicative approach (Dörnyei, 2005). However, in the last few years, due to the lack of validity of the criticism (its anti-egalitarian nature) leveled at the concept of language learning aptitude (Sáfár & Kormos, 2008), language learning aptitude has become an increasingly researched topic (Dörnyei, 2005).

A large proportion of most aptitude research has investigated the relationship between language aptitude and L2 achievement. Researchers have found that generally, language aptitude is positively related to L2 achievement, particularly in adults (Ehrman & Oxford, 1995; Horwitz, 1988; Parry & Child, 1990).

Ehrman and Oxford (1995) examined the relationships of a variety of ID variables to the end of training proficiency ratings, in speaking and reading for adults in intensive training in the US. They found a strong correlation between aptitude measured using the MLAT aptitude battery, and overall learning success in a communicative language learning environment.

The following more recent studies of DeKeyser (2000), Abrahamsson and Hyltenstam (2008) and Granena and Long (2013) discovered significant interaction between L2 attainment and language aptitude in adult learners. Abrahamsson and Hyltenstam (2008) explored DeKeyser's (2000) hypothesis that late learners with high analytical verbal abilities can achieve near native-like second

language proficiency. The authors looked at 42 L2 Swedish learners who were native Spanish speakers, who also considered themselves near-native in Swedish proficiency. The results showed that high language aptitude is highly predictive of second language proficiency, especially in everyday verbal interaction. Abrahamsson and Hyltenstam's (2008) findings were divergent to DeKeyser's (2000) hypothesis that aptitude will not be a significant predictor among early L2 learners. They found that language aptitude appeared to be necessary in adult near-native SLA and also advantageous in child SLA. Their findings were also different to those of DeKeyser (2000) and DeKeyser et al. (2010), since DeKeyser (2000) and DeKeyser et al.'s (2010) results did not show any significant interaction between proficiency and aptitude among early learners.

Granena and Long (2013) examined the interaction between language aptitude and ultimate morphosyntactic attainment. Aptitude was measured by the LLAMA aptitude test (Meara, 2005). The participants of the study were 65 early and late Chinese native speaker learners of L2 Spanish, and twelve native speakers also participated as controls. The results showed significant correlations were found between language aptitude, and lexis and collocation scores, in the late age of onset group (ages 16–29). Significant correlations were also found in language aptitude, when measured using the LLAMA test (Meara, 2005), and pronunciation scores (Granena & Long, 2013).

Li (2015) conducted a meta-analysis that looked at empirical research on the role of language aptitude in L2 grammar acquisition, and focused on the relationship between aptitude and morphosyntactic attainment. The meta-analysis consisted of 33 different studies, some of which were interactional studies that examined the relationships between aptitude and the effectiveness of instructional treatment. Other studies examined the relationships between aptitude and ultimate L2 attainment. Only studies conducted since publication of the MLAT were included in the meta-analysis, and studies published after May 2013 were not included. The meta-analysis included all studies that used traditional aptitude measures, and attitude tests that included the MLAT, PLAB, VORD, DLAB, and LLAMA, and tests that only measure one aptitude component. Li (2015) discovered that aptitude had an overall medium effect size in both predictive and interactional research. They also discovered

that young learners were more reliant on aptitude than older learners in predictive studies, although in interactional studies the opposite was found. The results of the meta-analysis showed an aggregated effect size of .31. Li (2015) discovered that aptitude affected high school participants' learning more than it affected university students' learning. Li (2015) hypothesises that this may be because high school students are less advanced learners, so aptitude plays a greater role at lower levels of proficiency. Li (2015) also found that aptitude showed a moderate interaction with L2 grammar learning and that aptitude scores and language analytic ability were more predictive of grammar learning than rote memory and phonetic coding ability. Li (2015) also found that the effectiveness of explicit instruction was more related to aptitude than implicit instruction. They perceive that this suggests in instructed settings, older learners are less likely to draw on aptitude than younger learners. Li (2015) concludes by stating that the importance of aptitude has been partially exaggerated and it is predictive of initial second language grammatical competence, but less so of the later stages of learning.

In a second meta-analysis that synthesised the results of 66 empirical studies, Li (2016) examined the construct validity of language aptitude, focused on the relationship between aptitude and other individual difference variables, as well as on the relationship between aptitude and SLA in terms of both general proficiency and specific skills. It concluded that aptitude was independent of other cognitive and affective factors; executive working memory was more strongly related with aptitude and aptitude components than phonological short-term memory. Aptitude measured using full-length tests was found to be a strong predictor of general L2 proficiency; however, it had lower predictive validity for L2 writing and vocabulary learning. Li (2016) notes that this indicates that writing could need a different skill set compared to those measured in traditional aptitude tests. Furthermore, the importance of organisation and content when evaluating learners' written performance may be a reason for the low predictive validity of traditional aptitude tests for writing. Aptitude was a strong predictor of L2 proficiency, with about 25% of the variance accounted for. The mean effect size was larger for high-school learners, than for university-level learners. Furthermore,

the results showed a weak relationship between aptitude and motivation and this would suggest that the two variables are separate constructs. Li (2016) also states that the important role played by aptitude in learning the linguistic aspects of L2 competence means that language aptitude cannot be dissociated from L2 writing competence. Li (2016) further concluded that language aptitude was more likely to be drawn upon in explicit rather than implicit instruction similar to Dörnyei and Skehan (2003, p. 600) who claimed that aptitude “presupposes a requirement that there is a focus on form”. The meta-analysis also looked at the relationship between aptitude and the ID variables of motivation (a weak correlation), anxiety (a negative correlation) and intelligence (a strong positive correlation). Li (2016) concludes that in future, empirical research needs to address the following: including a need to develop measures or include components that are sensitive to implicit learning that tends to occur in non-tutored settings, and the fact that the criterion variables, predictor variable, and covariates are measured or operationalised in different ways in primary research on aptitude.

To conclude, looking at the results of DeKeyser (2000), DeKeyser et al. (2010), Abrahamsson and Hyltenstam (2008) and the results of Granena and Long (2013), the research regarding the relationship between aptitude and long-term L2 achievement has produced varied findings for child learners, but not for adults. For adults, it is concluded that generally, language aptitude is positively related to L2 achievement. Carroll et al. (2010, p.19) pointed out that the results of the large amount of aptitude research “seem to offer strong support for the major claims about the nature of language aptitude”. However, the theoretical claims need to be tested through aggregation of the quantitative results, rather than impressionistic judgments (Li, 2015). Li’s (2015) meta-analysis concludes that the importance of aptitude has been rather exaggerated and that it is predictive of initial L2 grammatical competence, but less so of the later stages of learning. Li (2015) also argues that it is a construct, that affects learning outcomes in explicit conditions, and that future research needs to work on clarifying the construct of aptitude. Li’s (2016) meta-analysis, however, found that there was a strong association ($r = .49$) between aptitude and general L2 proficiency, a correlation that is impressive for a meta-analysis. In summary, the findings of Li’s (2015, 2016) meta-analyses, show

that aptitude significantly predicts L2 attainment in terms of both general proficiency and knowledge of morphosyntax. In general, the studies of aptitude show that aptitude measured using full-length tests is a strong predictor of L2 proficiency, but they have lower predictive validity for vocabulary learning and L2 writing.

2.3.2.3 Research on Language Learning Aptitude within the Interactionist Paradigm

There is another strand of research within the interactionist paradigm that instead of viewing aptitude as a fixed predictor of L2 success, treats it as a dynamic construct whereby its role differs as a function of the processing demands of different learning conditions (Robinson, 2002, 2005). This research focuses on aptitude-treatment interactions that try to show how different aptitude components may play different roles dependent on the instructional treatment. Researchers such as De Graaff (1997), Robinson (1997), and Williams (1999) all reported that differences in aptitude, as measured by subtests of the MLAT, resulted in learning variance in implicit and explicit learning conditions. De Graaff (1997), Robinson (1997), and Williams's (1999) experimental studies suggest that language aptitude may play a role under explicit learning conditions and that learning under implicit and incidental conditions is not affected by IDs in language aptitude.

Yilmaz (2013) investigated the role of working memory capacity and language analytic ability through the extent to which L2 learners benefit from two different types of oral feedback. The researcher found that explicit correction worked more favourably than recasts, only when the learners in the compared groups had working memory capacity or high language analytic ability.

A study by Sheen (2007) examined the effectiveness of direct and explicit metalinguistic WCF in interaction with language analytic ability as measured by a test based on a language analysis test developed by Otto, and used previously by Schmitt et al. (2004). This was a focused WCF study on English articles with 91 adult intermediate ESL learners of differing LI backgrounds. The study took place at a community college in the United States, and the students came from six intact classrooms

in the American Language Programme. The students came from both international and immigrant ESL populations, their ages varying from 21-56 years. They also came from different educational backgrounds, for example from those with high school diplomas, to doctoral degrees, and the three major groups were Korean, Hispanic, and Polish. The study comprised of three groups: a direct-only correction group, a direct metalinguistic correction group, (Sheen explains that direct metalinguistic correction indicates the location of an error and provides the correct form as well as including metalinguistic comments that explain the correct form) and a control group. The study found that both treatment groups performed much more favourably than the control group on the immediate post-tests, but the direct metalinguistic group performed better than the direct-only correction group in the delayed post-tests. The results also showed a significant positive correlation between students' gains and their aptitude for language analysis. Sheen notes that there was a stronger relationship with language analytic ability and acquisition in the direct metalinguistic group than in the direct-only group. These results show that written WCF that targets single linguistic features, improved learners' accuracy, especially when the learners had high language analytic ability. Sheen found direct metalinguistic WCF helped the learners with a higher aptitude for language analysis more than lower aptitude learners. Sheen's study, however, is limited in several ways; the writing task treatment was very short, and the study was a focused study on articles, limiting generalisability.

Kormos and Trebits (2012) examined the fluency, accuracy, syntactic complexity and lexical variety of performance in two types of written and spoken narrative tasks. The participants were 44 upper-intermediate learners of English in a Hungarian secondary school, aged between 15 and 18. The teachers of the students rated their level of proficiency as slightly above intermediate, thus corresponding to B1/B2 in the Common European Framework of Reference. Participants completed four narrative tasks: two involving cartoon descriptions, and two involving picture narration. The results showed a complex relationship between aptitude components and task performance, dependent on different conditions. The researchers found the components of aptitude that seemed to be most strongly related to the complexity and accuracy of production, were aptitude measured as

deductive ability and grammatical sensitivity, measured using the Hungarian version of the MLAT (HUNLAT). The results also showed that when writing, the participants used more varied vocabulary than they did in speech, but found similar performance for syntactic complexity.

Benson and DeKeyser's (2018) study compared the effects of direct and metalinguistic WCF on the simple past and present perfect tenses. To do so, it examined essays by 151 English second language learners from an academic English class at university. The study looked at to what extent learner differences in language-analytic ability mediated the effects of these two types of explicit WCF by using the LLAMA F, aptitude test. The participants in both feedback groups were given WCF on two essays, but the control group received general comments on content. Next, the participants in all groups completed two additional writing tasks to see if the WCF led to greater gains in accuracy compared to the control group. They found that both WCF groups performed better than the control group on new pieces of writing, and that in the long-term, direct feedback was more effective than metalinguistic feedback for the simple past tense. Furthermore, they found that learners with greater language aptitude benefited more from direct WCF while learners with lower language aptitude benefited more from metalinguistic feedback.

In conclusion, it is apparent that there is a relationship between aptitude and learning, and that different aptitude components demonstrate differential predictive validity for various aspects of learning. Explicit forms of WCF worked more favourably when the learners have greater working memory capacity and high language analytic ability. WCF targeting a single linguistic feature improves learners' accuracy, especially when the learners have high language analytic ability. Along with aptitude, another important ID variable that some believe is one of the leading predictors of success in learning a language, is that of attitudes and beliefs.

2.3.3 Attitudes and Beliefs

As well as aptitude, learner beliefs or attitudes is another variable that is categorised as an ID factor. Ellis notes that attitudes are dynamic and situated (2008) and are seen as a part of metacognitive knowledge (Flavell, 1987), that include all that individuals understand about themselves as learners and thinkers, including their needs and goals. Attitudes in SLA research started in the 1950s and continues to be researched at present (Dörnyei, 2001).

Attitudes do not have a clear and all-accepted definition. Attitude has been defined as a mental state that includes beliefs and feelings (Latchanna & Dagneu, 2009). Victori and Lockhart (1995, p. 224) defined attitudes as “general assumptions that students hold about themselves as learners, about factors influencing language learning, and about the nature of language learning and teaching”. Beliefs about language learning were defined as “opinions on a variety of issues and controversies related to language learning” (Horwitz, 1987 p. 120). Dörnyei (2005, p.214) distinguishes between ‘attitudes’ and ‘beliefs’. He states:

The main difference, in fact, between the conception of attitudes and beliefs is exactly that the latter have a stronger factual support whereas the former are more deeply embedded in our minds and can be rooted back in our past or in the influence of the modelling example of some significant person around us.

Brown (1994, p.168) pointed out that attitudes develop in early childhood and are the result of the interaction with friends, peers, parents, and society. Stern (1983) classified learner attitudes into three different types: (i) attitudes towards the community and people who speak the L2, (ii) attitudes towards the language concerned, and (iii) attitudes towards languages and language learning in general.

However, in the literature on attitudes, the boundaries between attitudes and learner beliefs are not clear-cut, and they have often been used interchangeably; therefore, attitudes and learner beliefs will be considered as synonymous during this study. Although the definitions are expressed differently, on closer inspection they are rather similar and useful to define the concept. The definition of attitudes by Victori and Lockhart, that attitudes are "general assumptions that students hold about

themselves as learners, about factors influencing language learning, and about the nature of language learning and teaching" (1995, p.224) will be applied to both attitudes and beliefs throughout the study.

Some believe that the attitude toward learning the language is one of the leading predictors of success in learning a language, and therefore this factor is important when designing language instruction and feedback (Hall, 2009). Studying the beliefs of second language learners is also important (Oxford, 2003). Weinburgh (1998) also notes that attitudes can influence behaviours, for example, reading books and speaking in a foreign language, and that learners' attitudes dictate whether they will be successful. Supportive and positive beliefs can help in overcoming problems and in that way, sustain motivation, while negative beliefs can lead to decreased motivation, anxiety, and frustration (Puchta, 1999).

According to (Barcelos, 2003), three different approaches to investigating learners' beliefs can be distinguished. First, the normative approach shows beliefs as preconceived notions, which can be found by using Likert-style questionnaires. An example is the Beliefs About Language Learning Inventory, known as the BALLI (Horowitz, 1987). Second, the metacognitive approach looks at learners' metacognitive knowledge about language learning as 'theories in action', according to Wenden (1999). These can be examined by means of the content analysis of learner self-reports in semi-structured interviews. Third, the contextual approach looks at learner beliefs as varying according to context (Barcelos, 2003).

Empirical Studies on Attitudes and Learner Beliefs

Researchers have realised the influence of attitudes and language beliefs in second and foreign language learning, and the way it can affect its progression. Their focus has been on students' beliefs about language learning and the effect they have on students' motivation, anxiety, and strategy use (Gregerson & Horwitz, 2002). A large proportion of the research on learner beliefs is related to describing and classifying the types of beliefs learners hold and the sources of beliefs. Wenden (2001)

argues that foreign and second language learner beliefs are not often studied in SLA. Ellis further points out that there are very few studies examining the relationship between beliefs and language learning (Ellis, 2008). Horwitz studied relationships between learning strategies, and goals of students and teachers for foreign or second language acquisition of each group, and thus created an instrument called the Beliefs about Language Learning Inventory (BALLI). The BALLI contributed to the growth of this research. There are now three BALLIs: one is used for ESL students, another for foreign language teachers, and one for foreign language students (Horwitz, 1990). Most studies of learner beliefs have focused on what belief learners have, and how learners' backgrounds have an effect on their beliefs (Tanaka & Ellis, 2003). Studies on learner beliefs show that they may vary due to age, cultural background, learning environment, stage of learning, and target language.

2.3.3.2 Studies Examining the Relationships between Attitude and Proficiency

It is widely claimed that learning beliefs have a large effect on learning (Horwitz, 1987; Mantle-Bromley, 1995; White, 1999), but there are few studies that examine the extent to which learning beliefs affect achievement. Among them, Park (1995) found that three variables predicted students TOEFL scores to some extent: one being a belief variable. Park's (1995) study examined 332 Korean university EFL students' beliefs about language learning, their language learning strategies, and the relationships among their beliefs, strategy use, and L2 proficiency. The learners who reported being confident in learning English and the intention of speaking to others in English, were more likely to use English actively and outside the classroom, and this was related to an improvement in L2 proficiency.

Mori (1999) found that strong beliefs in the ability to learn is inherited and cannot be improved by effort. An avoidance of ambiguity was related to lower L2 achievement, and learners who believed that it was easy to learn an L2, manifested higher levels of achievement. The subjects were 187 university students enrolled in Japanese language classes at various proficiency levels in the US, and

Mori examined the interaction between epistemological beliefs and beliefs about language learning, and most importantly the relationship between beliefs and L2 achievement. Furthermore, the study showed that low level and advanced learners have different beliefs.

During a 15-week study-abroad programme, Tanaka and Ellis (2003) examined changes in students' beliefs about language learning and their English proficiency. The participants were Japanese university students and their beliefs about language learning were measured by a questionnaire and their English proficiency was measured by the TOEFL test. The results showed statistically significant changes in the students' beliefs relating to analytic language learning (where students did not attach as much value on their teacher using their L1 to explain things in class), and experiential language learning and confidence (the students were less concerned about making mistakes, more confident in speaking English and were more satisfied with their progress). The researchers note that there was no relationship between beliefs relating to self-efficacy and confidence and proficiency, before and after a three-month period of study abroad, but that learners' beliefs about analytic learning were negatively related to proficiency.

In conclusion, attitudes and beliefs are significant in enabling learners to learn effectively. It has been claimed that having positive or negative attitudes towards a certain language can exert considerable influence upon a learner's performance and proficiency, but given the small amount of empirical research this claim is difficult to prove. As well as examining the relationship between attitude and proficiency, others have looked at the relationship between student attitudes and corrective feedback.

2.3.3.3 Empirical Research on Student Attitudes and Corrective Feedback

To understand the role of WCF in classrooms, it is essential to determine whether individual differences of such attitudes mediate the effects of different kinds of WCF. The following studies

attempt to examine what students' attitudes towards WCF are, and how these mediate the effects of WCF.

Lim (1990) examined the attitudes and beliefs towards feedback of secondary school students in Singapore, and found students had positive attitudes toward peer correction. Furthermore, students found their grammatical errors to be the most important followed by vocabulary, spelling, organisation of ideas and punctuation errors. Most importantly, Lim (1990) found the students stated that the primary burden for correcting errors should be the responsibility of their teachers.

A study by Hedgcock and Lefkowitz (1994) had similar results, and investigated the differences in writing contexts and student motivation at a U.S. university, by surveying students on their attitudes to CF, and how they influenced their views of text quality and writing processes. The results showed students were most concerned with grammatical accuracy, and that they had a positive attitude towards written corrective feedback. The participants were 110 ESL and 137 EFL students. Storch and Wigglesworth's (2010) study found that the effectiveness of WCF is affected by a student's attitudes, beliefs and objectives, and that these factors are unfortunately often omitted from most WCF research when seeing if the feedback is beneficial. They found students may be negatively affected by the feedback they receive and that "learners' attitude towards the feedback affects not only whether and how learners respond to the feedback provided, but ultimately whether there is long term learning" (Storch & Wigglesworth, 2010, p.44).

The most influential studies examining L2 learners' beliefs about grammar instruction are probably those by Schulz (1996, 2001). The first study by Schulz (1996) looked at the beliefs of U.S. adult foreign language students and teachers, about the role of grammar instruction and CF in language learning. They found that 90% of students thought it necessary to be corrected while speaking in class. However, only 34% of the teachers thought the same, showing differences between student and teacher beliefs. Regarding writing, around 90% of teachers and students agreed that errors should be explicitly corrected in writing. In another study by Schulz (2001), 607 Colombian foreign language students and 122 teachers; and 824 American foreign language students and 92 teachers,

answered a questionnaire to find out student and teacher perceptions concerning the role of explicit grammar instruction and corrective feedback. It found a relatively high agreement between students and teachers across cultures, but there were also some differences, the largest of which related to error correction; how and how often errors should be corrected. Schulz concludes that since these differences between student and teacher belief systems can impact student learning negatively, teachers should explore their students' perceptions to deal with potential conflicts between student beliefs and instructional practices.

Loewen et al.'s (2009) study investigated the beliefs of L2 learners regarding grammar instruction and CF, whereby 754 L2 students at an American university participated and filled in a questionnaire. The study investigated differences in beliefs among learners studying different target languages, and the results showed that among ESL learners and those studying a foreign language, there were varied beliefs about grammar instruction and error correction. It found that learners had a general view of the usefulness of grammar instruction, but some learners held negative views of grammar instruction. It was also found that learners of Arabic and Chinese were more positive about grammar instruction and also error correction than learners of other languages. Different students expressed varying beliefs, which Loewen et al. (2009) say may be due to their previous language learning contexts. Hyland and Hyland (2006) also note that ESL students from cultures where teachers are highly directive, expect teachers to comment on errors. Students may resent their teacher if they do comment or notice their errors. They further point out that it is also possible that some students may disregard cultural models as they have individual identities, and thus caution must be taken when generalising results of attitudes research.

When summarising the research on corrective feedback and attitudes and beliefs, the patterns emerging are that learners have a general view of the efficacy or usefulness of grammar instruction, and prefer more explicit forms of error correction. Furthermore, different students expressed varying beliefs, possibly due to their previous language learning contexts.

2.4 Conclusion and Gaps in Existing Research

The effectiveness of WCF has been debated in the literature due to the results yielded from studies being varied, but more recently, most researchers agree that WCF has a positive effect on students' writing. The literature has argued that both focused and unfocused WCF may help second language learners to improve the linguistic accuracy of their written productions (Bitchener, 2008; Bitchener & Knoch, 2010), and the debate surrounding whether WCF is effective or not has mostly moved on to analysing which type is the most effective (Bitchner & Storch, 2016). A review of studies that compare indirect and direct feedback showed mixed results, as some researchers have favoured direct WCF (de Jong, & Kuiken, 2008; Bitchener & Knoch, 2010; van Beuningen et al., 2012; Alimohammadi & Nejadansari, 2014) while others have found indirect WCF the most effective (Lalande, 1982; Eslami, 2014). Kang and Han's (2015) meta-analysis found that focused feedback had a much greater positive effect than unfocused feedback. They also discovered a greater effect resulting from direct feedback, compared to indirect feedback.

Most of the studies on WCF produced over the past ten years have concentrated on focused WCF, possibly because of practical reasons, and not because it is more effective than unfocused corrective feedback (Ferris, 2010; Bitchner & Storch, 2016). Many, therefore, have called for more studies that investigate the learning potential that can arise out of unfocused WCF (Xu, 2009; Van Beuningen, 2010; Bitchner & Storch, 2016). There is still no clear answer on whether focused or unfocused WCF is more effective at different levels of proficiency, and further research is needed (Bitchener & Ferris, 2012). Furthermore, to investigate the potential effects of WCF, Truscott and Hsu (2008) and Van Beuningen et al. (2012) argue that further research should compare independent written works instead of comparing an initial text to a text revision. This would enable researchers to examine if longer-term language development is taking place, rather than immediate effects that do not transfer to new tasks.

When analysing the effects of corrective feedback on student's written performance, researchers sometimes look at the variation in complexity, accuracy, and fluency (CAF) of student output. CAFs have become common in the literature and represent the three dimensions of L2 production and performance. CAFs do not constitute a research programme or a theory in themselves (Pallotti, 2009), but CAF research is of great value to L2 researchers because the principal dimensions of L2 performance can be captured by the notions of complexity, accuracy, and fluency (Ellis & Barkhuizen, 2005) and thus be used to determine if WCF has an effect on student performance. The operationalisation of CAFs are controversial, so a balanced approach drawing from the most up-to-date research and taking into consideration the differing points of view is necessary. Furthermore, out of all the studies on WCF, few studies take complexity into account when measuring the effects of WCF on writing and more are thus needed.

Recently there has been a large amount of research on the impact of ID variables on the effects of oral feedback, but studies on the relationship between ID variables and WCF are scarcer. L2 proficiency has a moderating effect on the efficacy of WCF (Bitchener, 2008; Ellis, 2009), with an increase in effect size as the proficiency level increases, and a negative effect for feedback when given to beginners (Kang & Han, 2015). The research on proficiency and WCF is still in its infancy, but the overall consensus is that it has a moderating effect on the efficacy of different feedback types.

Other ID variables, including language aptitude and WCF, have also been examined. Researchers have mostly found that students with higher aptitudes benefit more from WCF (Granena & Long, 2013; Li, 2015) - although research on the type of feedback, and if it is more favourable for students with high or low aptitude, has been mixed.

Attitudes or student beliefs are variables that are categorised as ID factors. Attitudes are dynamic and situated, and are viewed as a part of metacognitive knowledge (Ellis, 2008). Attitudes are significant in enabling learners to learn effectively, and thus having positive or negative attitudes towards corrective feedback can exert considerable influence upon a student's learning. Recent studies on WCF have called for more research to investigate learners' attitudes and beliefs regarding

WCF (Bitchener & Storch, 2016). As corrective feedback is provided for the benefit of the learners, obtaining a clearer idea about the effectiveness of WCF - and understanding their attitudes and preferences toward it, and its effects on CAF measures - are important.

Therefore, in light of the gaps in the literature on L2 writing and feedback for acquisition studies, as well as the lack of empirical evidence regarding the effects of unfocused WCF on CAF, the present study aims to examine the effects of unfocused direct, indirect, and metalinguistic WCF on the CAF of L2 English students' writing. It also investigates if the moderating variables of aptitude, attitudes, and proficiency affect the uptake of feedback by addressing the research questions that can be found in Chapter 3, next.

Chapter 3: Methodology

3.1 Introduction

This chapter outlines the research design and the instruments used to undertake this quasi-experimental study. First, the research questions are presented. The chapter then explains the ethical approval process. Following that, the setting and the demographic of the participants are detailed and the experimental treatment is described. The different instruments used in the study are then described. The pilot study methodology and its experimental set-up as well as the results of the pilot study that informed changes to the main study and its conclusions, are then presented. Next, complexity, accuracy, and fluency (CAF) measures and their operationalisation used in the main study are offered. Then, the timeline of the study is presented, and following that, the information on inter-rater reliability analysis and data processing are shown. Finally, preliminary data checks that were carried out are explained and last of all a summary is presented.

3.2 Research Questions

RQ1a. Does unfocused corrective feedback lead to an increase in the accuracy, complexity, and fluency of student writing on revised tasks, compared to no feedback?

RQ1b. Which of the following types of unfocused corrective feedback have a greater influence on the accuracy, complexity and fluency of student writing on revised tasks?

- (a) direct corrective feedback in the form of written corrections of errors on students' compositions;
- (b) indirect corrective feedback in the form of error underlining;

(c) metalinguistic feedback in the form of error codes with metalinguistic information.

RQ1c. Does unfocused corrective feedback lead to an increase in the accuracy, complexity and fluency of student writing on new tasks, compared to no feedback?

RQ1d. Which of the types of unfocused corrective feedback has a greater influence on the accuracy, complexity and fluency of student writing on new tasks?

RQ2. Is there a relationship between gains in accuracy, gains in complexity and gains in fluency on revised and new tasks?

RQ3. Is there a relationship between L2 proficiency and gains in complexity, accuracy and fluency on revised and new tasks?

RQ4. Is there a relationship between aptitude and gains in complexity, accuracy and fluency on revised and new tasks?

RQ5. Is there a relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy, and fluency on revised and new tasks?

3.3 Ethical Approval

To begin with, ethical approval was obtained from the University of Essex. Next, ethical approval was also applied for and obtained at The American University of Sharjah, where the research took place. The participants were handed out consent forms and participation information sheets that explained the study and what they would need to do if they decided to participate. Before giving their

consent, assurance was given to the participants that their anonymity would be preserved in all reports of the study, by deleting or changing details that might reveal their identities and giving each student an identifier number that would refer to any of the questionnaires they answered, tests they took or essays they wrote. The researcher also gave a talk that elaborated on the study and answered any questions participants might have. These signed consent sheets were then collected.

3.4 Setting

The study was conducted at a co-educational private university in the United Arab Emirates. Participants came from the lower of two levels of academic writing classes available for freshmen. Students take academic writing classes twice a week and the study was conducted during class time, but the writing tasks completed for the study were not part of the course and did not count for a grade.

3.5 Participants

In total, 139 English academic writing 001 students participated voluntarily in the study. The data was collected from four intact groups of 001 academic writing classes. All students are required to take an in-house placement test when entering the University, and would either be placed in Academic writing 101 or the lower level Academic writing 001 class. For the purposes of this study, 001 students were selected as they had the sufficient English writing proficiency to produce the length and level of writings required, but would also have errors in their compositions, so feedback could be given.

Out of the 139 participants, 71 were female and 68 male and the age of the participants was between 18 and 20 years old. The mean age of the participants was 18.6. Most were L1 speakers of Arabic (N = 115), while a small minority were L1 speakers of Urdu (N = 24). A total of 45 participants

reported having spent between 1 and 3 months in an English-speaking country, with most having spent time in the United Kingdom.

Most students reported that they speak Arabic and English at home (N= 74) and regarding the language spoken with friends, the majority (N= 111) spoke Arabic and English. The majority of students came from a school where English was the medium of instruction (N= 102) and the minority (N=36) came from a school where Arabic was the medium of instruction. The majority of students had studied English as a subject for 12 years (N= 121); 7 students had studied English for 8 years, 7 students had studied English for only 5 years; and 3 students had studied English for 16 years.

The participants were given the Oxford Quick Placement Test (UCLES, 2004) to find out their English proficiency levels, and were scored between 36 and 58 points. Following the Quick Placement Test’s interpretation of results, The Association of Language Testers in Europe levels ranged from lower intermediate to very advanced (6 students in lower intermediate, 50 students in upper intermediate, 65 students in advanced, 18 students in very advanced). The descriptive statistics of the students’ proficiency on the Oxford Quick Placement Test are presented in Table 3, below:

Table 3. Descriptive Statistics of the Oxford Quick Placement Test

	N	Minimum	Maximum	Max. Possible	Mean	Std. Deviation	Mean %
Oxford Quick Placement Test	139	36	58	60	48.81	5.033	81.35

3.6 Experimental Treatment

This section describes the methodology used in the study. The set up was quasi-experimental and the tasks were designed for experimental purposes only and, therefore, were not part of the standard syllabus. All tasks, however, were administered during class periods. In total there were three groups with treatment conditions where corrective feedback was given: direct WCF n= 40, indirect WCF n=36, metalinguistic WCF n=34; and a control group n =35 that did not receive WCF. The writing tasks the participants needed to complete were persuasive/argument essays that answered a prompt. The topic of the essay was assigned to the students from a list of topics. The participants wrote a practice essay, a pre-test essay and a post-test essay, which was a re-write of the pre-test. They also wrote a delayed post-test essay, which was a new topic. For all the essays, students wrote an argument/ persuasion topic and had 10 minutes planning time and 30 minutes writing time.

The participants in the study were made up of four intact groups. The four intact groups thus became the direct feedback group, the indirect feedback group, the metalinguistic feedback group and a control group. All linguistic errors were identified by the researcher, as illustrated in the transcribed examples below.

The students in the direct feedback group received direct unfocused CF on their argument/persuasion essays that took the form of identifying both the errors, and giving the target form on the writings they produced. Their errors were corrected by the researcher by crossing out the erroneous forms and providing the corresponding target forms above the errors as illustrated in Example 1 below:

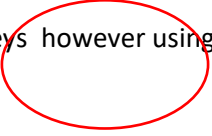
Example 1

Children are introduced to technology in ~~there~~^{eir} childhood.

The students in the indirect feedback group received indirect unfocused feedback on their argument/persuasion essays on all linguistic errors that took the form of circling the error with a red pen as illustrated in example 2 below:

Example 2


Other species of monkeys however using the technology available today have a less painful experience.



Students in the metalinguistic group received feedback on their argument/persuasion essays in the form of metalinguistic error codes. The metalinguistic error codes were chosen as they were the most commonly used codes among teachers, and comparable to other studies using metalinguistic error codes. The most frequent grammatical errors that students would be able to self-correct had their own code, for example S/V; however, more infrequent grammatical mistakes were marked as Gr, meaning grammatical error (See Appendix A) on all linguistic errors as in example 3 below:

Example 3

Children using iPhones are having a lower level of concentration.



The students who received metalinguistic error codes were given a legend that explained what each code represented, and were also instructed on how to use and interpret the error codes. The list of error codes can be found in Appendix A.

The control group did not receive any feedback on their argument/persuasion essays. In all cases, the WCF was provided by the researcher, because having the same researcher providing all feedback ensures greater consistency of feedback.

3.7 Instruments

The instruments used in this study comprised of a questionnaire to elicit student preferences for the essay topics, an attitudes and a language background questionnaire, the Oxford Quick Placement Test, and the LLAMA_B and F aptitude tests.

3.7.1 Essay Topic Questionnaire

A questionnaire with the list of the most frequent topics students would write about when given the choice to write about any subject they pleased, was generated from past student essays collected from previous offerings of the writing course. This list was then narrowed down to the twenty most popular topics (Appendix B). Students were then given the list of twenty topics and next to each topic, participants could choose from:

I am interested in this topic; or

I am neither interested in this topic nor not interested in this topic; or

I am not interested in this topic.

The results of this questionnaire narrowed the topics down to the three most popular (*Children should not be given smartphones; Animal testing should be banned; Going to university doesn't always lead to success*) and had the intention of allowing students to write about something they were interested

in, rather than the researcher choosing a topic. The rationale for this was due to the possibility that if a student is not interested in a topic, it could negatively affect their writing performance.

3.7.2 Language Background Questionnaire

The language background questionnaire was a pen-and-paper based questionnaire comprising of 10 questions. Students were given around ten minutes to fill in the questionnaire and could ask questions if they had any doubt. The questionnaire asked about their mother tongue, how many years they had studied in an English medium school and other such questions about their experience of English (Appendix C). The results are reported in section 3.5, above.

3.7.3 Attitudes Questionnaire

The attitudes questionnaire was a pen-and-paper questionnaire (Appendix D) that attempted to elicit what students' attitudes toward corrective feedback were by asking 25 questions. 22 of the questions used a Likert scale, and the other three were open answers. Questions elicited students' attitudes by asking such things as 'I consider error correction useful' and 'I always look at the corrective feedback given by the teacher'. The coding is as follows: 1 represents the most positive attitude toward corrective feedback (strongly agree); and a 5 represents the least positive attitude towards corrective feedback (strongly disagree). Statements that represented a negative attitude toward corrective feedback used reverse coding.

3.7.4 The Oxford Quick Placement Test

The Oxford Quick placement test is used to test grammar and vocabulary, and is usually given to students to place them in their appropriate levels. In this study, it was used to measure proficiency. It is known as a reliable and time saving test. The test has two versions, a pen and paper version and a computer-based version. The computer-based version has a listening component while the pen and paper version does not. Due to the context of the study being on WCF the pen and paper version - without the listening component - was used. The pen and paper test consists of 60 multiple choice questions and answers. Test takers are required to answer all the questions in 30 minutes. The results are reported in Table 3, above.

3.7.5 The LLAMA Aptitude Test

The language aptitude test used in this study was the LLAMA language aptitude test version 2.0 (Meara, 2005). Recently LLAMA 3.0 was been released as a beta version, but was not available when this study was conducted. This test is a computer-based test and includes the following four subtests:

LLAMA B: a test of vocabulary learning;

LLAMA D: a test of sound recognition;

LLAMA E: a test of sound-symbol association;

LLAMA F: a test of grammatical inferencing.

The LLAMA is loosely based on the components that appear in Carrol & Sapon's (1959) MLAT, but has a more user friendly interface. The LLAMA test was chosen since the MLAT is designed for L1 speakers of English, while the LLAMA aptitude test is mostly language-neutral and apart from the MLAT, the LLAMA aptitude test is the most widely used aptitude test in L2 learning research at the

moment (Roehr-Brackin, 2020). Results on the validity and reliability of the LLAMA are mixed. An exploratory validation study of the LLAMA by Granena and Long (2013) using a 186 participant sample from three different language backgrounds, yielded acceptable levels of reliability. Granena (2013) and Rogers et al. (2016) also found positive results. A more critical stance toward the validity of the LLAMA test was adopted by Bokander & Bylund (2020). They suggest that researchers using the LLAMA battery must treat their results with caution after their results showed that only the LLAMA B produced scores that fit a latent trait model with sufficient accuracy, and thus the LLAMA could be refined further.

Although the LLAMA battery is comprised of four sub-tests, only two (LLAMA_B and F) were chosen for the purpose of this study. The rationale for this is that due to the study being about the effects of written corrective feedback and the tasks the participants had to produce were writing tasks, the LLAMA_B relating to vocabulary learning and LLAMA_E relating to grammatical inferencing were the two components of aptitude that were the most relevant to the study.

3.7.5.1 LLAMA_B

The first subtest used was the LLAMA_B test. This test is a vocabulary learning task which measures the test taker's ability to learn relatively large amounts of vocabulary in a relatively short amount of time. The test presents a set of 20 fictional objects on the screen that represent words taken from a Central American language which are assigned to the pictures only for the purposes of the test. During the study phase, students need to learn the names of as many of the twenty objects as they can in two minutes. They can then click the objects as many times as they wish, but are not allowed to take notes. During the testing phase, a word is then presented and they need to match the word with the object. Test-takers score five points for each object that is correctly identified by its name. LLAMA_B scores range from 0-100. A score of 0-20 is a poor score, 25-45 is an average score, 50-70 is a good score

and 75-100 is a very high score that few people score unless they are using a mnemonic system. The user interface of the LLAMA_B is shown a Figure 1.

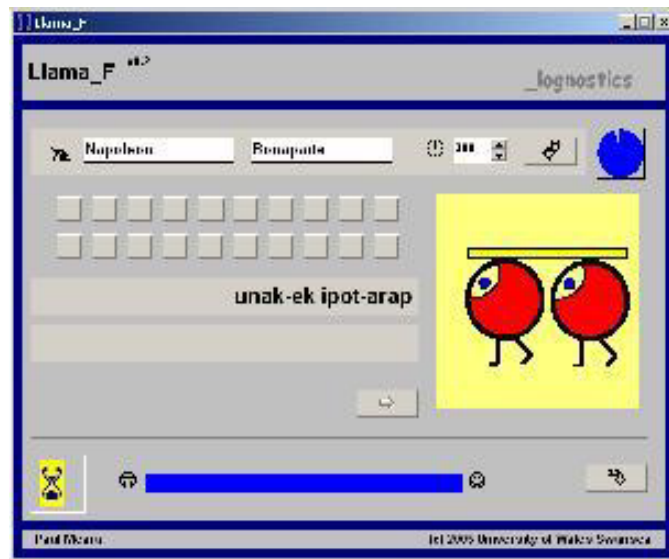


Figure 1. LLAMA_B User Interface

3.7.5.2 LLAMA_F

The LLAMA_F is a grammar inferencing test. It shows a series of pictures depicting shapes and objects and a sentence in an artificial language, which describes each picture. During the study phase, students have five minutes to work out the grammatical patterns of the artificial language. They must click on the button and a picture and a sentence that describes it will be displayed. Unlike the LLAMA_B, in the LLAMA_F, notes can be taken. During the testing phase, the programme displays a picture and two sentences; one is correct and the other contains an error. The test taker must select

the correct sentence. Five points are given for a correct answer and five points are deducted for a mistake. This reduces the impact of guessing. The scores can range from 0-100. 0-15 is a very poor



score, 20-45 is an average score, 50-65 is a good score, and 75-100 is a very high score. The user interface of the LLAMA_F is shown at figure 2.

Figure 2. The LLAMA_F User Interface

3.8 Pilot Study

The purpose of including a pilot study was to test and improve the instruments that would be used in the main study, and to discover any unwanted issues – such as participant attrition due to the class attendance policy, whether the instruments were too easy or difficult, and to test the reliability of the questionnaire. Discovering if any of the research instruments had drawbacks that could be improved would be of paramount importance for the large-scale study.

3.8.1 Pilot Study Methodology and Experimental Set-up

The pilot study was conducted at the same institution with the same level of participant as the main study. In the pilot study, there were 68 participants, 43 females and 25 males. Most were L1 speakers of Arabic (N = 61), while a small minority were L1 speakers of Urdu (N = 7). Regarding the language spoken with friends, the majority (N= 43) spoke Arabic and English, 13 spoke only Arabic, 9 spoke English and 2 spoke Urdu and English and 1 spoke only Urdu. A total of 23 participants reported having spent between 1 and 3 months in an English-speaking country, with most having spent time in England. The majority of students came from a school where English was the medium of instruction (N= 58) and a small minority (N=8) came from a school where Arabic was the medium of instruction. The majority of students reported that they spoke Arabic and English at home (N= 33), 29 students only spoke Arabic at home; 5 students spoke Urdu, and one student spoke only English at home.

The participants were given the Oxford Quick Placement Test to find out their English proficiency levels, and were scored between 14 and 58 points. Following the Oxford Quick Placement Test's interpretation of results, their ALTE (The Association of Language Testers in Europe) levels ranged from beginner to very advanced. The students were randomly split into four groups by assigning each student a number between 1 and 4, so that there were students from different experimental groups in each intact class. The different groups consisted of three treatment groups: a group receiving direct feedback, a group receiving indirect feedback and a group receiving metalinguistic feedback. There was also a control group. The experimental set-up, timeline and instruments are shown in Table 4 and followed a similar experimental set-up as the main study. All of the research was conducted during class periods and took six sessions.

Table 4. Experimental Set-up and Instruments

Session 1 Week 1	Session 2 Week 2	Session 3 Week 3	Session 4 Week 4	Session 5 Week 5	Session 6 Week 6
Students chose topic for essay from list of 20 topics (5 minutes)	LLAMA_B aptitude test (30 minutes to set up and complete)	Oxford Quick Placement Test (30 minutes)	Pre-test writing task persuasive essay 1 (20 minutes)	Treatment (Corrective feedback given to CF groups and essays returned) Students look at the feedback and ask questions if they do not understand. (10 minutes)	Post-test writing task persuasive essay 2 (20 minutes)
Language background questionnaire (15 minutes)	LLAMA_F aptitude test (30 minutes to set up and complete)				
Attitudes questionnaire (15 minutes)					

The pilot study and the main study have many similarities, but there were also differences regarding the experimental set-up. In the pilot study, students did not write a text re-write, received fewer feedback sessions, and also wrote fewer essays overall. In the main study, compared to the pilot study, there was an extra initial essay and WCF session, whose data was not included in the study, but would allow the participants to become used to the different forms of feedback and to give them more exposure to the WCF. Furthermore, the pilot study lacked the dedicated feedback workshops that were included in the main study. A further difference was that in the pilot study, the participants' essays were returned to them a week later in the fifth session and they were given ten minutes to look at them. They were encouraged to ask questions if they did not understand the feedback, before leaving the classroom - meanwhile, in the main study, a different process was used. In the pilot study,

students in the control group were let out of class early as they did not receive any feedback, whereas in the main study a different approach was adopted.

The participants' essays were then coded for Fluency, Accuracy and a range of complexity measures (Table 5). The operationalisation of CAF are controversial, so a balanced approach drawing from the most up-to-date research and looking at what measures other researchers have used was necessary. The justification for choosing multiple measures for complexity was that when measuring complexity, one measure is not enough (Bulté and Housen, 2012), and each dimension requires one or more different measures appropriate for that dimension.

Table 5. CAF Measures, their Operationalisation and Other Studies that Have Used the Measure

Measure and other studies that have used this measure	Operationalisation
Fluency	total number of words in the text
Accuracy	proportion of error free t-units (a <u>main clause</u> plus any <u>subordinate clauses</u> that may be attached to it)
Overall syntactic complexity Bygate (2001); Ishikawa (2007); Kawauchi (2005); Mochizuki & Ortega (2008)	Mean length of T-units (number of words/number of t-units)
Sentential syntactic complexity as used in the studies of Ellis & Yuan (2005); Ishikawa (2007); Iwashita et al. (2008); Kawauchi (2005); Kuiken et al. (2005); Kuiken & Vedder (2007); Mochizuki & Ortega (2008); Sangarun (2005); Sercu et al. (2006); Yuan & Ellis (2003); Adel & Alwi (2014).	Clauses per t-unit
Sentential syntactic complexity Ali Mohammad Fazilatfar et al. (2014); Ishikawa (2007); Iwashita et al. (2008); Kuiken et al. (2005); Kuiken & Vedder (2007); Adel & Alwi (2014).	Dependent clause ratio (number of dependent clauses / number of clauses)
Subsentential syntactic complexity Ishikawa (2007);	Mean length of clause in words
Lexical diversity Calculated using the L2 syntactic complexity analyser (SCA) Ai & Lu (2010); Lu (2012) Ellis & Yuan (2004); Yuan & Ellis (2003); Chan et al. (2015); Lu & Ai (2015); Lorenzo and Rodriguez (2014); Wang and Slater (2016).	Type token ratio (TTR) (the ratio of different words to total words used)
Lexical diversity Calculated using the L2 syntactic complexity analyser (SCA) Ai & Lu (2010); Lu (2012); Ellis & Yuan (2004); Yuan & Ellis (2003); Chan et al. (2015); Lu & Ai (2015); Lorenzo and Rodriguez (2014); Wang and Slater (2016).	Mean Segmental Type token ratio (TTR) (For each segment the TTR is calculated and then averaged for all segments.)

Lastly, when all the data was collected, it was entered into SPSS version 24.

3.8.2 Results of the Pilot Study that Informed Changes to the Main Study

The first issue the pilot study addressed was to see how the CAF measures correlated and if the large number of CAF measures could be reduced. The pilot study originally included six measures of complexity; however, correlations of the CAF measures and a principal component analysis (PCA) were run as pre-checks, and the results of this can be seen in Appendix E. A 4 factor solution was deemed the most appropriate as unlike the three factor solution that was also tried, the four factor solution would allow loadings of fluency and accuracy on a factor by themselves, unlike the three factor solution where they loaded negatively on the same factor as the measures of lexical diversity. The loadings of the variables onto the factors are bolded in Table 33 in Appendix E. The four-factor solution was run, as the fourth eigenvalue was close to one, and the scree plot did not suggest a clear break after the three factors.

The CAF measures were then factor analysed using principal component analysis with Varimax (orthogonal) rotation. The analysis yielded four factors explaining a total of 88.959% of the variance for the entire set of variables. Factor 1 was named Syntactic Complexity, due to the high loadings by the following items: Syntactic complexity (mean length of t-unit), Syntactic complexity sentential (clauses per t-unit) and Syntactic complexity sentential (dependent clause ratio). This first factor explained 31.341% of the variance. Factor 2 was named Lexical Diversity due to the high loadings by the following items: Lexical Diversity (TTR) and Lexical Diversity (mean segmental TTR). This second factor explained 22.772% of the variance for the entire set of variables. Factor 3 was named Subsentential Syntactic Complexity due to the high loading of subsentential syntactic complexity (mean length per clause). This factor explained 18.735% of the variance for the entire set of variables. Factor 4 was named Accuracy due to the high loading of accuracy. This factor explained 16.101% of

the variance for the entire set of variables. The four-factor solution can be seen at Appendix E along with the scree plot.

The correlations and PCA showed that syntactic complexity sentential (dependent clause ratio) and Lexical Diversity (TTR) had very high correlations, and were basically measuring the same thing as two other measures of complexity: Lexical Complexity (clauses per t-unit), and Lexical Diversity (mean segmental TTR) - thus eliminating them would be appropriate. Therefore, syntactic complexity sentential (dependent clause ratio) and Lexical Diversity (TTR) were eliminated from further analyses, and Lexical complexity (clauses per t-unit) and Lexical Diversity (mean segmental TTR) were kept. Mean segmental TTR is less sensitive to text length than TTR (Johnson 1944) as well as the fact that Mean Segmental TTR and clauses per t-unit correlated better with proficiency than the measures that were eliminated.

The second issue the pilot study examined was whether the instruments were sound or would need changing for the main study. To accomplish this, preliminary data checks were carried out on the pilot study data. 68 students were present at the start of the study; however, certain students were absent during the writing tasks and, therefore, the number of students present for all sessions was 58, when taking attrition into account. The descriptive statistics for the LLAMA B and F aptitude test are presented in Table 6 and show that this instrument was working as intended:

Table. 6 Descriptive Statistics for the LLAMA B and F

Test Type	N	M	SD	Min	Max
LLAMA B	58	54.49	20.39	10	100
LLAMA F	58	42.87	24.80	0	90

First, reliability analyses for the attitudes questionnaire that used a Likert scale, were carried out using Cronbach's alpha. The alpha value for the attitudes questionnaire was .44. Therefore, the questionnaire was deemed to not be sufficiently reliable, so could not be used to generate a summary score for a variable representing 'attitude'. Thus, the questions could only be used individually when analysing the data. This questionnaire needed to be changed for the main study, so the questionnaire was analysed at item level and some questions were modified or eliminated based on the reliability score.

The assumptions of Normality of Distribution for all variables were checked by using the Kolmogorov-Smirnov and Shapiro-Wilk test (Appendix F). The results of both tests with the alpha level set at 0.05 show that only 8 variables out of 40 had a normal distribution, so normality could not be assumed, and non-parametric statistics were used. Due to these results, it was deemed necessary to increase the number of participants in the main study to try to achieve normality and to be able to use more powerful parametric statistics.

In order to find out if the students in the four groups began the study with similar writing proficiency, and due to the non-normal distribution of the data, the Kruskal-Wallis test was used (Appendix G). The results showed that the four groups were similar, with no significant difference between the groups, regarding all CAF measures, using an alpha level of 0.05. These results suggest that all treatment groups had a comparable accuracy, complexity, fluency, and L2 proficiency level at the beginning of the data collection. Thus, it can be assumed that any differences found later on in the study are not related to initial differences between the treatment groups. The descriptive statistics can be seen in Tables 7, 8, 9 and 10 below.

Table 7. Descriptive Statistics for Fluency

Treatment groups	N	M (pre-test)	SD (pre-test)	M (post-test)	SD (post-test)
Direct	12	213.46	58.80	183.46	68.52
Indirect	15	182.65	77.30	194.88	74.33
Metalinguistic	15	186.81	77.80	160.30	68.80
Control	16	193.94	83.66	184.13	70.10
Total	58	193.10	74.743	180.81	69.92

Table 8. Descriptive Statistics for Accuracy

Treatment groups	N	M (pre-test)	SD (pre-test)	M (post-test)	SD (post-test)
Direct	12	.2510	.16172	.1992	.14619
Indirect	15	.1531	.16632	.2644	.20120
Metalinguistic	15	.2713	.16887	.2544	.21404
Control	16	.2600	.21288	.3213	.24719
Total	58	.2255	.18170	.2638	.20791

Table 9. Descriptive Statistics for Syntactic Complexity Overall

Treatment groups	N	M (pre-test)	SD (pre-test)	M (post-test)	SD (post-test)
Direct	12	18.2591	3.97763	19.7528	10.54561
Indirect	15	20.4527	6.21468	23.1771	8.52069
Metalinguistic	15	21.4281	13.85310	18.6541	7.59931
Control	16	19.7891	6.54836	18.2392	3.85445
Total	58	20.1227	8.67508	19.9178	7.77750

Table 10. Descriptive Statistics for Lexical Diversity

Treatment groups	N	M (pre-test)	SD (pre-test)	M (post-test)	SD (post-test)
Direct	12	.6755	.25626	.7273	.19713
Indirect	16	.7727	.04464	.7620	.03610
Meta linguistic	16	.7081	.22275	.6363	.29915
Control	16	.5613	.31616	.5606	.33195
Total	60	.6781	.24044	.6652	.25730

As well as leading to a change in the instruments in the main study, the pilot study also informed changes in the main study's research design. When examining if corrective feedback leads to an increase in the accuracy, complexity, and fluency of student writing on new tasks, compared to no feedback, the pilot study found that there were no significant differences between the feedback groups and the control group. Due to the distribution of the data being non-normal the Friedman test was used to compare the effect of corrective feedback on students' writing fluency, accuracy and complexity. The results can be seen in Appendix H. This lack of effect of WCF feedback on the CAF

measures in the pilot study, led to increasing the number of written corrective feedback sessions students received in the main study, to see if doing this would lead to an effect. In the pilot study, participants only received WCF once. This may have caused problems in all treatment groups, but especially in the group receiving indirect feedback or those receiving metalinguistic information as a form of error correction. Even though the participants were instructed on how to use and interpret the error codes, it may be possible that they might not have had enough opportunity to become accustomed to this type of WCF. It is therefore possible that the effects of metalinguistic as well as the other forms of WCF would have been greater if a design with more than one treatment session was conducted. The indirect feedback may have also proved problematic for the students as they needed to try to discover what their mistakes were, and they may have been reluctant to do it in their own time, so having a dedicated post-feedback session would give these students in-class time to try to work out what their mistakes were. Not only the indirect feedback group, but all feedback groups may benefit more from the WCF if there is a post-feedback support session where they have time to review their feedback and ask the teacher questions about it. It may be that the students did not pay attention to the feedback and notice the feedback. The main study's research methodology was thus adapted with these issues in mind.

3.8.3 Conclusions of Pilot Study

The pilot study aimed to test and assess the research design, so improvements could be made for the main study. The results of the pilot study showed that in general, corrective feedback did not lead to an increase in the fluency, accuracy, or complexity of student writing on new tasks. The results using nonparametric statistics did not reveal any significant difference between the treatment and the control groups when examining gain scores, which shows that unfocused feedback had no impact on accuracy, fluency, or complexity. There are a multitude of possible reasons why the groups receiving WCF had no statistical difference from the control group regarding the accuracy, fluency and complexity measures - and this could be due to the experimental design of the study. In summary, the

pilot study showed there were many changes regarding the research instruments and design that needed to be made for the main study.

3.9 Complexity, Accuracy and Fluency Measures

In the main study, the CAF measures used to measure proficiency, their operationalisation, the programme used to calculate them, and examples from the data set are explained in the subsequent paragraphs. The participants' essays were measured for the following CAF variables that can be seen in Table 11.

Table 11. CAF Measures, their Operationalisation, the Way They Were Calculated and Examples from the Data Set

Measure and studies that have used this measure	Operationalisation and coding conventions	How they were calculated	Examples from data set
Fluency Chandler (2003) Wang & Salter (2016)	Total number of words in the text	L2 Syntactical Complexity Analyzer (LCA) Lu (2010)	Children should not be given an iPhone. Its not good for these. = 12 words Today is rainy. = 3 words
Accuracy Wolfe-Quintero et al. (1998) Jiang (2013) Abdollahzadeh & Kashani (2012)	1. Proportion of error free t-units (a <u>main clause</u> plus any <u>subordinate clauses</u> that may be attached to it) = number of error-free T-units divided by the total number of T-units A Clause was defined as a unit containing a subject and a finite verb, which includes independent, nominal, adverbial, and adjective clauses, but not non-finite verb phrases. A Dependent clause was defined as a finite nominal, adverbial, and adjective clause. 2. Errors per 100 words	Calculated by hand	Children should not be given an iPhone. Its not good for these. = 0.5 error free t-units. Today is rainy, but tomorrow will be sunny. = 2 error free t-units.
Storch (2009) Errors per total words	3. Lexical errors per 100 words		I look television every day.
Tai, H.-Y. (2015). Errors per total words	4. Grammatical errors per 100 words		I like to eaten all the time.
	5. Spelling and punctuation errors per 100 words		The boy. dud his hemwurk..

Measure and studies that have used this measure	Operationalisation and coding conventions	How they were calculated	Examples from data set
Complexity by phrasal elaboration; Lu and Ai (2015); Lorenzo & Rodriguez (2014); Wang & Slater (2016)	Complex nominals per clause (number of complex nominals divided by the number of clauses) A complex nominal is defined as: a) nouns with modifiers b) nominal clauses c) gerunds and infinitives that function as subjects	L2 Syntactical Complexity Analyzer (LCA) Lu (2010)	Children should not be given an iPhone. Its not good for these. = 0 complex nominal per clause He is not a real man. = 1 complex nominal per clause
Sentential syntactic complexity as used in the studies of R.Ellis & Yuan (2004),(2005); Ishikawa (2007); Iwashita et al. (2008); Kawauchi (2005); Kuiken et al. (2005); Kuiken & Vedder (2007); Mochizuki & Ortega (2008); Sangarun (2005); Sercu et al. (2006); Yuan & Ellis (2003); Adel & Alwi (2014). Wang and Slater (2016)	Clauses per t-unit	L2 Syntactical Complexity Analyzer (LCA) (Lu, 2010)	Children should not be given an iPhone. Its not good for these. = 1 clauses per t-unit The students struggled writing their essays, and asked the teacher for help. = 2 clauses per t-unit
Lexical diversity vocd-D Schmid & Jarvis (2014); Treffers-Daller (2013) Sadeghi & Dilmaghani (2013)	Voc-d is based on the predicted decline of the TTR, as the sample size increases. This mathematical curve is compared with empirical data from a text sample. For calculating Voc-D, information from the whole text sample is used. The higher the value of Voc-D, the higher the lexical diversity.	Coh-Metrix McNamara, Graesser, McCarthy, & Cai (2014)	The program needs a certain amount of text that would not fit in the table to run a suitable measure for voc-D

Measure and studies that have Used this measure	Operationalisation and coding conventions	How they were calculated	Examples from data set
<p>MTLD (Measure of Textual Lexical Diversity) Schmid & Jarvis (2014); Crossley, et al. (2009) Šišková (2012)</p>	<p>The mean length of sequential word strings in a text that maintain a given TTR value where during the calculation process, each word of the text is evaluated sequentially for its TTR. Calculated as the average number of running words in a text that remain above a certain type-token ratio (usually .72) (Schmid & Jarvis, 2014)</p>	<p>Coh-Metrix McNamara, Graesser, McCarthy, & Cai (2014) McCarthy & Jarvis (2010)</p>	<p>The program needs a certain amount of text that would not fit in the table to run a suitable measure for MTLD</p>

CAFs are ways of measuring language performance; however, there are controversies regarding CAF research. Some researchers note that reported research may not contribute to the accumulated knowledge because the findings cannot be compared (Norris & Ortega, 2009). Others propose examining CAF research from a dynamic or complex systems point of view (Larsen-Freeman, 2009), but, most researchers have come to a consensus that using different ways of measuring CAF and the different ways CAF are defined, are areas that researchers need to tackle. The operationalisation of CAF are controversial, so a balanced approach drawing from the most up-to-date research and taking into consideration the differing points of view is necessary.

3.9.1 Fluency

Fluency has been operationalised in different ways and in this study, the measure of total number of words in the text was chosen. This was done because numerous studies have used this way to measure fluency, including Chandler (2003) and Wang & Salter (2016). The programme used for measuring fluency was the L2 Syntactical Complexity Analyzer (LCA) (Ai & Lu, 2010).

3.9.2 Accuracy

Accuracy is usually regarded as the most straightforward construct of CAF and refers to the amount of conformity to certain language usage norms. In this study it was operationalised as proportion of error free t-units, following numerous studies that have measured it in this way, including Wolfe-Quintero et al. (1998); Jiang (2013) and Abdollahzadeh & FardKashani (2012). An additional measure of errors per total words, following studies such as Storch (2009) and Tai, H.-Y. (2015) was also used. Errors per total words was also chosen as it is

often used in CAF studies to measure accuracy and would be a complimentary measure. Both of these accuracy measures were calculated by hand.

3.9.3 Complexity

Complexity can be measured in a myriad of ways, but in this study it was broken down into general syntactic complexity, complexity by phrasal elaboration, complexity by subordination, sentential syntactic complexity and lexical diversity. The justification for choosing these measures is that research on the construct of syntactic complexity has shown it to be a multi-dimensional construct. Therefore, when measuring complexity, one measure is not enough and each dimension requires one or more different measures appropriate for that dimension. For example, Bulté and Housen (2012), Lu (2010), and Norris and Ortega (2009) recommend that the following measures should at least be incorporated: general syntactic complexity, complexity by subordination, and complexity via phrasal elaboration.

Lexical diversity is the variety of unique words in a text in relation to number of words. When a text has a high score in lexical diversity it has diverse language and when words are frequently repeated in a text, lexical diversity is low. A way to measure lexical diversity is type-token ratio (TTR), but since the measurement of TTR is influenced by text length, using the measures of voc-D and the MTL D index (Measure of Textual Lexical Diversity) this problem can be solved by using estimation algorithms (McNamara et al., 2014). If a text has a higher score of MTL D or Voc-D it is more likely to be more difficult, complex, and more advanced (McCarthy & Jarvis 2010). Schmid and Jarvis, (2014) recommend that researchers should consider using MTL D, vocd-D (or HD-D) in their research, rather than any single index, as Lexical Diversity can be measured in a variety of ways.

Although in the pilot study, the complexity measures were reduced by way of a PCA as some were measuring the same thing, in the main study, the complexity measures used were

more sophisticated and measured different types of complexity; and thus multiple measures of complexity were included.

Two automated measurement programmes were used to measure complexity in this study. The rationale for using them was that calculating all the essays by hand would have been too time consuming and also, an accurate measure of lexical diversity could only be achieved by using automation due to the complex mathematical calculations involved. The programme used for measuring the syntactic and sentential complexity measures was the L2 Syntactical Complexity Analyser (LCA) (Ai & Lu, 2010); and Coh-Metrix (McNamara et al., 2005; Graesser & McNamara, 2011; McNamara et al., 2014). Previous studies that used the L2 Syntactical Complexity Analyser (LCA) and Coh-Metrix in the last five years have all investigated L2 learner writing in English. 47 articles have used Coh-Metrix, but not all included the complexity component. Sixteen studies used LCA, and four used both programmes (Polio & Hyung-Jo, 2018) - and this growing body of research adds to the acceptability of using automated measures of complexity.

The L2 syntactic complexity analyser (L2SCA) (<http://aihaiyang.com/software/l2sca>) was developed by Xiaofei Lu, and detailed in Lu (2010). It is available free online and was developed for what Lu classifies as advanced writers of English. The programme was tested using the Written English Corpus of Chinese learners, that includes 3,554 essays that were written by university students who were English majors in China. The programme was developed to automate the syntactic analysis of L2 English texts using 14 measures (Lu, 2010).

This study also used the computational tool Coh-Metrix (McNamara et al., 2005; Graesser & McNamara, 2011; McNamara et al., 2014). It draws together a range of techniques and resources that have been developed within the field of Natural Language Processing launched in 2003 at the University of Memphis. McNamara et al. (2014) note that “It is arguably the broadest and most sophisticated automated textual assessment tool currently

available on the Web” (p. 2) and was originally developed as a tool for automatically assessing text readability. The version used in this study was the online web tool version of Coh-Metrix (www.cohmetrix.com). It employs a series of databases that provide a wide range of statistically referenced linguistic information (108 different categories). The two categories it measured in this study were lexical diversity measures comprising of MTLN and Voc-D.

3.10 Procedure

The study followed a timeline that can be seen in Table 12.

Table 12. Timeline of the Experiment

Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7	Session 8	Session 9	Session 10	Session 11
Week1	Week 2	Week 3	Week 4	Week 4	Week 5	Week 5	Week 6	Week 7	Week 13	Week 14
Students chose topic for essay from list of 20 topics (5 minutes)	LLAMA_B aptitude test (30 minutes to set up and complete)	Oxford Quick Placement Test (30 minutes)	Practice essay writing task persuasive (20 minutes)	Corrective feedback given back to students and WCF feedback workshop	Pre-test Writing task Essay 1 (20 minutes)	Treatment (Corrective feedback given to CF groups and essays returned) (5 minutes)	Post-test writing task persuasive Essay 1 re-write of text (20 minutes)	Feedback given(Corrective feedback given to CF groups and essays returned)	Delayed Post-test writing task persuasive essay 2 (20 minutes)	Feedback given(Corrective feedback given to CF groups and essays returned) (5 minutes)
Language background questionnaire (15 minutes)	LLAMA_F aptitude test (30 minutes to set up and complete)					Feedback workshop (40 minutes) students try to correct all the mistakes on their essays. The teacher will be present to answer any questions	Feedback workshop (40 minutes) students try to correct all the mistakes on their essays. The teacher will be present to answer any questions (post-feedback session) collect text revisions	Feedback workshop (40 minutes) students try to correct all the mistakes on their essays. The teacher will be present to answer any questions (post-feedback session) collect text revisions		Feedback workshop (40 minutes) students try to correct all the mistakes on their essays. The teacher will be present to answer any questions (post-feedback session) collect text revisions
Attitudes questionnaire										

In Session 1, the students were given an online questionnaire, using the SurveyMonkey platform to elicit which essay topic would be the most interesting from a given list and had the intention of allowing students write about something they were interested in, rather than the researcher choosing a topic.

In the same session, all participants then filled in the language background questionnaire that asked them about their English language learning experiences and how they use, or have used English in their day-to-day life (Appendix C). Participants then filled in the attitudes toward WCF questionnaire (Appendix D). The instruction was given to students orally to make sure that each participant could understand how to respond to the items. The participants were also made aware that there were no right or wrong answers, and were only asked to report their own opinions.

In session 2, the researcher met with the participants and they took two sub-tests of the LLAMA aptitude test (Meara, 2005), sub-test B (a vocabulary learning task) and F, (a grammatical inferencing task). The LLAMA tests were taken in a computer lab at the university where the research took place, and the researcher noted down the scores on an Excel sheet.

Session 3 consisted of the participants taking the paper-based version of the Oxford Quick Placement test, a language proficiency test.

In Session 4, students wrote the first argument/persuasion essay. The students chose an argument/ persuasion topic from the list, but the data from the essay was not used for the research. The rationale behind having this first essay, and not including the data in the study, was to allow the participants to become used to the different forms of feedback and to give them more exposure to the WCF, thus allowing the corrective feedback to have a larger effect. When the essays were written, corrective feedback was given to the participants in the treatment groups. The control group did not receive feedback and were given an alternative task.

In Session 5, students had a post-feedback workshop session for 30 minutes where they were asked to correct their essays. Students in the direct group could ask questions if they did not understand, and since the corrections were already provided, they had the least correcting to do. Students in the indirect group could ask questions, but the answers the researcher gave could only be very simple, such as “yes” and “no”. This was intentional, so as not to turn the answer into a form of direct feedback, and the correct form could not be given to students. The students had to try to correct the mistakes they had made on their papers and were allowed to use the Internet, or ask their friends. Students in the metalinguistic group tried to correct their essays using the error codes and error code legend given to them. They were allowed to ask questions, but the correct form could not be given to the students, to make sure the feedback did not become a form of direct feedback. Students in the control group, however, were given feedback on content and set another activity which was to read a passage on current events that had no connection to the study, and would not give them any advantage over the feedback groups regarding the grading on their coursework.

In Session 6, students wrote the first argument/persuasion essay (pre-test) that would be used in the research. They were assigned a topic from the most popular essay topics on the list. The topics were assigned to make sure that participants received a different prompt to the one they wrote about the first time. The researcher then put WCF on the essays of the treatment groups and the control group did not receive feedback.

The participants’ essays were printed out and returned to them a week later in Session 7 with the feedback, and they were yet again given a post-feedback workshop in the same manner as in the fifth session.

Session 8 was conducted a week later, and the participants were told to bring back their original essay with the corrections they had made during the feedback workshop session. They were once more instructed to re-write the essay as well as they could. Unlike other studies

where the students were allowed to look at the corrections and copy them word by word while completing the re-write, in this study the students were given 15 minutes to look at their essays and double-check the corrections they had made in the previous session. The essays were then taken away and they were given 30 minutes to re-write them as best they could, to see if any of the feedback would be retained in the short-term. The re-write (re-test) was then collected by the researcher.

In session 9, the students' text re-writes were returned with corrective feedback and they were yet again given a post-feedback workshop.

Five weeks later, Session 10 consisted of the students writing the final essay (Delayed post-test). This essay was an argument/persuasive essay and students could choose a topic from the list of the most popular topics (Appendix B), but it could not be an essay on a subject they had previously written about. The purpose of the (post-test) was to examine if the students could improve their writing in the context of a new topic. The final essay was then collected and the essay was coded.

Session 11 was conducted in week 14, and the students' texts were returned with corrective feedback. Feedback needed to be given as the students request feedback on every essay they write. They were once again given a post-feedback workshop.

When all the data was collected, preliminary data checks needed to be carried out.

3.11 Inter-rater Reliability Analysis and Data Processing

The study used a second marker who coded 20% of the scripts for inter-rater consistency for the overall accuracy measurement. Fluency and complexity used automated measures, thus a second marker using the same programmes would return the same measurements, therefore,

only accuracy which was calculated by hand needed a second marker. The percentage of agreement between the two markers was 95%.

3.12 Preliminary Data Checks

Before proceeding with the statistical tests, preliminary data checks needed to be carried out. First, the assumptions of Normality of Distribution of the instruments were checked by using the Kolmogorov-Smirnov test (Appendix I). The results of the tests of normality are shown in Tables 38-40. The results with the alpha level set at 0.05 show that only 10 variables out of 36 have a normal distribution, so normality cannot be assumed, and non-parametric statistics were used.

Next, the reliability analysis of the Attitudes Questionnaire was conducted using Cronbach's alpha. The alpha value for the attitudes questionnaire was .692. With question 16 ("I feel anxious about receiving corrective feedback") deleted the alpha coefficient was .712. Thus, question 16 was deleted and the questionnaire could be used to generate a summary score for a variable representing 'attitude'. Reliability for the LLAMA B and F tests was not assessed due to the fact that the version 2.0 of the LLAMA test that is freely available online to researchers, does not give scores at item level. The new version 3.0 of the LLAMA test does, however, but was not available at the time the research was conducted.

Following this, checks then were then run on a random sample of thirty of the student essays to examine if punctuation errors would affect the automated complexity measures using the L2SCA. If a text had multiple commas, this could cause a higher reading in clauses per unit for instance. However, due to the proficiency level of the students in the study being rather high, including multiple commas in a grammatically incorrect way would be very rare; however, to verify that this kind of anomaly did not occur often, fifteen of the thirty essays

were chosen randomly, and all punctuation mistakes were corrected. These corrected essays were again run through the L2SCA. The remaining fifteen essays were left as they were without correcting punctuation errors. A paired samples t-test showed that for the complexity measures there was no significant difference in the scores for non-corrected and corrected essays using an alpha level set at 0.05. The results for mean length of t-unit was $p = .117$; complex nominals per clause was $p = .567$; and clauses per t-unit was $p = .086$.

Similar checks were run on another random sample of essays to check if spelling mistakes inflated Voc-D and MTL D, because an automated program would count a misspelt word as a new word, and thus a text would appear to be more lexically diverse. With this in mind, a random sample of thirty student essays was selected and fifteen of them were randomly chosen for correction of all spelling errors. However, due to the fact that spelling mistakes are very minimal at the proficiency level of the participants in this study, there were not many spelling mistakes to correct. These corrected essays were again run through the L2SCA. The remaining fifteen essays were left as they were without correcting spelling errors. The corrected fifteen texts and other fifteen uncorrected texts were then compared using paired samples t-tests, but the results were non-significant using an alpha level of 0.05. Voc-D: $p = .112$, MTL D: $p = .137$.

Next, pre-test CAF measures were correlated with proficiency measured using the Oxford Quick Placement Test. As some CAF measures had normally distributed data (Appendix I), but others did not, correlations using non-parametric Spearman correlation were used. When both variables were normally distributed, Pearsons' R was used instead and is denoted by the letter 'r'. The results can be seen in Table 13 below.

Table 13. Correlations: (Spearman) Pre-test CAF Measures and Proficiency

	Oxford Quick Placement Test	Total Number of words	Proportion of error free t-units	Errors per/100 words	Lexical Errors /100	Grammatical Errors/100	Spelling and Punctuation/100	Mean Length of t-units	Complex Nominals/clause	Clauses /t-unit	Voc-D
Total Number	.125										
Of words	p= .141										
Proportion of	-.254**	.002									
Error free t-units	p= .003	p= .799									
Errors/100 words	-.298**	-0.96	-.739**								
	p=000	p= .260	p= .000								
Lexical Errors/100	-.219**	.043	-.271**	.451**							
	p=010	p= .617	p= .001	p= .000							
Grammatical Errors /100	-.174*	-1.02	-.456**	.507**	.279**						
	p= .041	p= .234	p= .000	p= .000	p= .001						
Spelling and Punctuation/100	-.188*	-0.71	-.639**	.832**	.121	.128					
	p= .027	p= .408	p= .000	p= .000	p= .159	p= .135					
Mean Length of t-units	-.032	.087	-.398**	.204*	.011	.162	.208*				
	p= .708	p= .313	p= .000	.016	p= .894	p= .057	p= .015				
Complex Nominals/ Clause	-.045	-.065	r = -.127	.129	.025	.084	.090	.339**			
	p= .601	p= .452	p= .137	p= .130	p= .773	p= .329	p= .292	p= .000			
Clauses/ t-unit	-.136	.108	-.425**	.220**	.058	.123	.262*	.801**	-.054		
	p=.111	p= .208	p= .000	p= .009	p= .500	p= .149	p= .002	p= .000	p= .528		
Voc-D	.116	.158	r = -.035	.109	.033	.023	.108	-0.36	r = -.044	-.052	
	p=.177	p= .063	p= .682	p= .204	p= .699	p= .785	p= .209	p= .672	p= .608	p= .548	
MTLD	.034	.037	-.050	.057	-.018	-.014	.073	-.005	.078	-.062	.793
	p= .691	p= .665	p= .557	p= .508	p= .835	p= .637	p= .393	.952	p= .364	p= .467	p= .000

** p < .01. * p < .05.

The correlation matrix (Table 13) shows that for the pre-test, students' proficiency measured using the Oxford Quick Placement test had significant negative correlations with all the accuracy measures (the proportion of error-free t-units was reverse coded). This means the higher the proficiency of the student, the less mistakes they made when writing essays and accuracy should correlate with the Oxford Quick Placement test. The fluency measure did not have significant correlations with proficiency and none of the complexity measures had significant correlations with proficiency, but this is because fluency and complexity are not represented in the version of the Oxford Quick Placement test taken by the students.

To see if it would be possible to reduce the number of variables for the main study, inter-correlations of the CAF measures were performed. Unsurprisingly, the five accuracy measures correlated significantly with each other, with the proportion of error free t-units and errors per 100 words having the strongest negative correlation of $-.739$. This is understandable since spelling and punctuation make up part of the errors that would be included in total errors per 100 words, and thus they are in fact sub-scores of the measure of errors per 100 words.

One of the accuracy measures (Proportion of error free t-units which was reverse coded), was significantly negatively correlated with two of the complexity measures (Mean length of t-units and clauses per t-unit). This shows that students who produced more accurate writing also wrote less complex writing, although these correlations were weak. The other accuracy measure (Errors per 100 words) has a significant positive correlation with a complexity measure, so students who wrote less accurate writing also wrote more complex writing, but this correlation was very weak. Spelling and punctuation errors also have a positive significant correlation with some of the complexity measure of mean length of t-units, showing that students who write more accurate writing also write more complex writing. Some of the syntactic complexity measures (Mean length of t-units and Complex nominals per clause and Mean length of t-units) have strong correlations with each other, which is unsurprising, since

they all measure syntactic complexity. Voc-D and MTLD were also very strongly correlated as they are both measures of lexical density. All other correlations were non-significant. These analyses show, due to their inter-correlation of the CAF measures, that factor analysis would be a useful procedure to try to reduce the number of variables for the main study.

3.12.1 Principal Component Analysis

The pre-test CAF measures were then factor analysed using principal component analysis with Varimax (orthogonal) rotation. Table 14 shows the principal component analysis for a 5 factor solution, Figure 3 shows the scree plot and Table 15 shows the rotated component matrix.

Table 14. Principal Component Analysis 5 Factor Solution

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.361	30.550	30.550	3.361	30.550	30.550	2.811	25.550	25.550
2	1.839	16.722	47.273	1.839	16.722	47.273	1.978	17.985	43.535
3	1.479	13.448	60.720	1.479	13.448	60.720	1.836	16.687	60.221
4	1.132	10.287	71.007	1.132	10.287	71.007	1.165	10.592	70.814
5	1.101	10.013	81.020	1.101	10.013	81.020	1.123	10.207	81.020
6	.777	7.059	88.079						
7	.596	5.415	93.495						
8	.366	3.327	96.822						
9	.200	1.815	98.636						
10	.149	1.358	99.995						
11	.001	.005	100.000						

Extraction Method: Principal Component Analysis.

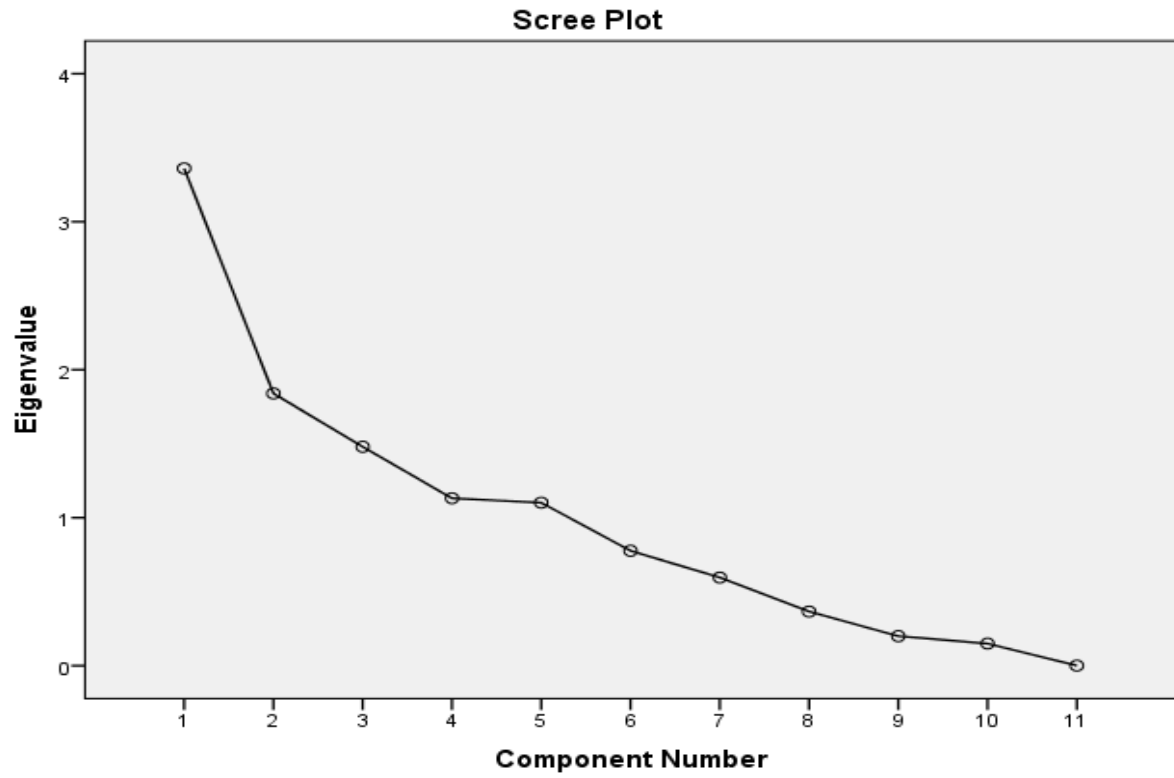


Figure 3. Scree Plot

Table 15. Rotated Component Matrix

	Rotated Component Matrix ^a				
	Component				
	1	2	3	4	5
Total number of words in the text pre test	-.170	.242	.070	-.256	.759
Proportion of error free t-units pre test	-.741	-.420	.002	-.009	.000
Errors per 100 words pre test	.949	.196	.071	-.004	-.046
Lexical errors per 100 words pre test	.520	-.170	-.089	.248	.614
Grammatical errors per 100 words pre test	.729	-.137	.039	.227	.124
Spelling and punctuation errors per 100 words pre test	.691	.384	.100	-.206	-.349
Mean length of T-units pre test	.145	.841	.021	.354	.086
Complex nominals per clause pre test	.063	.134	.021	.893	-.094
Clauses per t-unit pre test	.151	.885	-.058	-.079	.050
voc-D pre test	.059	-.014	.951	-.079	.080
MTLD pre test	.041	-.018	.947	.097	-.067
Extraction Method: Principal Component Analysis.					
Rotation Method: Varimax with Kaiser Normalisation.					
a. Rotation converged in 7 iterations.					

Table 14 shows the analysis yielded five factors, with an eigenvalue greater than 1, explaining a total of 81.02% of the variance for the entire set of variables. Factor 1 was named Accuracy due to the high loadings by the following variables: Errors per 100 words, Grammatical errors per 100 and Lexical errors per 100 words, and Spelling and punctuation errors per 100 words. This first factor explained 30.55% of the variance. Factor 2 was named Complexity 1 due to the high loadings by the following variables: Mean length of t-units and Clauses per t-unit. This second factor explained 16.8% of the variance for the entire set of variables. Factor 3 was named Lexical Diversity due to the high loading of the variables MTL D and Voc-D. This factor explained 13.45% of the variance for the entire set of variables. Factor 4 was named Complexity 2 due to the high loading of Complex nominals per clause. This factor explained 10.287% of the variance for the entire set of variables. Factor 5 was named Fluency due to the high loading of the variable Total number of words in the text.

The factor analysis showed it would be possible to combine variables to arrive at a composite score with five new variables representing accuracy, complexity 1 and 2, lexical diversity and fluency.

3.12.2 Data Transformation and Normalisation

To create the five composite variables in relation to the solution proposed by the principal component analysis, the variables' scores first needed to be changed into Z scores, so they could be combined according to the factors they loaded on. The final solution of the PCA proposed 1 fluency, 1 accuracy, 2 syntactic complexity and 1 lexical diversity composite variable. These five new pre-test measures were then checked for normality using a Kolmogorov-Smirnov and a Shapiro-Wilk test, but some were found to be non-normal (Appendix J). To be able to run parametric statistics they needed to be transformed, thus they

were first changed to t scores and depending on their skewness they were transformed using either SQRT or Log10 transformations in SPSS (MacDonald, 2014). For the CAF measures by group after transformation at the pre-test, 17 out of 20 variables were normally distributed; for the re-test by group, 18 out of 20 variables were normally distributed and for the post-test by group, 17 out of 20 variables were normally distributed.

The following other variables were also transformed in the same way as the CAF measures. The attitudes summary score, the LLAMA and Oxford Quick Placement Test (QPT) by group, had 14 out of 16 variables normally distributed after transformation.

Next, gain scores were calculated from pre to re, re to post and pre to post-test and checked for normality. By group, the transformed data was normal for 17/20 variables for pre to re, re to post 16/20 and for pre to post 18/20.

In total, the proportion of normally distributed variables was deemed large enough to assume normality, and since parametric statistics have much more power than non-parametric statistics, it was decided to use parametric statistics. Others, for instance Norman (2010), state the acceptability of using parametric statistics with non-normal distributions and in addition, the Central Limit Theorem (Polyà, 1920) shows that, for sample sizes greater than 5 or 10 per group, the means are approximately normally distributed regardless of the original distribution.

3.13 Data Analysis

The data was analysed using SPSS version 24 and different statistical tests were used to answer the research questions. First, research questions 1 a,b,c and d were answered in one model using a repeated measures MANCOVA. Pearson correlations were then used to answer research questions 2, 3, 4 and 5.

3.14 Summary

In summary, the focus of this study was to examine the effects of direct, indirect and metalinguistic written corrective feedback on the complexity, accuracy and fluency (CAF) of English as a foreign language students' academic writing. Furthermore, it attempted to discover if the moderating variables of aptitude, attitudes and proficiency have a role to play in the uptake of unfocused written corrective feedback. The research design was a quasi-experimental quantitative model and the total time of the experiment was fifteen weeks. In total, 139 English academic writing 001 students participated voluntarily in the study. The data was collected from four intact groups of 001 academic writing classes and the participants in the study were made up of four intact groups. The four intact groups thus became the direct feedback group, the indirect feedback group, the metalinguistic feedback group and a control group. These three feedback groups were given four rounds of WCF and the control group did not receive any WCF. To begin with, ethical approval was obtained, a pilot study was conducted and issues that appeared in the pilot study were noted and informed changes to the main study's methodology. Inter-rater reliability analysis and preliminary data checks were also carried out. When the data collection was finished, the data was analysed to answer the research questions which are presented in detail in the next chapter.

Chapter 4: Results

4.1 Introduction and Overview

The focus of this study was to examine the effects of direct, indirect and metalinguistic written corrective feedback on the complexity, accuracy and fluency (CAF) of English as a foreign language students' academic writing. Furthermore, it attempted to discover if the moderating variables of aptitude, attitudes and proficiency have a role to play in the uptake of unfocused written corrective feedback. In total, 139 English academic writing 001 students participated voluntarily in the study and the data was collected from four intact groups of 001 academic writing classes. The different groups consisted of three treatment groups: a group receiving direct feedback $n= 40$, a group receiving indirect feedback $n= 36$ and a group receiving metalinguistic feedback $n= 34$. There was also a control group $n= 35$. This chapter first presents the descriptive results for the CAF variables, and then presents the descriptive results for the ID variables. It then analyses the effects of different types of unfocused feedback on CAF in revised and new texts by answering RQ1a, RQ1b, RQ1c, and RQ1d. The chapter then examines the relationships among the CAF variables by answering RQ 2. Finally, the chapter examines the relationships between CAF variables and ID variables (covariates) by answering RQ 3, 4, and 5.

The research questions the study attempts to answer are:

4.1 RQ1a. Does unfocused corrective feedback lead to an increase in the accuracy, complexity and fluency of student writing on revised tasks, compared to no feedback?

4.2 RQ1b. Which of the types of unfocused corrective feedback have a greater influence on the accuracy, complexity and fluency of student writing on revised tasks?:

(a) direct corrective feedback in the form of written corrections of errors on students' compositions;

(b) indirect corrective feedback in the form of error underlining; and

(c) metalinguistic feedback in the form of error codes with metalinguistic information

4.3 RQ1c. Does unfocused corrective feedback lead to an increase in the accuracy, complexity and fluency of student writing on new tasks, compared to no feedback?

4.4 RQ1d. Which of the types of unfocused corrective feedback has a greater influence on the accuracy, complexity and fluency of student writing on new tasks?

4.5 RQ2. Is there a relationship between gains in accuracy, gains in complexity and gains in fluency on revised and new tasks?

4.6 RQ3. Is there a relationship between L2 proficiency and gains in complexity, accuracy and fluency on revised and new tasks?

4.7 RQ4. Is there a relationship between aptitude and gains in complexity, accuracy and fluency on revised and new tasks?

4.8 RQ5. Is there a relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy and fluency on revised and new tasks?

4.2 Descriptive Results for the CAF Variables

This section presents the descriptive statistics of the CAF gain scores by group. The following CAF gain scores are presented for the pre-test to re-test, and pre-test to post-test: fluency gain scores (Figure 4 and 5), accuracy gain scores (Figure 6 and 7), complexity gain scores (Figure 8 and 9), lexical diversity gain scores (Figure 10 and 11), and complex nominals per clause (Figure 12 and 13). Revised texts are represented by gains from the pre-test to the re-test and new texts are represented by gains from the pre-test to the post-test. The descriptives by test (not for gains) of the composite variables, which are Z scores, are also presented at Appendix K for supplementary information. It is important to note that MANCOVAs for re-test to post-test were ran as a supplementary analysis and are presented at Appendix L, although they did not yield any significant or meaningful results.

The box plots for fluency gain scores (Figure 4) suggest that from the pre-test to the re-test, students in the direct and metalinguistic groups wrote less on the text revision, but students in the indirect and control groups had gains in fluency. Figure 5 shows that for the pre-test to the post-test, students in the direct, indirect, metalinguistic and control groups had gains in fluency on new texts. However, when looking at gains on new texts, it can be seen that there were only very small gains. Looking at Figures 4 and 5, it can be seen that there are some outliers and some extreme outliers, and that the range for the pre-test to re-test is noticeably smaller than the range for the pre-test to post-test. Furthermore, when looking at Figure 4, the descriptives show that the students in the direct and indirect group performed much more homogeneously, with most students performing similarly (although there was one outlier each in the direct and indirect group) compared to the metalinguistic and the control group.

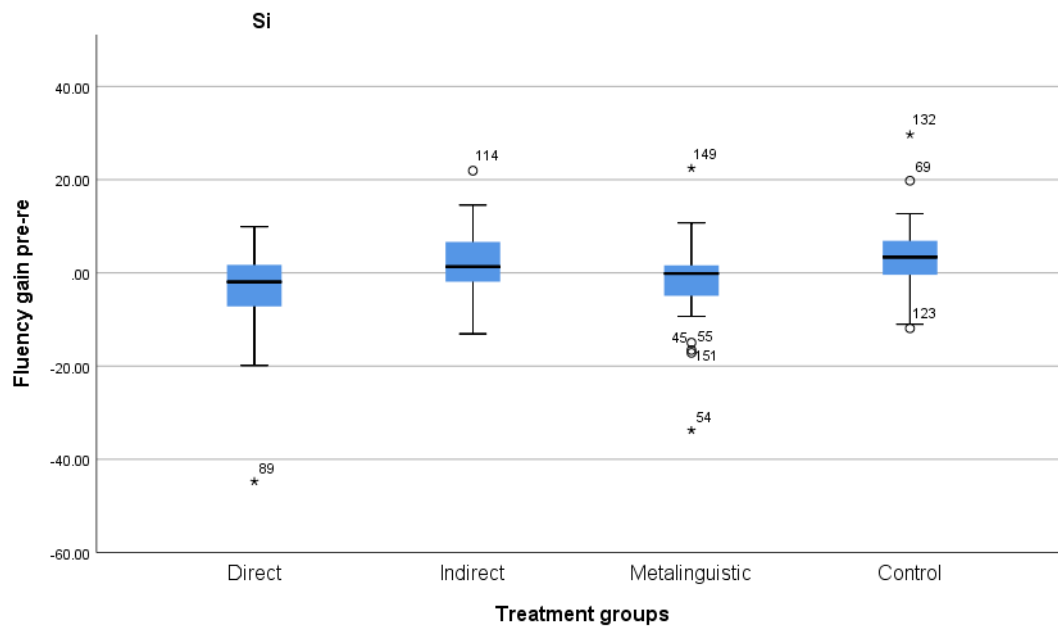


Figure 4. Box Plots of Fluency Gain Scores Pre-test to Re-test (revised task)

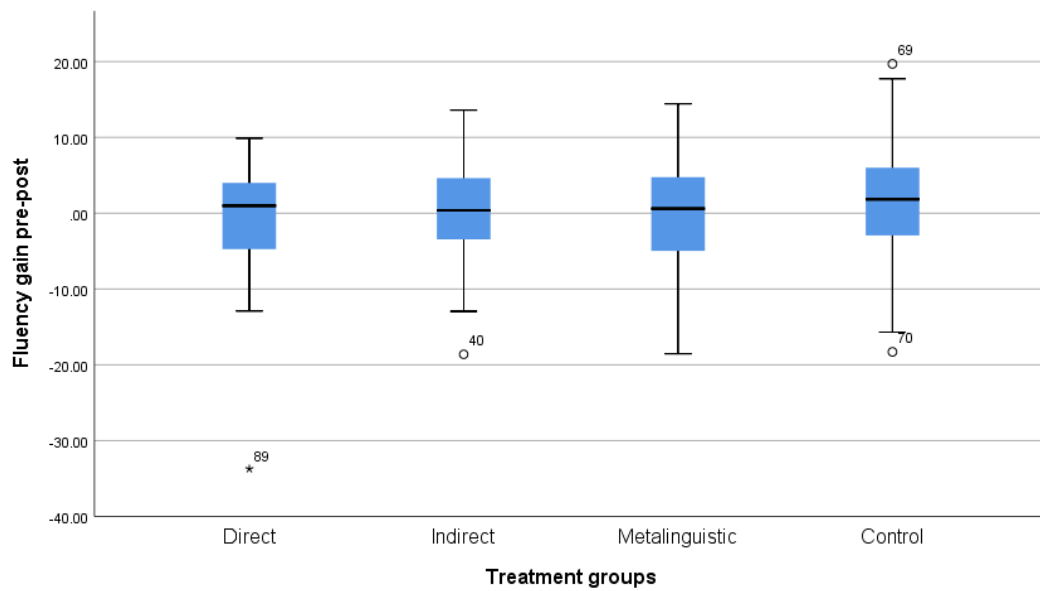


Figure 5. Box Plots of Fluency Gain Scores Pre-test to Post-test (new task)

The box plots for accuracy (Figure 6, Figure 7) suggest that for the text re-write and when writing the new texts, the direct and control groups had positive gains in accuracy and thus made fewer mistakes in their writing, but the indirect and metalinguistic groups had negative gains and thus made more mistakes. Looking at Figures 6 and 7, it can also be seen that there are some outliers, but no extreme outliers and the ranges appear to be very similar in both figures - unlike the ranges for fluency that varied greatly between the pre to re-test and pre-test to post-test.

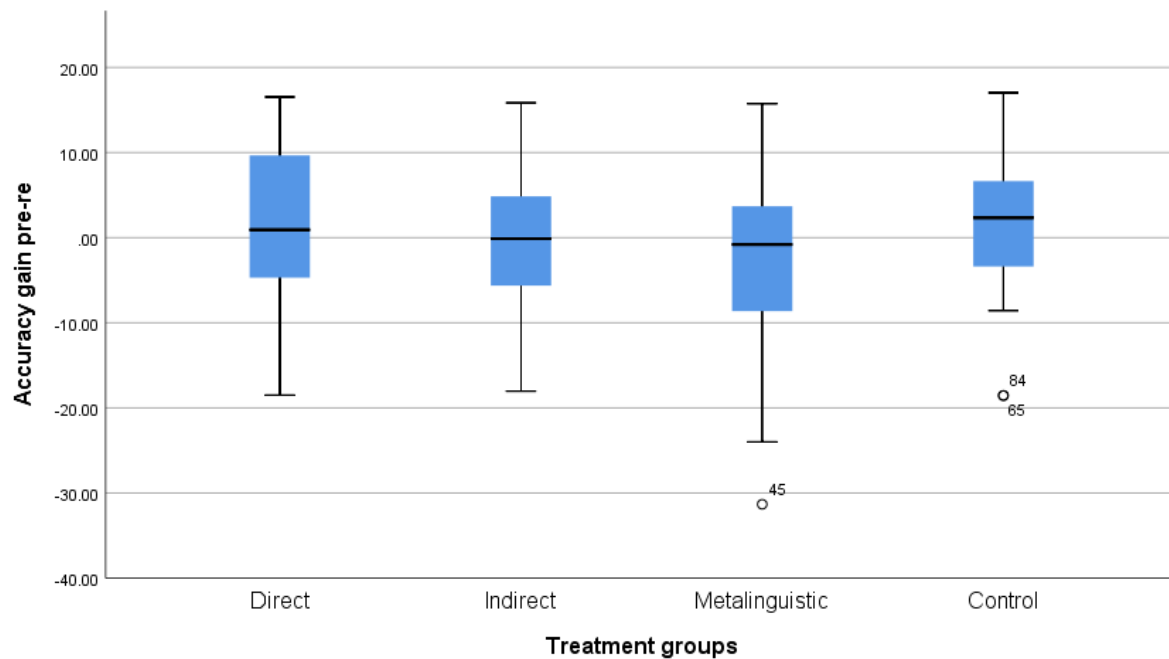


Figure 6. Box Plots of Accuracy Gain Scores for Pre-test to Re-test (revised task)

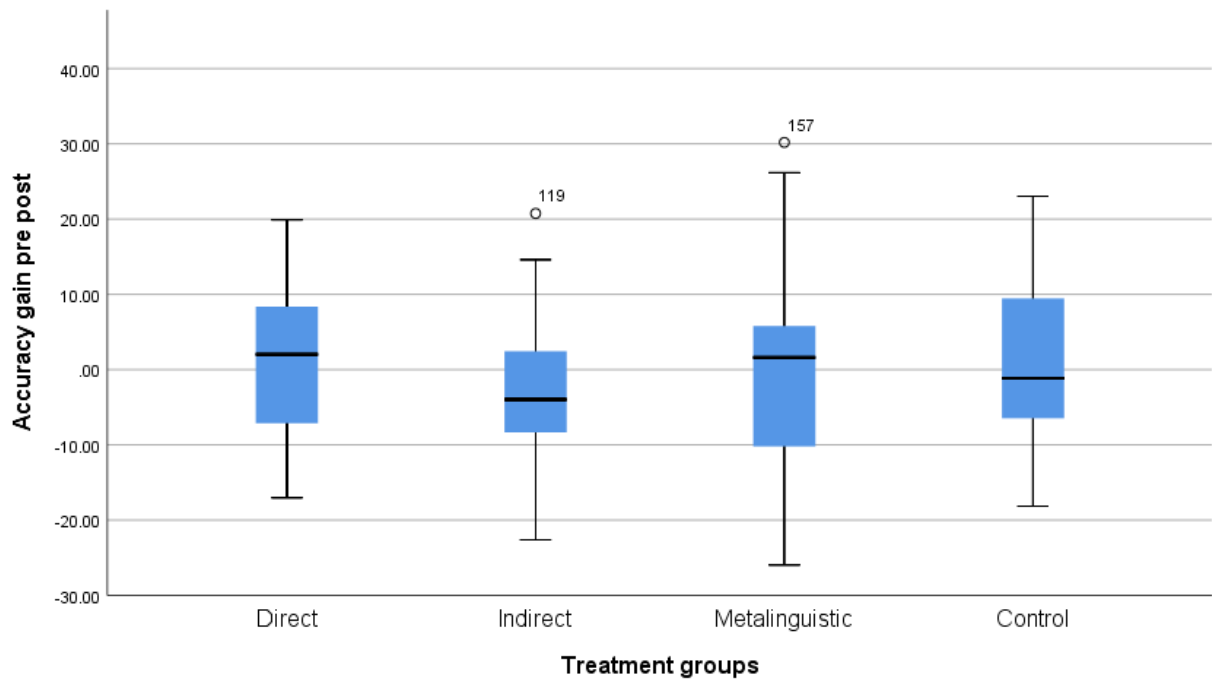


Figure 7. Box Plots of Accuracy Gain Scores for Pre-test to Post-test (new task)

The box plots for complexity gain scores (Figure 8, Figure 9) show that only the direct group had losses in complexity on the text re-write, but the other groups had gains. However, when writing new texts, only the indirect group had losses in complexity, but all other groups had gains in complexity. The figures show that again there are some outliers and some extreme outliers. The pre-test to post-test gains have a larger range than the pre-test to re-test gains. It can also be seen from Figure 8, that the students' performance for pre-test to re-test gains was not homogenous in all groups; the control group being the least homogenous. For the pre-test to the post-test, the lack of homogeneity was apparent for the control group due to the amount of outliers.

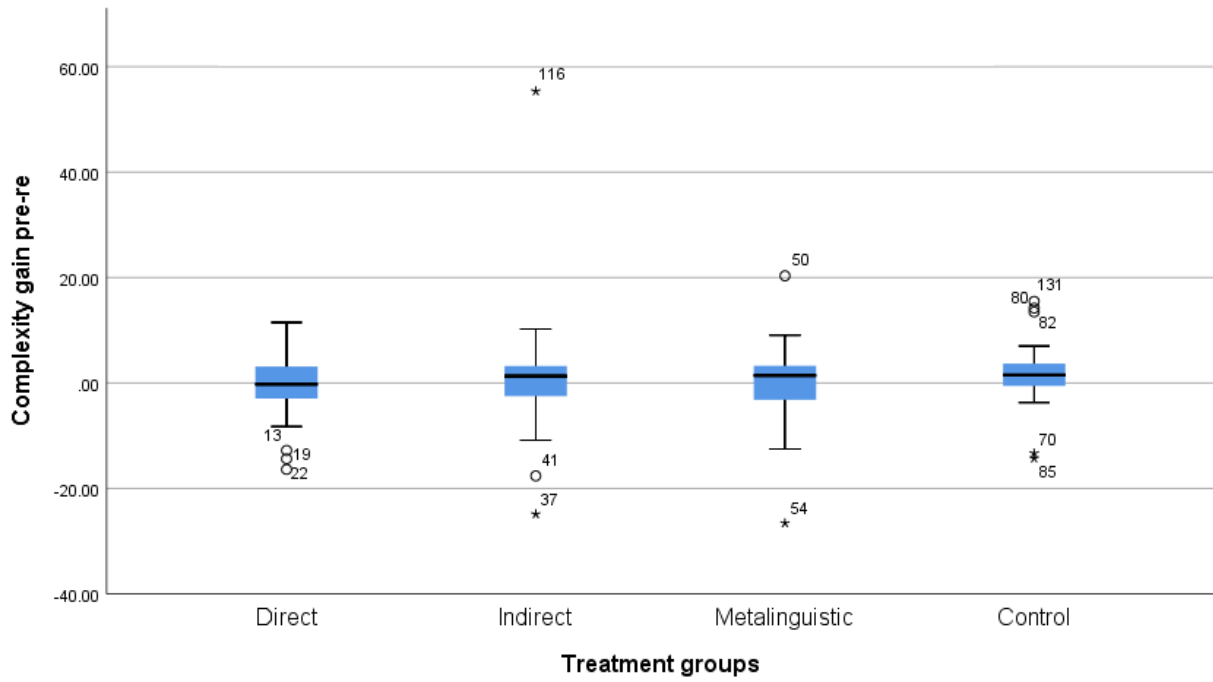


Figure 8. Box Plots for Complexity Gains Pre-test to Re-test (revised task)

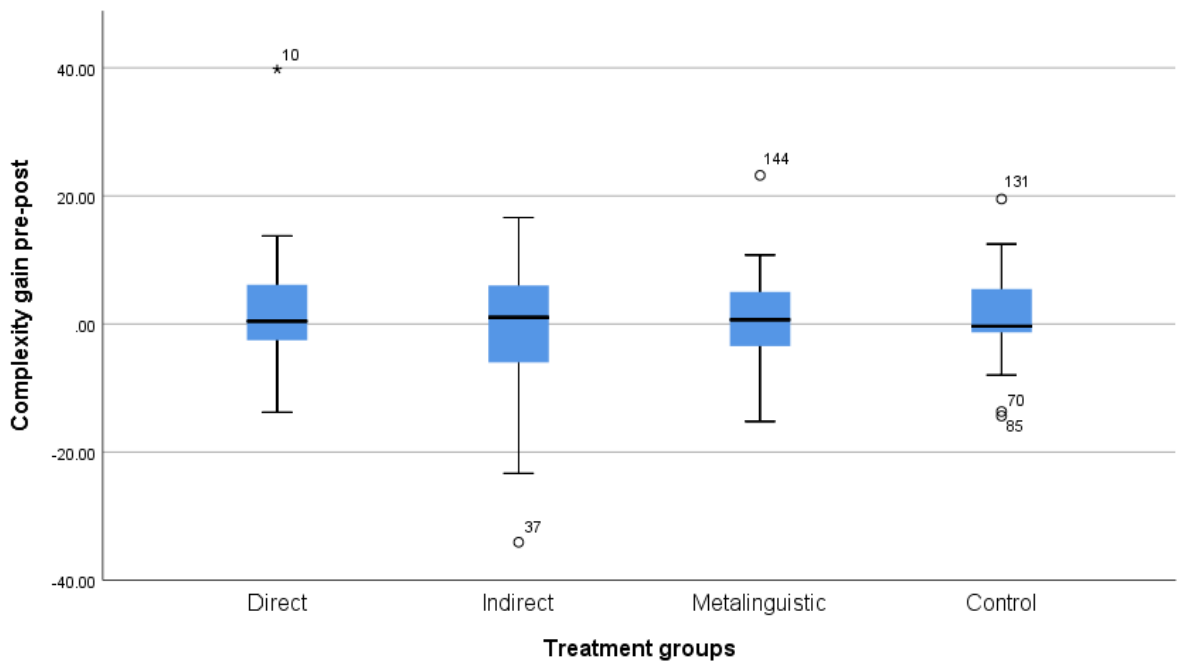


Figure 9. Box Plots for Complexity Gains Pre-test to Post-test (new task)

The descriptive statistics and box plots for gains in complex nominals per clause, is presented separately from the other CAF measures in Figure 10 and Figure 11, as complex nominals per clause did not load onto the composite variable (Complexity 1) with the other complexity measures, and would only load on a variable by themselves. The descriptives show that on the text re-write, the direct and metalinguistic groups had negative gains, but the indirect and control groups had positive gains. On new texts, the direct and control groups wrote less complex nominals per clause while the indirect and metalinguistic groups wrote more. The pre-test to post-test gains shows that there were some outliers; even so, there were no outliers in the pre-test to re-test gains. The pre-test to re-test gains also have a smaller range than the pre-test to post-test gains. Looking at Figure 10, it can be seen that the indirect and the control group displays the least homogeneity due to the amount of outliers.

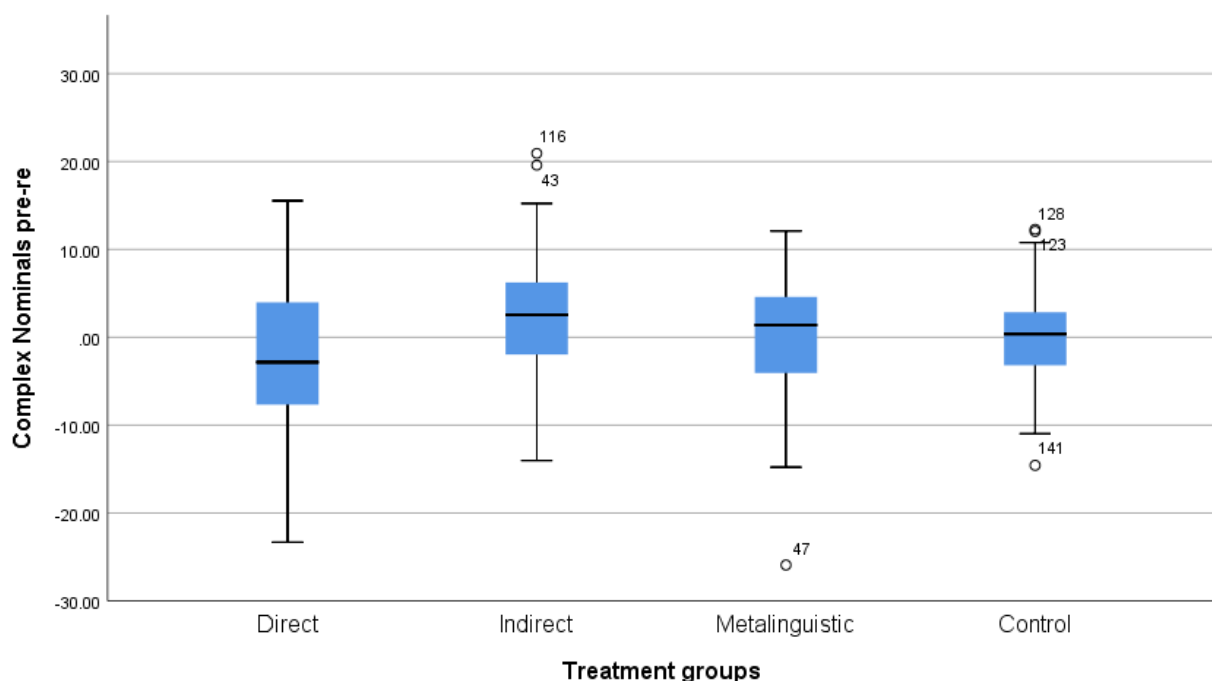


Figure 10. Box Plots for Complex Nominals per Clause Gains Pre-test to Re-test (revised task)

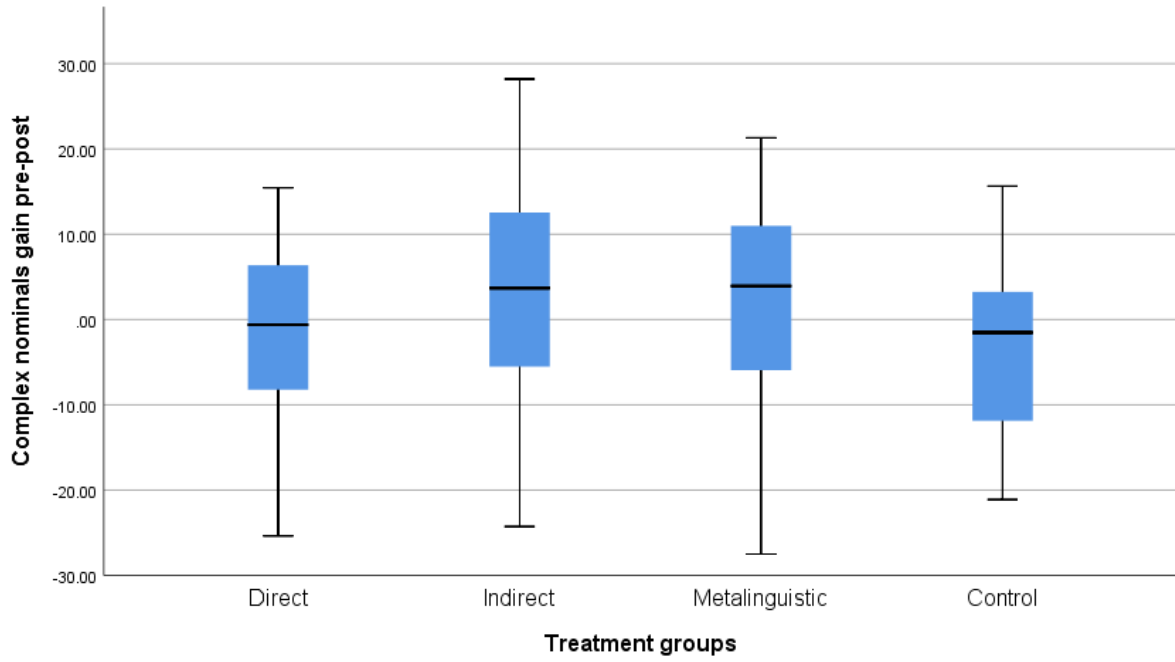


Figure 11. Box Plots for Complex Nominals per Clause Gains Pre-test to Post-test (new task)

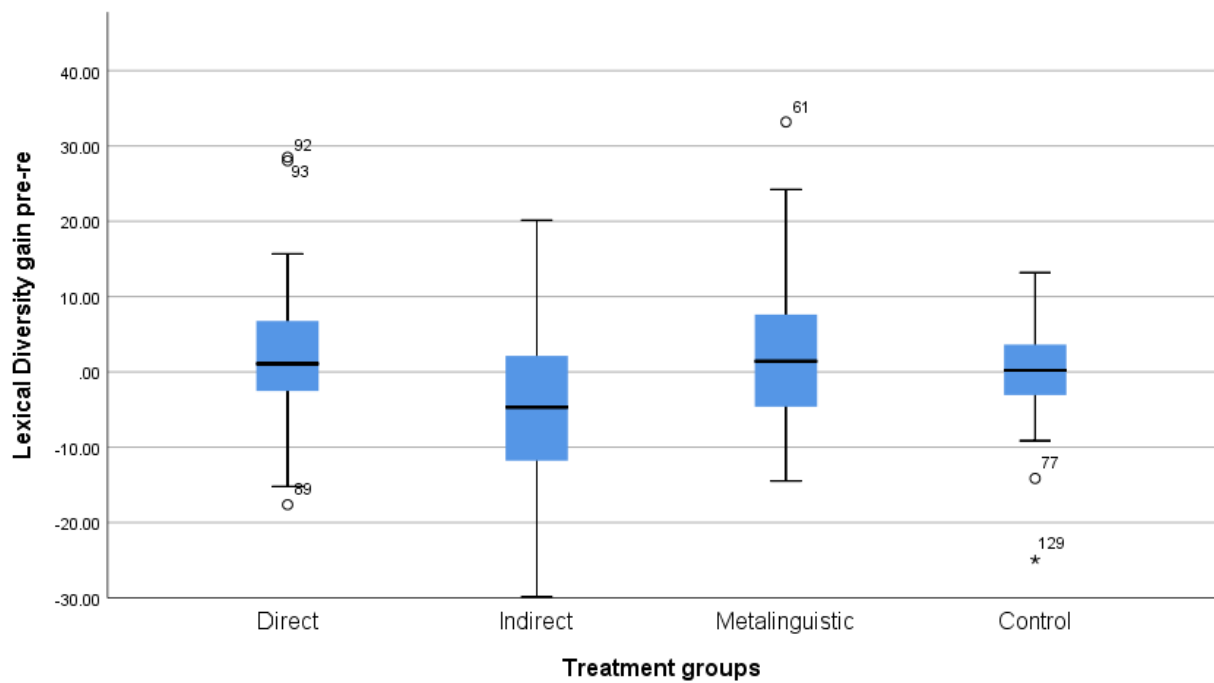


Figure 12. Box Plots for Lexical Diversity Gains Pre-test to Re-test (revised task)

The descriptive statistics for lexical diversity gains for the text re-write (Figure 12) show that the indirect and the control groups produced less lexically diverse writing, whereas the direct and metalinguistic groups wrote more lexically diverse writing. On new texts (Figure 13), the metalinguistic and control groups had losses in lexical diversity whereas the indirect and direct groups had gains. The range appears to be similar for both the pre-test to post-test gains and pre-test to re-test gains, and both have some outliers, although only the pre-test to post-test gains have one extreme outlier. Looking at Figure 13 it can be seen on gains from the pre-test to the post test, that the metalinguistic and the control group are the least homogenous of the four groups due to the large number of outliers.

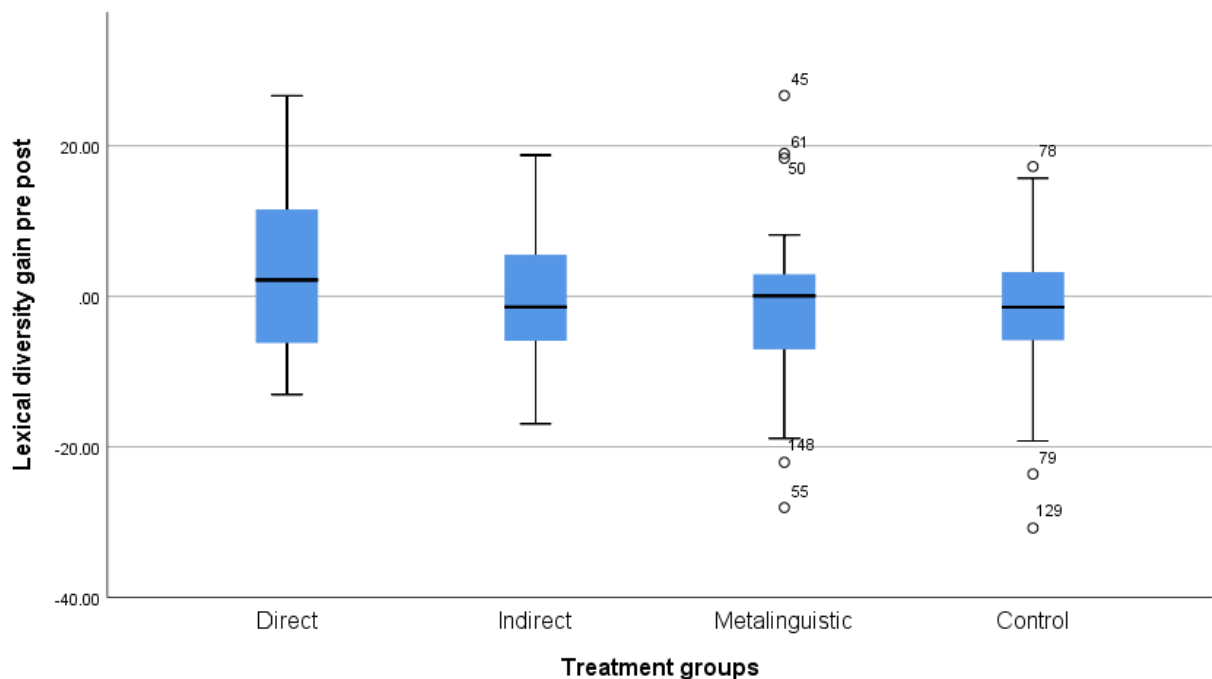


Figure 13. Box Plots for Lexical Diversity Gains Pre-test to Post-test (new task)

From the descriptives, the pattern that appears is that ranges in general for the pre-test to re-test gains are smaller than the ranges for the pre-test to the post-test gains. This suggests that the students benefit from WCF in a way that is more similar on text re-writes, but on new texts, since the gains are less clustered and the range is larger, the benefit of WCF appears to be more varied on new texts. Another general pattern that appears from the descriptives is that the metalinguistic and control groups are in general the least homogenous of the groups due to the number of outliers.

4.3 Descriptive Results for the ID Variables

The descriptive statistics for the ID variables (covariates) for the whole sample are now presented below in Table 16. Table 16 shows the descriptive statistics for the covariates: Proficiency using the Oxford quick placement test, the LLAMA_B and F, and the attitudes summary score. It is important to note that LLAMA_B had a smaller range than LLAMA_F and thus the students performed more homogeneously on LLAMA_B. Table 16 also presents further information of the descriptive statistics.

Table 16. Descriptive Statistics of the Covariates for the Whole Sample

Descriptive Statistics							
	N	Range	Min	Max	Mean	Std. Deviation	Variance
Oxford Quick Placement Test	139	22	36	58	48.81	5.033	25.332
LLAMA F Aptitude test	139	100	0	100	50.50	23.562	555.179
LLAMA B Aptitude test	139	95	0	95	52.63	21.747	472.946
Attitudes	139	2.18	1.06	3.23	1.91	0.39	0.15
Valid N (listwise)	139						

Looking at Table 16, it can be seen that the mean of the Oxford Quick Placement Test is 48.81 and this represents a score of advanced, according to the Oxford Quick Placement Test’s manual (University of Cambridge, 2001). There is a rather small spread for the Oxford quick placement test as the standard deviation was only 5.033 and the scores for the students cluster around the mean. An average score of 50.50 on the LLAMA_F Aptitude test and a 52.63 average score on LLAMA_B represents a “good score” according to the LLAMA manual (Meara, 2005). However, the LLAMA aptitude tests both had rather large standard deviations showing that the scores did not cluster and there was a spread of scores.

The descriptives of the covariates by group are now presented at figures 14, 15, 16 and 17.

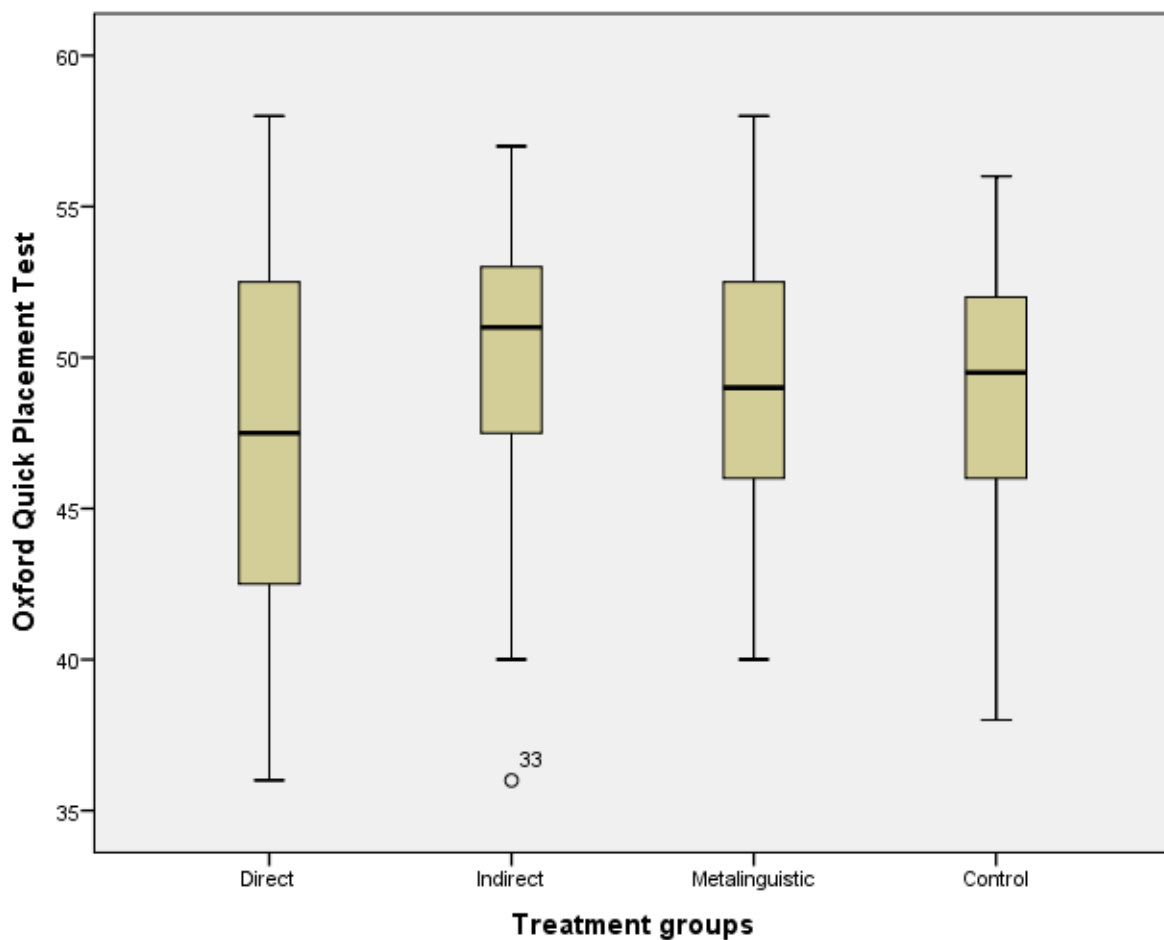


Figure 14 . Box Plots for the Oxford Quick Placement Test by Group

The box plots for the Oxford Quick Placement test show that the students in the direct group had the lowest mean score, while the students in the indirect group had the highest. Only the indirect group had an outlier, but the range for the direct group shows that it was the least homogenous of the groups.

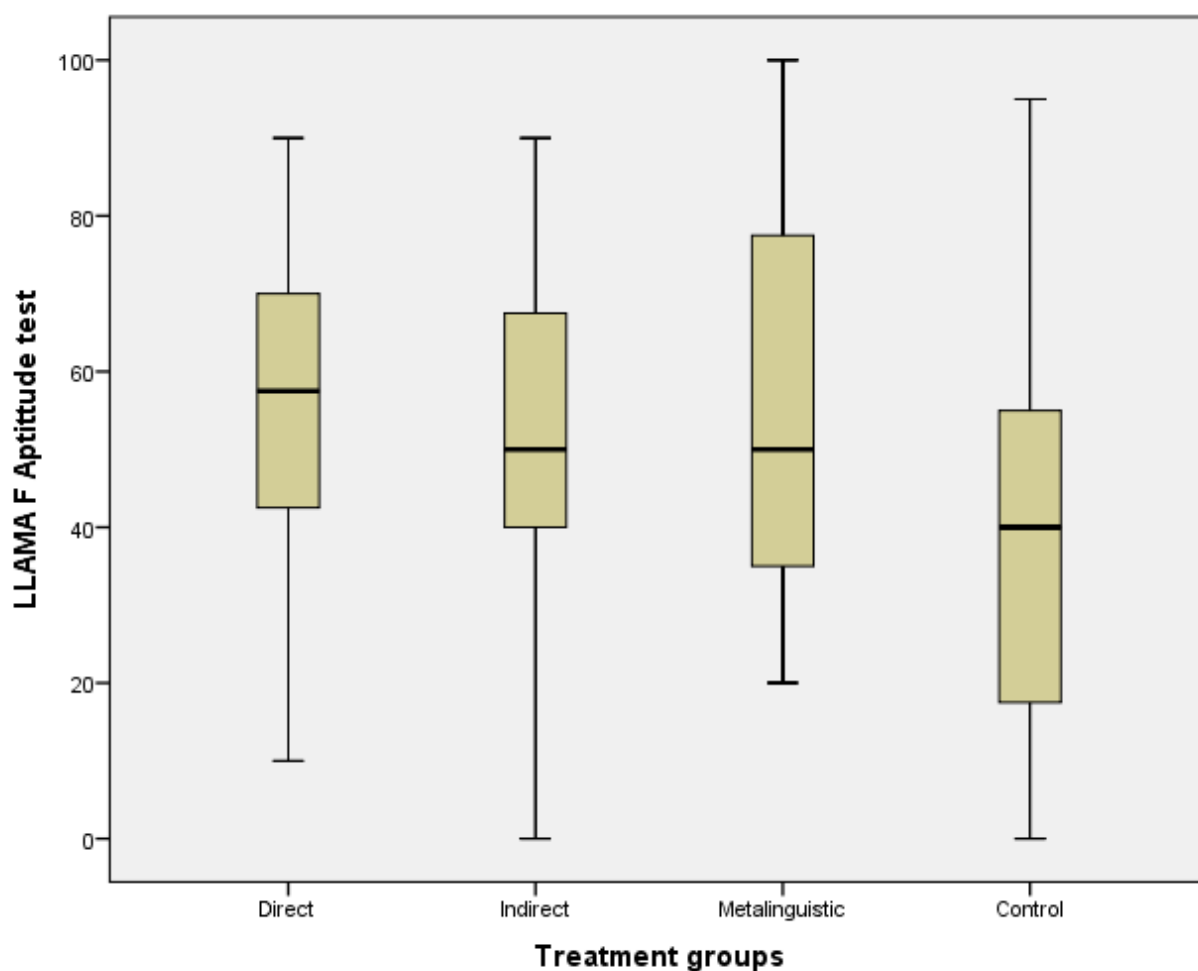


Figure 15. Box Plots for the LLAMA_F Aptitude Test by Group

The box plots for the LLAMA_F aptitude test show the students in the control group had the lowest mean score, while the students in the indirect group had the highest. None of the groups had an outlier, and the range for the direct and indirect groups shows they were more homogenous than the metalinguistic and control groups.

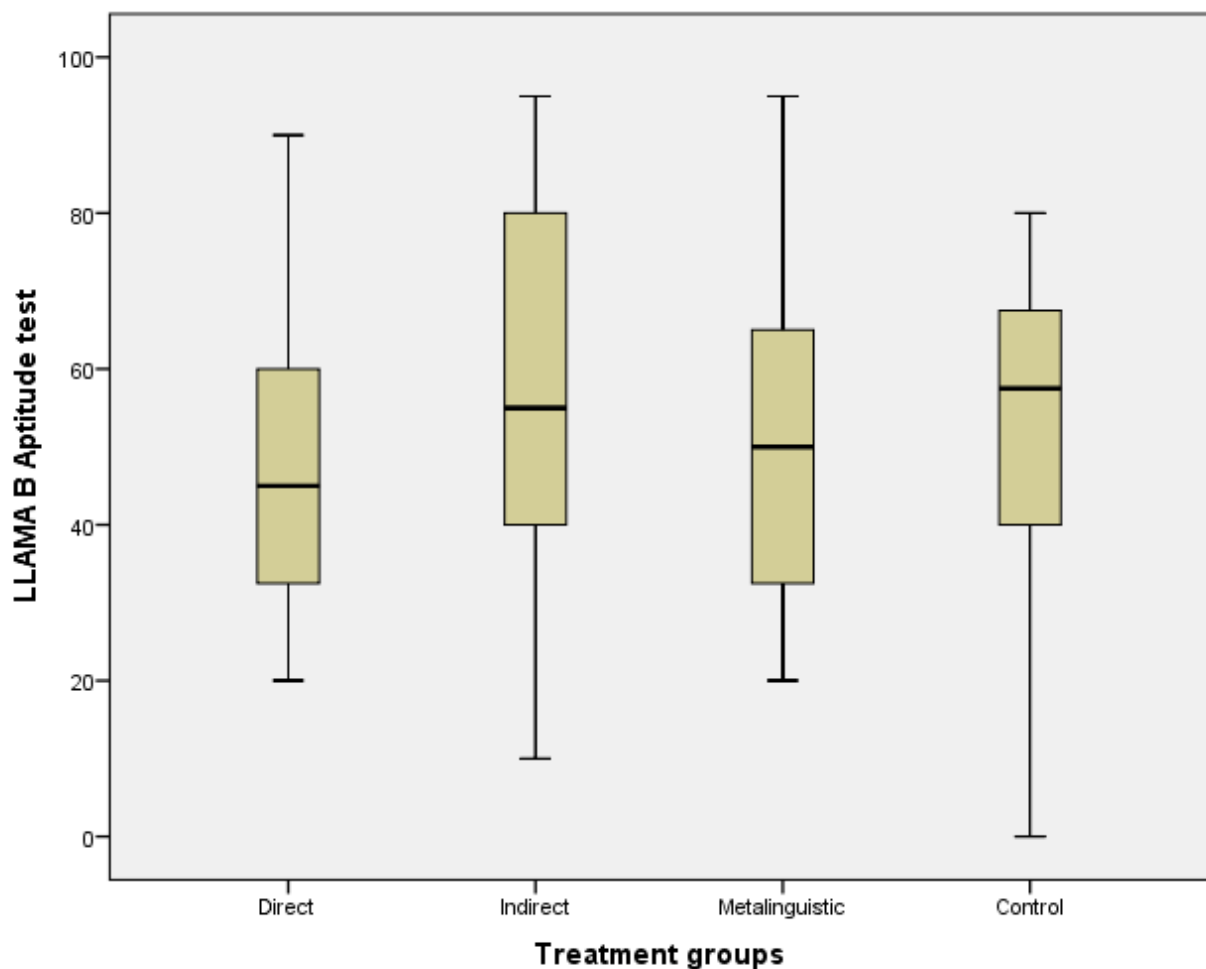


Figure 16. Box Plots for the LLAMA_B Aptitude Test by Group

The box plots for the LLAMA_B aptitude test show the students in the control group had the highest mean score, while the students in the direct group had the lowest. None of the groups had an outlier, and the range for the direct, metalinguistic and control group were very similar; however, the indirect group's range was the largest showing that this group was the least homogenous.

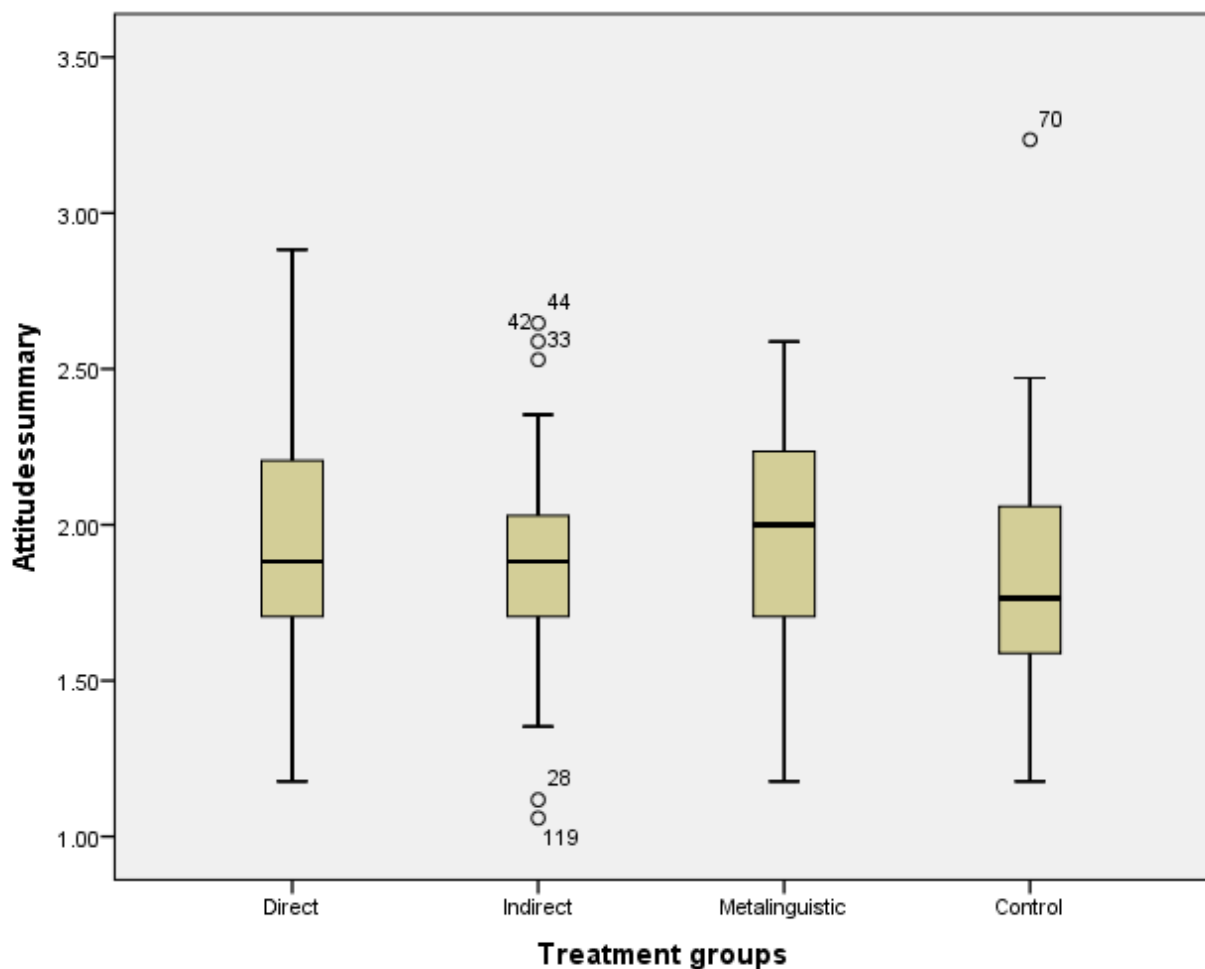


Figure 17. Box Plots for the Attitudes by Group

The box plots for attitudes show that the students in the metalinguistic group had the most positive attitudes towards WCF, while the students in the control group had the least positive attitudes. Although the indirect group did not have the largest range, the descriptives suggest it is the least homogenous group due to the large amount of outliers present, compared to the other groups.

4.4 Tests for any Initial Between-group Differences in Language Proficiency, Aptitude and Attitudes

In order to find out if the students in the four groups began the study with similar writing proficiency, aptitude and attitudes, the Kruskal-Wallis test was used on the pre-test measures to test for any initial between-group differences in language proficiency, aptitude and attitudes. The Kruskal-Wallis test was used due to the non-normal distribution of the data (Appendix I), The results are presented below in Table 17.

	P value
Proficiency (Oxford Quick Placement Test)	.268
LLAMA B	.208
LLAMA F	.102
Attitudes	.355

Table 17. Kruskal-Wallis Test: Language Proficiency, Aptitude and Attitude Measures

The results showed that the four groups were similar, with no significant difference between the groups regarding all measures using an alpha level of 0.05 (Table 17). These results suggest that all treatment groups had a comparable L2 proficiency, aptitude, and attitudes at the beginning of the data collection. Thus, it can be assumed that any differences found later on in the study are not related to initial differences between treatment groups.

4.5 Effects of Different Types of Unfocused Feedback on CAF in Revised and New Texts

To answer research questions 1 a,b,c and d in one model, MANCOVAs were run on pre-test to re-test gain scores (which measured the gains in CAF on revised tasks), and pre-test to post-test gain scores (which measured the gains in CAF on new tasks). The dependent variables were continuous, linear, and normally distributed variables with equal variance–covariance matrices between the groups; thus, they met the MANOVA assumptions (Laerd Statistics, 2019). In addition, the two groups had parallel lines with homoscedasticity, and therefore, had equal slopes and variances (see Table 18 and Table 19 for pre-test to re-test gain scores, and Table 20 and Table 21 pre-test to post-test gain scores) which met the additional assumptions for MANCOVA (Table 22).

Table 18. Box's Test of Equality of Covariance Matrices

Box's Test of Equality of Covariance Matrices^a

Box's M	69.150
F	1.428
df1	45
df2	37628.406
Sig.	.031

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

Table 19. Levene's Test of Equality of Error Variances

Levene's Test of Equality of Error Variances^a				
	F	df1	df2	Sig.
Fluency gains pre re	.413	3	130	.744
Accuracy gains pre re	2.313	3	130	.079
Complexity gains pre re	.657	3	130	.580
Lexical diversity gains pre re	2.058	3	130	.109
Complex nominal gains pre re	1.273	3	130	.287

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

Table 20. Box's Test of Equality of Covariance Matrices

Box's Test of Equality of Covariance Matrices^a	
Box's M	48.865
F	1.009
df1	45
df2	37628.406
Sig.	.455

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

Table 21. Levene's Test of Equality of Error Variances^a

Levene's Test of Equality of Error Variances ^a				
	F	df1	df2	Sig.
Fluency gains pre post	.638	3	130	.592
Accuracy gains pre post	1.698	3	130	.171
Complexity gains pre post	1.806	3	130	.149
Lexical diversity gains pre post	.840	3	130	.474
Complex nominal gains pre post	.507	3	130	.678

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

4.1 RQ1a. Does unfocused corrective feedback lead to an increase in the accuracy, complexity and fluency of student writing on revised tasks, compared to no feedback?

4.2 RQ1b. Which of the types of unfocused corrective feedback has a greater influence on the accuracy, complexity and fluency of student writing on revised tasks?

To answer RQ 1a and RQ1b, a MANCOVA was run on pre-test to re-test gain scores. The MANCOVA for pre-test to re-test gains showed that Box's M test was significant (Table 18). Box's M test is a test that also checks normality and some non-normal distributions in the variables would make Box's M significant. In this case, the data used in the study did have some non-normal distributions. This would have caused Box's M test to be significant, thus proceeding with the MANCOVA with a significant result was deemed acceptable. When

interpreting the results, Pillai's trace was used rather than Wilks' Lambda due to the significance of Box's M test and Pillai's trace being more robust to violations of Box's M test (Olson, 1974). The results of the MANCOVA are presented in Table 22 and 23 (see bolded rows in Table 22 and 23). See Appendix M for the full version of Table 22 and 23.

Table 22. MANCOVA Pre-test to Re-test

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.012	.300 ^b	5.000	122.000	.912	.012
OQPT	Pillai's Trace	.019	.470 ^b	5.000	122.000	.798	.019
LLAMAB	Pillai's Trace	.017	.433 ^b	5.000	122.000	.825	.017
LLAMAF	Pillai's Trace	.040	1.025 ^b	5.000	122.000	.406	.040
Attitudes summary	Pillai's Trace	.019	.477 ^b	5.000	122.000	.793	.019
Group	Pillai's Trace	.310	2.863	15.000	372.000	.000	.103

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Table 23. Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Group	Fluency gains pre re	1237.291	3	412.430	5.442	.001	.115
	Accuracy gains pre re	400.303	3	133.434	1.675	.176	.038
	Complexity gains pre re	114.181	3	38.060	.563	.640	.013
	Lexical diversity gains pre re	1355.065	3	451.688	4.850	.003	.104
	Complex nominal gains pre re	346.957	3	115.652	1.850	.141	.042

The answer to RQ1a is that there were significant effects on fluency and lexical diversity (decreases), but overall there were no significant increases in CAF on revised tasks compared to the group receiving no feedback.

The results show that there was a statistically significant difference between the feedback groups after controlling for LLAMA B, F, attitudes and the QPT, $F(15, 372.000) = 2.863, p < .005$, Pillai's Trace = .310, partial $\eta^2 = .103$ (Table 22: see bolded row in the table). The effect size is medium according to Cohen (1988). The MANCOVA shows that there was a significant difference between gains from the pre-test to the re-test of the WCF groups for fluency, $F = 5.442, p < .005$ partial $\eta^2 = .115$ and lexical diversity, $F = 4.850, p < .005$ partial $\eta^2 = .104$ (Table 22: see bolded rows in the table; the first bolded row is fluency, the second bolded row is lexical diversity) both have medium effect sizes according to Cohen (1988). Accuracy, complexity and complex nominals per clause; however, were non-significant.

Follow up pairwise comparisons using a Bonferroni post-hoc test are presented in Table 24 and the estimated marginal means are shown in Figures 19 and 20.

Table 24. Pairwise Comparisons

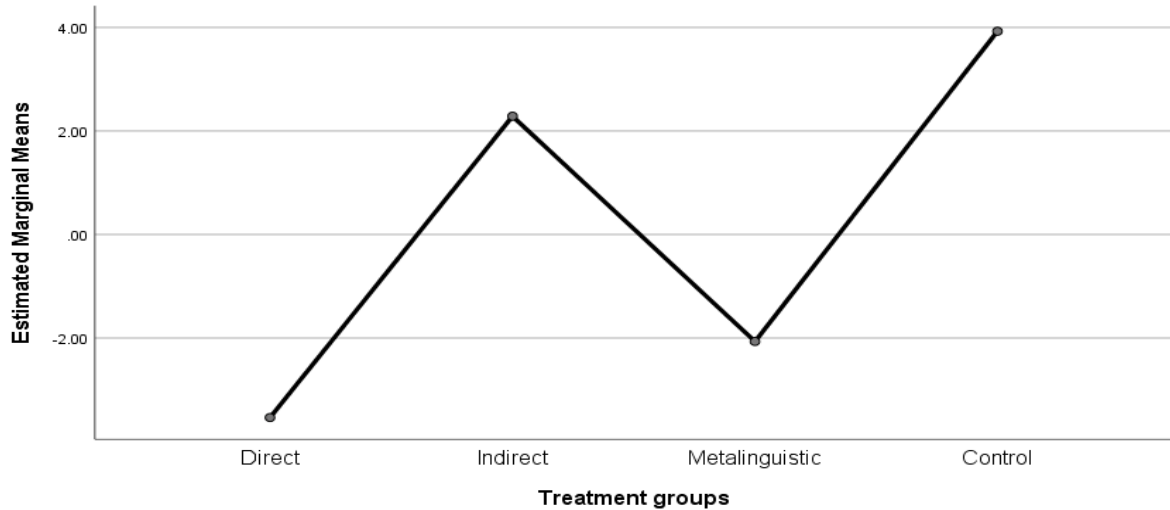
Dependent Variable	(I) Treatment groups	(J) Treatment groups	Mean		Sig. ^b	95% Confidence Interval for Difference ^b	
			Difference (I-J)	Std. Error		Lower Bound	Upper Bound
Fluency gains pre re	Direct	Indirect	-5.821*	2.071	.034	-11.369	-.274
		Metalinguistic	-1.470	2.084	1.000	-7.053	4.112
		Control	-7.463*	2.116	.003	-13.133	-1.794
	Indirect	Direct	5.821*	2.071	.034	.274	11.369
		Metalinguistic	4.351	2.140	.264	-1.381	10.083
		Control	-1.642	2.123	1.000	-7.328	4.044
	Metalinguistic	Direct	1.470	2.084	1.000	-4.112	7.053
		Indirect	-4.351	2.140	.264	-10.083	1.381
		Control	-5.993*	2.215	.046	-11.927	-.059
	Control	Direct	7.463*	2.116	.003	1.794	13.133
		Indirect	1.642	2.123	1.000	-4.044	7.328
		Metalinguistic	5.993*	2.215	.046	.059	11.927
Lexical diversity gains pre re	Direct	Indirect	<u>7.602*</u>	<u>2.304</u>	<u>.007</u>	<u>1.430</u>	<u>13.774</u>
		Metalinguistic	-.554	2.318	1.000	-6.765	5.656
		Control	3.348	2.355	.945	-2.960	9.655
	Indirect	Direct	<u>-7.602*</u>	<u>2.304</u>	<u>.007</u>	<u>-13.774</u>	<u>-1.430</u>
		Metalinguistic	<u>-8.156*</u>	<u>2.381</u>	<u>.005</u>	<u>-14.534</u>	<u>-1.779</u>

Dependent Variable	(I) Treatment groups	(J) Treatment groups	95% Confidence Interval for				
			Mean		Sig. ^b	Difference ^b	
			Difference (I-J)	Std. Error		Lower Bound	Upper Bound
		Control	-4.255	2.362	.443	-10.581	2.072
	Metalinguistic	Direct	.554	2.318	1.000	-5.656	6.765
		Indirect	<u>8.156*</u>	<u>2.381</u>	<u>.005</u>	<u>1.779</u>	<u>14.534</u>
		Control	3.902	2.464	.695	-2.700	10.503
	Control	Direct	-3.348	2.355	.945	-9.655	2.960
		Indirect	4.255	2.362	.443	-2.072	10.581
		Metalinguistic	-3.902	2.464	.695	-10.503	2.700

Based on estimated marginal means

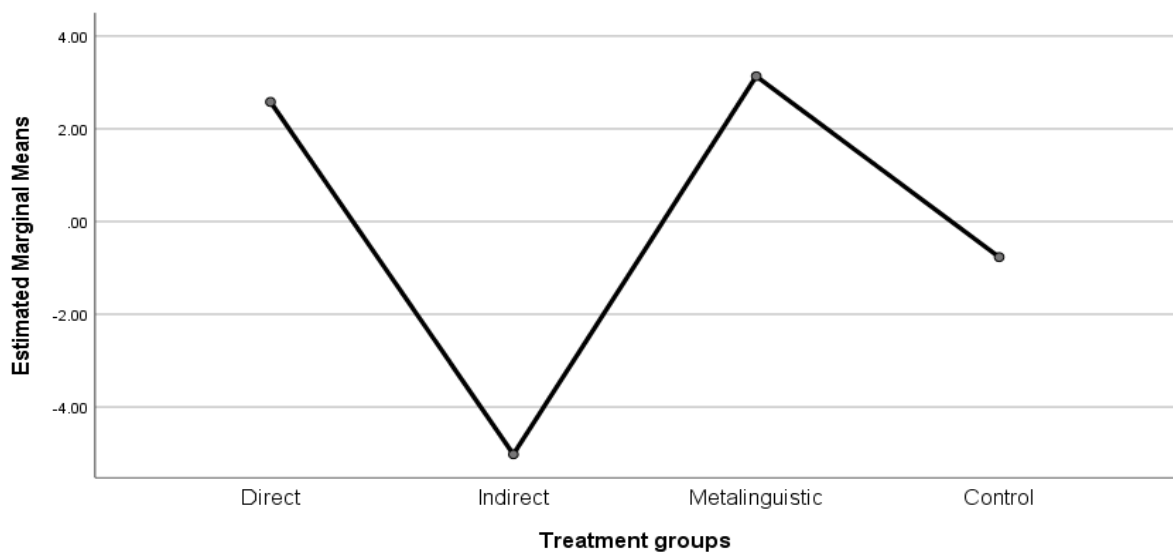
*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.



Covariates appearing in the model are evaluated at the following values: Oxford Quick Placement Test = 48.80, LLAMA F Aptitude test = 50.69, LLAMA B Aptitude test = 52.64

Figure 108. Estimated Marginal Means of Fluency Gains Pre-test to Re-test



Covariates appearing in the model are evaluated at the following values: Oxford Quick Placement Test = 48.80, LLAMA F Aptitude test = 50.69, LLAMA B Aptitude test = 52.64

Figure 19. Estimated Marginal Means of Lexical Diversity Gains Pre-test to Re-test

Follow-up pairwise comparisons using a Bonferroni post-hoc test (Table 24) revealed that for fluency, there was a significant difference between the indirect and direct feedback groups, $p = .034$ (see Table 24 first and third bolded rows). Students in the indirect group ($m = 2.24$) wrote more than the direct group ($m = -3.46$), showing that direct feedback had an effect of

reducing the amount students wrote (it is important to note as mentioned above that students in the indirect feedback group also wrote less than the control group). There was also a significant difference between the direct ($m = -3.46$) and the control ($m = 3.89$) groups, $p = .003$ (see second and fifth bolded rows in Table 24) where the control group wrote more than the direct group, showing that again direct corrective feedback lowered the amount students wrote. Furthermore, there was a significant difference between the metalinguistic feedback group ($m = -2.07$) and the control group, $p = .046$ ($m = 3.89$) (see fourth and sixth bolded row in Table 24) where the metalinguistic group wrote less than the control group, showing that metalinguistic corrective feedback also lowered the amount students wrote (Figure 18) (it is important to note as also mentioned in the previous paragraph above, that students in the indirect feedback group also wrote less than the control group). A summary is presented in Table 25.

Table 25. Summary of Significant Differences 1

CAF Measure	Summary
Fluency	Indirect > Direct
	Control > Direct
	Control > Metalinguistic

For lexical diversity, the MANCOVA revealed there was a significant difference between the direct ($m = 2.31$) and the indirect ($m = -4.95$) feedback group, $p = .007$ (see Table 24: the first and second underlined rows) where the direct group had significantly higher lexical diversity than the indirect feedback group. Furthermore, there was a significant difference between the

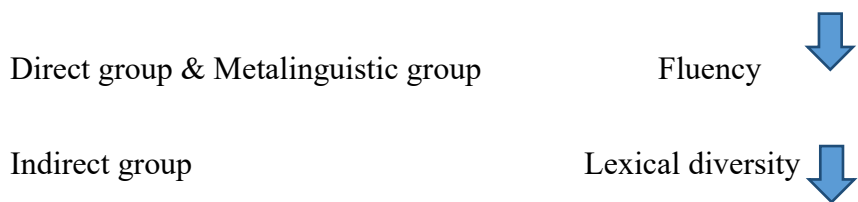
indirect (m= -4.95) and the metalinguistic (m= 2.82) feedback group, p=.005 (see Table 24: the third and fourth underlined rows) where the metalinguistic group wrote more lexically diverse writing than the indirect group (Figure 19). This shows that indirect feedback lowered students' lexical diversity in both cases. A summary is presented at Table 26.

Table 26. Summary of Significant Differences 2

CAF Measure	Summary
Lexical diversity	Direct > Indirect
	Metalinguistic > Indirect

In summary, to answer RQ1b: fluency had losses rather than gains for the direct group and metalinguistic group compared to the control group. Thus, direct and metalinguistic feedback have the effect of reducing the amount students write on revised tasks. Regarding lexical diversity, there was also a significant difference between the indirect, direct and metalinguistic group showing that indirect feedback lowered the students lexical diversity.

For revised tasks:



4.3 RQ1c. Does unfocused corrective feedback lead to an increase in the accuracy, complexity and fluency of student writing on new tasks, compared to no feedback?

4.4 RQ1d. Which of the types of unfocused corrective feedback has a greater influence on the accuracy, complexity, and fluency of student writing on new tasks?

To examine the increase in CAF measures on new tasks to answer RQ1c and RQ1d, a MANCOVA was run on pre-test to post-test gains and is presented at Table 27. The full version of Table 27 can be seen at Appendix N.

Table 67. Multivariate Tests

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Wilks' Lambda	.983	.421 ^b	5.000	122.000	.834	.017
QPT	Wilks' Lambda	.974	.640 ^b	5.000	122.000	.670	.026
LLAMAB	Wilks' Lambda	.979	.532 ^b	5.000	122.000	.752	.021
LLAMAF	Wilks' Lambda	.980	.492 ^b	5.000	122.000	.781	.020
Attitudes summary	Wilks' Lambda	.935	1.698 ^b	5.000	122.000	.140	.065
Group	Wilks' Lambda	.824	1.631	15.000	337.190	.064	.062

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

The MANCOVA shows there was no statistically significant difference between the feedback groups on the combined dependent variables after controlling for LLAMA B, F, attitudes and the QPT (Table 27) although, it is important to note that for group it was approaching significance, $p = .064$. When interpreting the results of the previous MANCOVA to answer RQ 1a and b, Pillai's trace was used, rather than Wilks' Lambda due to the significance of Box's M test, but in this MANCOVA, Wilks' Lambda is used, as Box's M test was not significant. The between-subjects effects can be seen in Table 28 below and the full version of Table 28 can be seen in Appendix N.

Table 28. Between-Subjects Effects

Tests of Between-Subjects Effects						
Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Group	Fluency gain pre post	35.399	3	11.800	.188	.904
	Accuracy gain pre post	387.661	3	129.220	1.121	.343
	Complexity gain pre post	305.659	3	101.886	1.333	.267
	Complex nominal gain pre post	1008.649	3	336.216	2.774	.044
	Lexical diversity gain pre post	664.593	3	221.531	2.133	.099
	Lexical diversity gain pre post	13897.903	134			

The between subject effects (Table 28) showed that there was a significant effect of group for complex nominals per clause, $F = 2.774$, $p < .005$. Therefore, follow up pairwise comparisons using a Bonferroni post-hoc test were conducted to see where the difference was. The pairwise comparisons, however, did not show any significant differences between the groups. In summary, to answer RQ1c; corrective feedback does not lead to an increase in CAF measures compared to no feedback on new tasks.

In summary, to answer RQ1d, on new tasks, none of the feedback types had a greater effect than another or no feedback.

4.6 Relationships among the CAF Variables and Relationships between CAF Variables and ID Variables

To answer the remaining research questions:

- a. RQ2. Is there a relationship between gains in accuracy, gains in complexity and gains in fluency on revised and new tasks?
- b. RQ3. Is there a relationship between L2 proficiency and gains in complexity, accuracy and fluency on revised and new tasks?
- c. RQ4. Is there a relationship between aptitude and gains in complexity, accuracy and fluency on revised and new tasks?
- d. RQ5. Is there a relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy and fluency on revised and new tasks?

Even though some variables had non-normal distribution, in total, the proportion of normally distributed variables was deemed large enough to assume normality and since parametric statistics have much more power than non-parametric statistics, it was decided to use Pearson correlations to answer RQs 2, 3, 4 and 5.

Table 29. Correlations Pre-test to Re-test Gains

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains pre re	Accuracy gains pre re	Complexity gains pre re	Lexical diversity gains pre re
LLAMA F Aptitude test	r	.117							
	Sig. (2- tailed)	.179							
LLAMA B Aptitude test	r	-.052	.189*						
	Sig. (2- tailed)	.551	.028						
Attitudes summary Score	r	-.029	.056	.050					
	Sig. (2- tailed)	.739	.523	.566					
Fluency gains pre re	r	.011	-.063	.060	.012				
	Sig. (2- tailed)	.904	.467	.490	.888				
Accuracy gains pre re	r	-.003	-.142	-.077	-.051	.145			
	Sig. (2- tailed)	.975	.100	.375	.559	.094			
	r	-.103	-.033	.077	.025	.149	.158		

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains pre re	Accuracy gains pre re	Complexity gains pre re	Lexical diversity gains pre re
Complexity gains pre re	Sig. (2- tailed)	.237	.705	.379	.773	.085	.068		
Lexical diversity gains pre re	r	-.014	-.106	-.089	-.069	.048	.052	-.081	
	Sig. (2- tailed)	.869	.221	.309	.425	.582	.554	.351	
Complex nominal gains pre re	r	.068	-.085	.119	.037	-.034	.048	.073	-.059
	Sig. (2- tailed)	.436	.326	.172	.675	.696	.580	.405	.497

Table 30. Correlations Re-test to Post-test Gains

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains re post	Accuracy gains re post	Complex nominal gains re post	Complexity gains re post
LLAMA F Aptitude test	r	.117							
	Sig. (2-tailed)	.179							
LLAMA B Aptitude test	r	-.052	.189*						
	Sig. (2-tailed)	.551	.028						
Attitudes summary score	r	-.029	.056	.050					
	Sig. (2-tailed)	.739	.523	.566					
Fluency gains re post	r	.020	.060	-.007	-.159				
	Sig. (2-tailed)	.817	.491	.938	.066				
Accuracy gains re post	r	.160	.154	-.070	-.061	.005			
	Sig. (2-tailed)	.066	.075	.422	.483	.953			
Complex nominal gains re post	r	.027	.134	-.022	-.055	.075	.187*		
	Sig. (2-tailed)	.758	.123	.802	.524	.392	.030		
Complexity gains re post	r	.056	.126	-.128	-.180*	.212*	.137	.101	
	Sig. (2-tailed)	.518	.146	.141	.038	.014	.114	.244	
	r	.002	.135	.077	-.008	.011	-.015	.039	-.141

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains re post	Accuracy gains re post	Complex nominal gains re post	Complexity gains re post
Lexical diversity gains re post	Sig. (2-tailed)	.980	.120	.374	.930	.904	.863	.652	.105

4.5 RQ2. Is there a relationship between gains in accuracy, gains in complexity and gains in fluency in revised and new tasks?

The correlations show that for the pre-test to re-test data (Table 29), there were no significant correlations. For the re-test to the post-test correlations (Table 30), fluency and complexity showed a weak positive correlation, $r = .212$ $p = .014$. This shows that it is possible for students to improve in both fluency and complexity without a trade-off occurring on new tasks. There was also a weak positive correlation between accuracy and complexity when measured as complex nominals per clause, $r = .187$ $p = .030$ which shows that students can improve in both accuracy (due to the reverse coding of Proportion of error-free t-units to create a composite variable representing accuracy, higher accuracy means better performance) and complexity without trade-offs occurring. For the pre-test to post-test data (Appendix O), there were no significant correlations.

4.6 RQ3. Is there a relationship between L2 proficiency and gains in complexity, accuracy and fluency on revised and new tasks?

To answer this research question, Pearson correlations were run (Tables 29, 30, and Appendix O) and show that none of the variables were significantly correlated.

4.7 RQ4. Is there a relationship between aptitude and gains in complexity, accuracy and fluency on revised and new tasks?

The Pearson correlations (Tables 29, 30, and Appendix O), show that for all the data there were no significant correlations.

4.8 RQ5. Is there a relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy, and fluency on revised and new tasks?

The Pearson correlations (Table 29) show that for the pre-test to re-test data, there were no significant correlations. However, for the re-test to post-data (Table 30), there was a weak negative correlation with complexity, $r = -.180$ $p = .038$. Thus, students who have positive attitudes towards corrective feedback actually produce less complex writing. For the pre-test to post-test data (Appendix O), there was a weak negative correlation with fluency, $r = -.172$ $p = .047$ and, therefore, students who have positive attitudes toward corrective feedback wrote less than those who have negative attitudes towards it.

4.7 Summary of Results

In summary, the study examined the effect of WCF generally on CAF, on text revisions. First, RQ 1a. attempted to discover if unfocused WCF generally led to an increase in the accuracy, complexity, and fluency of student writing on revised tasks, compared to receiving no feedback. The results showed that there were no significant gains in CAF on revised tasks compared to the control group.

Second, RQ 1b. attempted to discover which of the different types of unfocused corrective feedback had a greater influence on the accuracy, complexity, and fluency of student writing on revised tasks. The direct group and the metalinguistic group also had negative gains in fluency rather than positive gains compared to the control group, thus the results showed

that direct and metalinguistic feedback lower fluency, while they increase lexical diversity. Indirect feedback increases fluency, while it lowers lexical diversity.

Thus, it was concluded that direct and metalinguistic feedback have the effect of reducing the amount students write on revised tasks. Regarding lexical diversity, there was also a significant difference between the indirect, direct and metalinguistic group, showing that indirect feedback lowered the students' lexical diversity. In effect, a form of trade-off condition that is dependent on feedback type was found.

Third, RQ 1 c. asked if unfocused corrective feedback leads to an increase in the accuracy, complexity, and fluency of student writing on new tasks, compared to no feedback. RQ1d. asked which of the types of unfocused corrective feedback had a greater influence on the accuracy, complexity, and fluency of student writing on new tasks. The results showed that WCF generally did not lead to an increase in CAF measures compared to no feedback on new tasks and that none of the feedback types had a greater effect than each other, or no feedback.

RQ 2 also aimed to explore the interrelationship of CAF measures and the correlations showed that for the pre-test to re-test data, and pre-test to post-test data, there were no significant correlations. However, for the re-test to the post-test correlations, fluency and complexity showed a weak positive correlation. This demonstrates that it is possible for students to improve in both fluency and complexity without a trade-off occurring. Furthermore, there was also a weak positive correlation between accuracy and complexity, which shows that students can both improve in accuracy and complexity without trade-offs occurring.

RQs 3, 4 and 5 looked at the correlations between IDs and CAF, and showed no significant correlations between L2 proficiency and gains in CAF, or aptitude and gains in CAF on both revised and new tasks. When looking at the relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy and fluency on revised and new tasks, the results showed that for the pre-test to re-test data, there were no significant

correlations - although for the re-test to post-data, correlations were found. First, there was a weak negative correlation with complexity. Thus, students who had positive attitudes towards corrective feedback wrote less complex writing. Second, for the pre-test to post-test data there was a weak negative correlation with fluency. Therefore, students who had positive attitudes toward corrective feedback wrote less on the post-test. Thus, both significant results involving attitudes relate to new texts.

Chapter 5: Discussion

The present study has examined the effects of unfocused direct, indirect, and metalinguistic written corrective feedback (WCF) on the complexity, accuracy, and fluency (CAF) of 139 (L1) Arabic or Urdu (L2) English students' writing. It has also investigated if the moderating variables of aptitude, attitudes and proficiency affect the uptake of feedback. Students in four intact groups were designated as feedback groups, together with a control group. They wrote argument essays and were given four rounds of feedback and feedback support sessions over fourteen weeks; learners in the control group received no feedback or support sessions. The discussion of the results will first focus on the interrelationship of the CAF measures (RQ 2), then the relationships between CAF variables and ID variables (RQ 3,4,5) and finally the effects of WCF on CAF (RQ 1 a,b,c,d).

The study also aimed to explore the interrelationship of CAF measures. The correlations show that for the pre-test to re-test data, and pre-test to post-test data, there were no significant correlations. However, for the re-test to the post-test correlations, fluency and complexity shows a weak positive correlation. This demonstrates that it is possible for students to improve in both fluency and complexity without a trade-off occurring. Furthermore, there is also a weak positive correlation between accuracy and complexity, which shows that students can both improve in accuracy and complexity without trade-offs occurring.

The correlations between IDs and CAF showed no significant correlations between L2 proficiency and gains in CAF, as well as aptitude and gains in CAF on both revised and new tasks. When looking at the relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy and fluency on revised and new tasks, the results showed that for the pre-test to re-test data, there were no significant correlations - although for the re-test to post-data, correlations were found. First, there was a weak negative correlation with

complexity. Thus, students who had positive attitudes towards corrective feedback wrote less complex writing. Second, for the pre-test to post-test data there was a weak negative correlation with fluency, therefore, students who had positive attitudes toward corrective feedback wrote less on the post-test. Thus, both significant results involving attitudes relate to new texts.

The study also examined the effect of WCF generally on CAF on text revisions. First, RQ1a. attempted to discover if unfocused WCF generally led to an increase in the accuracy, complexity and fluency of student writing on revised tasks, compared to no feedback. The results showed that there were no significant gains in CAF on revised tasks compared to the control group.

Second, RQ 1b. attempted to discover which of the different types of unfocused corrective feedback had a greater influence on the accuracy, complexity, and fluency of student writing on revised tasks. The direct group and the metalinguistic group also had negative gains in fluency rather than positive gains compared to the control group. Thus, the results show that direct and metalinguistic feedback lower fluency, while they increase lexical diversity. Indirect feedback increases fluency, while it lowers lexical diversity.

Thus, it was concluded that direct and metalinguistic feedback have the effect of reducing the amount students write on revised tasks. Regarding lexical diversity, there was also a significant difference between the indirect, direct and metalinguistic group, showing that indirect feedback lowered the students' lexical diversity. In effect, a form of trade-off condition that is dependent on feedback type was found.

Third, RQ1c. asked if unfocused corrective feedback leads to an increase in the accuracy, complexity and fluency of student writing on new tasks, compared to no feedback. RQ1d. asked which of the types of unfocused corrective feedback had a greater influence on the accuracy, complexity, and fluency of student writing on new tasks. The results showed that

WCF generally did not lead to an increase in CAF measures compared to no feedback on new tasks and that none of the feedback types had a greater effect than each other, or no feedback.

5.1 Connected Growers: The Interaction of Gains in CAF

The relationship between CAF is very complex and in this study, when looking at gains in accuracy, gains in complexity and gains in fluency on revised and new tasks, the results showed that for the pre-test to re-test data there were no significant correlations and thus there was no trade-off or connected growth. For the re-test to the post-test correlations, fluency and complexity showed a weak positive correlation. This shows that it is possible for students to improve in both fluency and complexity without a trade-off occurring, and display connective growth. There was also a weak positive correlation between accuracy and complexity when measured as complex nominals per clause, which shows connected growers, and thus students can both improve in accuracy and complexity without trade-offs occurring.

The present study displayed connected growers rather than the trade-off effects mostly seen in oral studies (Bygate, 2001; Yuan & Ellis, 2003). In writing there is increased planning time and monitoring (Ellis & Yuan, 2005; Michel, 2017) and thus as this study dealt with written tasks, it could explain why there were no trade-off effects. Studies with similar results displaying connected growers include that of van Beuningen et al. (2008, 2012). The authors noted that negative trade-offs in CAF measures were more of an issue with oral feedback than in the offline handling of WCF, thus positive correlations of CAF measures, such as the correlation between complexity and fluency as found in this study are possible. Complexity and fluency could increase together, due to the fact that students were improving their academic writing by writing essays and being taught how to improve complexity as part of the taught part of the academic writing course.

In the present study, a positive correlation with accuracy and complexity was found and this can be linked with Robinson's Cognition Hypothesis (2011) as he points out that tasks will either promote fluency, or complexity and accuracy, which is the same as Skehan's (2009) primary trade-off. According to Robinson (2011), simple tasks are expected to promote fluency, but not complexity or accuracy, and complex tasks promote accuracy and complexity, but not fluency. This study does not actually test Robinson's Cognition Hypothesis directly because it does not include manipulation of task complexity; however, the results, to some extent, support Robinson's (2011) Cognition Hypothesis, in that accuracy and complexity both increased. However, Robinson's Cognition Hypothesis also expects a trade-off between fluency and complexity, which was not found here.

Connected growers were also found in other similar studies. In the study of Van Beuningen et al. (2012) on new texts, students who were given WCF had better accuracy than those of learners who were given extra opportunity to practice their writing skills but were not given WCF, and the WCF did not lead them to produce lexically or structurally less complex writing.

From a dynamic systems theory (DST) approach, the positive correlations in this study can be explained by the fact that there can be supportive growth (Larsen-Freeman, 2009), and all variables in the system are interrelated, so all changes will affect all the other parts of the system, although not always positively. Linguistic subsystems and dimensions of language proficiency interact in ways that are supportive, competitive and conditional, and the development in one of these subsystems may be dependent on the development in another (van Geert & Steenbeek, 2005a). Robinson and Mervis (1998) note that it is important to understand that there are not only static relations between variables, but also that relations change throughout the course of development and given the dynamic nature of the system and interconnected components, the outcome of their interactions can be non-linear too. De Bot

(2008), and Spoelman and Verspoor (2010), argue that cognitive resources are limited, but connected and can also be compensatory and they do not always result in trade-off effects and thus, the results of this study could be explained by that fact. Therefore, complexity and fluency as connected growers are theoretically and practically possible, because they require fewer attentional resources than unconnected subsystems (Spoelman & Verspoor, 2010).

In conclusion, in writing tasks, connected growers are not unusual and the results of this study fall into line with previous studies demonstrating connected growers. Furthermore, the positive correlations in this study can be explained by the fact that there can be supportive growth according to DST and the correlation between complexity and accuracy can also be supported by Robinson's Cognition Hypothesis (2011).

5.2 The Interaction of IDs and CAF

IDs are an under-researched area in WCF research, and this study has explored if there is a relationship between the IDs of L2 proficiency, aptitude and attitudes and gains in CAF.

5.2.1 L2 Proficiency and Gains in CAF

When examining the relationship between proficiency and gains in CAF, the present study has looked at the correlations between L2 proficiency and gains in complexity, accuracy, and fluency on revised and new tasks. The results show that there are no significant correlations.

This result demonstrates that L2 proficiency was not related to the uptake of WCF in this study, and this result is complementary to the MANCOVA that showed L2 proficiency did not have effects. However, before answering the research questions, correlations were ran between L2 proficiency and the pre-test CAF measures to see if they were valid (as reported in the methodology; it is important to note that in the methodology, unlike the results chapter, the

CAF measures were not gains in CAF, but pre-test CAF measures). The correlation for pre-test students' proficiency measured using the Oxford Quick Placement test, revealed significant negative correlations with all the accuracy measures. This means the higher the proficiency of the student, the fewer mistakes they made when writing essays and this result is as expected. The fluency measure; however, did not have significant correlations with proficiency and none of the complexity measures had significant correlations with proficiency. As far as the researcher is aware, there are no similar studies that look at correlations between CAF and L2 proficiency, apart from one study by Van Beuningen et al. (2012) that looks at proficiency measured using a receptive vocabulary test. Van Beuningen et al. wanted to establish whether indirect WCF could be more beneficial for learners with higher levels of metalinguistic awareness as they may be more able to use indirect corrections, and find out if the proficiency level (which they called the educational level) would mediate the efficacy of indirect WCF. They assumed higher-level pupils might be better equipped with metalinguistic knowledge than the lower level ones. Learners with lower levels of metalinguistic competence might have less ability to correct their own errors based on WCF (Ferris, 2004; Hyland & Hyland, 2006; Sheen, 2007). Their findings, however, showed that there was no significant correlation between the effectiveness of the WCF treatments and learners' educational level. Van Beuningen et al. (2012) point out that the probable reason that they did not discover a significant interaction between WCF efficacy and pupils' educational level is that the variance between the levels included in their study was not large enough. In the present study, it could also be argued that the students' proficiency levels were high, and the difference between the levels of participants was again not large enough for a correlation to be found as it is not possible to find a correlation if there are very clustered scores. Even the students classed as lower proficiency were still proficient enough in English that the WCF given would not be beyond their level of competence to be able to correct their errors.

In conclusion, although in this study there were no significant correlations between L2 proficiency and gains in CAF, it could be possible that the participants' proficiency levels were high enough in general, relative to the difficulty of the WCF, to not have an effect. The second reason could be purely methodological; due to the fact that there was not enough variance between the students in terms of proficiency, a correlation was not possible to be found as the scores were highly clustered.

5.2.2 Aptitude and Gains in CAF

Regarding the relationship between aptitude and gains in complexity, accuracy and fluency on revised and new tasks, the results showed that there were no significant correlations.

Few studies exist that investigate how the different components of language aptitude relate to L2 production on written tasks. Therefore, a study with similar results and using a similar sample of participants to this one, as far as the author knows, is not in existence. Studies with different results include Kormos and Trebits (2012) who looked at 44 upper-intermediate learners of English in a Hungarian secondary school. They measured aptitude as deductive ability and grammatical sensitivity, measured using the Hungarian version of the MLAT (HUNLAT) and found that it positively related to gains in accuracy and complexity performance. Another study by Benson and DeKeyser (2018) looked at essays by 151 learners of English as a second language to investigate the effect of WCF. The differences between their study and the present one, is that Benson and De Keyser's (2018) study used focused WCF, and compared the effects of direct and metalinguistic WCF on the simple past and present perfect tenses. However, similar to this study, the context was an academic English class at a university and the participants also made use of the LLAMA_F aptitude test to measure aptitude. They found that learners with greater language aptitude derived more benefit from

direct WCF, but learners with lower language aptitude had more benefit from metalinguistic feedback. In Benson and De Keyser's (2018) study, relationships between WCF and L2 aptitude were found, but there were no relationships found between them in this study. This is most likely due to the difference between the type of WCF used. Benson and De Keyser's (2018) study used focused feedback on only a select part of grammar, while this study has utilised unfocused feedback. In the present study, the lack of gains for the feedback groups could be due to the same reason as the lack of correlation for proficiency. It could be that the WCF was not demanding enough for the students, so cognitive resources were not taxed and the level of aptitude a student had made no difference. This argument should be supported by a correlation between L2 proficiency and aptitude, but there were no significant correlations between the two. Therefore, it is likely that the reason for this could be due to the clustering of L2 proficiency scores preventing any kind of correlations to emerge.

In conclusion, there were no significant correlations regarding the relationship between aptitude and gains in CAF on revised and new tasks, and furthermore, there are very few related studies for comparability.

5.2.3 Attitudes and Gains in CAF

Regarding the relationship between students' attitudes toward corrective feedback and gains in complexity, accuracy and fluency on revised and new tasks, the results showed that for the pre-test to re-test data, there were no significant correlations, but for the re-test to post-data, there was a weak negative correlation with complexity. Thus, students who had positive attitudes towards corrective feedback actually wrote less complex writing. For the pre-test to post-test data there was a weak negative correlation with fluency; therefore, students who had positive attitudes toward corrective feedback produced fewer words.

In general, students who have positive attitudes toward WCF may value it more, and thus may pay more attention to it. Thus, the results of this study could be related to the theory of Krashen's (1982) Monitor hypothesis. The Monitor hypothesis (Skehan, 1998) states that corrective feedback could reduce fluency if it caused the learners to monitor their production more carefully, and the students with positive attitudes toward the feedback may monitor it more. Polio (2012) also argues that WCF develops learners' declarative knowledge and helps learners to monitor the wrong information to make sure that errors would not get into procedural knowledge and become automatic, thus those with positive attitudes would be more likely to monitor their production. Truscott (2007) claims that corrective feedback makes learners aware of the errors they committed, and it may be that this awareness creates a motivation for students to avoid the corrected constructions. Positive attitudes may suggest particularly careful attention to WCF, and the unfocused nature of the feedback may then lead to excessive caution on the part of the students as there are so many issues to consider, making them write less and choose simpler language. This pattern was found on pre-post and re-post; therefore on new tasks only, which are more challenging and thus it affects the more challenging situation of new tasks, but not rewriting.

Studies with different results are scarce, especially regarding WCF. As far as the author is aware, there are no studies that explore how attitudes towards WCF mediate the uptake of WCF with regards to CAF. The closest study to this one is an oral study by Havranek and Cesnik (2001), which looked at English as foreign language students, whose L1 was German. They compared the effects of recasts, repetition and recasts, and elicitation, by measuring the effect of error correction on performance in a subsequent test. The results showed that WCF benefitted learners who had high language proficiency and positive attitudes toward error correction and thus, found a positive correlation between accuracy and attitudes toward feedback. Due to Havranek and Cesnik's (2001) study being on oral feedback – and this study

focusing on WCF, the difference between the two results of the two studies, is in the differing kinds and methods of corrective feedback given.

In conclusion, the present study found a weak negative correlation between attitudes towards WCF and complexity that can be explained by the fact that positive attitudes toward WCF may suggest particularly careful attention to WCF and thus excessive caution on the part of the students, making them write less and choose simpler language.

5.3 The Effects of WCF Generally on Text Revisions

When examining the effect of unfocused WCF generally on CAF on text revisions, the results showed that there was a statistically significant difference between the feedback groups after controlling for aptitude (LLAMA B, F), attitudes and proficiency, and the effect size is medium according to Cohen (1988). The MANCOVA showed that there was a significant difference between gains from the pre-test to the re-test of the CF groups for fluency and lexical diversity, and both have medium effect sizes according to Cohen (1988). Accuracy, complexity and complex nominals per clause; however, were non-significant.

First, the results of this study showed that there were no effects of the WCF on accuracy on text revisions. Unlike the present study, Truscott and Hsu (2008) and Van Beuningen et al.'s (2008) studies of unfocused feedback found that on text revisions, students who received WCF in the form of underlining errors improved in accuracy. The main difference between the study of Truscott and Hsu (2008) and the present one, however, is that in Truscott and Hsu's study, students had access to their original texts with the feedback while writing the text revision. In this study, the original texts were taken away just before the students re-wrote the text, so they could not copy their original text. The similarities between the study of Van Beuningen et al. (2008), and the present one is that the participants L1 was Arabic. The difference between their study and this one, is that in their study, just like in Truscott and Hsu's (2008) study, the

students copied their texts and had the original version with the WCF on hand when revising the text. This difference in methodology between Van Beuningen et al.'s (2008) study and the present one would most likely explain the variance in results.

Van Beuningen et al. (2012) in another study of unfocused feedback, also found gains in accuracy on text revisions. The methodology of their study was rather similar to the present study, but there are some factors that could explain the different results. First, the L1 of the participants in their study was different than in the present study. Second, the age of the participants - secondary school students - was different to those in the present study. Ferris and Roberts (2001) also found that WCF, in this case direct and indirect feedback, improved the overall accuracy of text revisions. Their study involved giving indirect WCF (codes and underlining). The result showed that ESL college students who received WCF produced significantly improved revised texts with fewer errors than those who did not receive feedback. It is important to note that unlike the present study, the study of Ferris and Roberts has been classed as a study of unfocused WCF by some researchers, such as Pourdana et al. (2021), because it only looked at five grammatical error categories. Thus, the classification of the work as a study of unfocused WCF, could be questioned and could explain the difference in results.

Bonilla López et al.'s (2018) study on unfocused feedback exhibited different results to this study, whereby the participants in their study improved in accuracy on text revisions. The major difference between this study and that of Bonilla López et al. (2018) is the way accuracy was measured. Bonilla López et al. (2018) used the number of errors that were successfully corrected during text revision divided by the total number of errors in the initial text, whereas this study used the percentage of error free t-units and errors per 100 words. This could be a reason for a small amount of variance in results regarding accuracy, but not the main reason as both involved relative error counts. Another difference between the studies could be that the participants in Bonilla López et al.'s (2018) study had a mean English proficiency level that

was lower intermediate as measured by the Oxford Quick Placement Test, while in this study the mean proficiency level also measured by the Oxford Quick Placement Test was advanced. A further difference was that in Bonilla López et al.'s (2018) study, the participants had been told from the start of the study that, the texts they wrote could become drafts of future graded compositions at the end of the study whereas in the present study, the texts the participants wrote, would not count towards their grade for the course.

Another recent study by Nicolás–Conesa et al. (2019), also found that the groups receiving unfocused WCF outperformed the control group regarding accuracy on text revisions. In this study that took place at a Spanish university, 46 English majors participated in a pretest–posttest design, with two treatment groups who were given direct or indirect WCF and asked to process it via written languaging. Languaging is in its written form, and defined as “any language noted by learners to reflect on their language use, with or without metalinguistic terminology” (Ishikawa, 2013, p. 220). The study also included a control group that wrote and rewrote their texts but also engaged in languaging. The main methodological difference between this study and the one of Nicolás–Conesa et al. (2019), is that theirs included languaging for all groups. They note that “the potential combined effect of WCF and languaging in our data is certainly a concern, given that we did not include a group that received WCF and did not engage in languaging” (p.853). Since there was no control group, it is impossible to tell if the improvement was due to the WCF or languaging, and this could explain the difference in results between this study and theirs.

Regarding complexity and lexical diversity, Van Beuningen's (2012) study looked at if WCF would improve structural complexity and lexical diversity in text revisions and found that the practice group that did not receive WCF (but wrote a new text instead) displayed structurally less complex writing than the writing of pupils which received direct or indirect WCF. They also found that all pupils who revised their text based on the WCF outperformed

the practice group regarding the measure of lexical diversity. However, the fact that the pupils in the practice group did not write a text re-write, but wrote a new text, makes the comparability of this study and theirs problematic because it is difficult to regard the practice group as a real control group.

In conclusion, there are studies on unfocused WCF that have found improvements in accuracy on text revisions, but the present study has not found gains in accuracy on text revisions. All the other studies have an element in common in that there is a difference in their methodologies, and this study, which likely lead to their variance in results. Some studies allowed students to have access to their original texts with the feedback while writing the text revision, or were classed as unfocused WCF. However in reality, they looked at only a few grammatical error categories, had lower proficiency levels and/or allowed the WCF drafts to be included in their grades, or combined WCF with languaging. There is also a lack of comparable studies on the effects of WCF on fluency, complexity, and lexical diversity.

5.4 The Effects of Different Types of WCF on Text Revisions.

When looking at which of the types of unfocused corrective feedback has a greater influence on the fluency and lexical diversity of student writing on revised tasks, the results showed that direct and metalinguistic feedback lower fluency, but not lexical diversity. The results also showed that indirect feedback lowers lexical diversity, but not fluency. A form of trade-off condition that is dependent on feedback type was found, which has implications for teaching.

Indirect feedback is potentially easier to ignore, possibly due to it being less explicit than direct feedback and in the direct and metalinguistic feedback groups, the increased monitoring could have reduced fluency. Furthermore, the students may have also engaged more with the direct and metalinguistic feedback as it requires less effort to determine what the error is than

indirect feedback. There can also be a negative impact of unfocused WCF on student motivation (Lee, et al., 2018). The lack of progress in fluency for the direct and metalinguistic feedback group, but not the indirect group, could be due to the simple fact that the more red ink on the page, the more motivation is negatively affected. Thus, when working on building fluency with students, using unfocused indirect WCF on text re-writes would be the preferable option.

Another important result was that indirect feedback lowered lexical diversity, but direct and metalinguistic feedback did not. This is possibly because direct and metalinguistic feedback would provide students with the correct word - or in the case of metalinguistic feedback, let the students know the error was due to the word chosen, but indirect feedback would not. Therefore, the positive effect of improving lexis due to the direct and metalinguistic feedback may be larger than the possible effects WCF has on lowering lexical diversity, due to increased caution and monitoring by students. In the indirect group, the positive effect might not be strong enough and thus lexical diversity would only be lower in this group. As elaborated on in the previous paragraph, Van Beuningen's (2012) study examined if WCF would improve structural complexity and lexical diversity in text revisions, and found that the practice group that did not receive WCF had structurally less complex writing than the writing of pupils who received direct or indirect WCF. However, as explained above, the fact that the practice group wrote a new text rather than a re-write, makes it difficult to consider this as a control group.

Currently, there are no previous studies on fluency on revised tasks, that are comparable with this study and there are very few studies on the effects of WCF on complexity and lexical diversity on revised texts.

5.5 The Effects of WCF Generally and the Effects of Different Types of WCF on New Texts.

The results of the effects of WCF on CAF on new texts, show that generally WCF does not lead to an increase in CAF, compared to no feedback on new tasks. When looking at which of the types of unfocused corrective feedback has a greater influence on the accuracy, complexity, and fluency of student writing on new tasks, the results show that on new tasks, none of the feedback types has a greater effect than each other, or no feedback.

Studies with similar results are rather scarce, since relatively few studies have examined different feedback types on the efficacy of unfocused feedback on new texts. Although some look at general measures of accuracy, there are none that have looked at all three CAF components. In Truscott and Hsu's (2008) study of unfocused WCF, although learners did show improvement in accuracy from unfocused indirect feedback in the revision task, they did not show any improvement on a new writing task. Truscott and Hsu (2008) thus concluded that successful error reduction on the revision cannot be taken as evidence of long-term learning. In another study by Van Beuningen et al. (2008), the direct WCF group had gains in accuracy in new texts, but the gains were not significant. The authors explain the result by the fact that there may have been a reduction of students' motivation as the study continued. Other studies exhibit results that - although not exactly the same as this study - are similar in certain aspects. Hartshorn et al. (2010) compared the effects of indirect unfocused WCF on fluency. They discovered that the treatment group was slightly less fluent than the control group in the post-test, (where participants wrote a new text). Hartshorn et al. (2010) note that a reason for this is that when students try to write more accurately, the fluency of their writing may be inhibited due to monitoring their production more carefully. Although this result is different to this study, which did not find a significant result, the similarity is that the WCF did not lead to improvements in fluency in either study.

Another study of unfocused WCF on new texts that had similar results to this study, whereby none of the feedback types fared better than the other or the control group, is that of

Karim and Nassaji (2018) which only looked at accuracy. In Karim and Nassaji's (2018) study, students were randomly divided into four groups: direct; underline only; underline and metalinguistic; and a control group. Participants produced three pieces of writing from different picture prompts and revised them over a three-week period. On the sixth week, all WCF groups wrote a new text from a new picture prompt. The results were that on new texts, all accuracy gains were non-significant. Karim and Nassaji (2018) suggest that this could be due to unfocused feedback which targeted a broad range of linguistic features, and thus may have caused cognitive overload for learners and interrupted their feedback processing.

Few studies comparing feedback types have investigated the effectiveness of unfocused feedback on grammatical accuracy or other CAF measures, but those with different results to this study include Van Beuningen et al. (2012), who found that unfocused feedback led to improved accuracy on new texts. In their study, the positive effect of unfocused feedback was visible on new texts, written four weeks later. The contrast between this study's findings and those reported by Van Beuningen et al. (2012) could be due to the research design, especially with regards to the writing tasks. Van Beuningen et al.'s (2012) adolescent participants wrote a short e-mail for twenty minutes to a friend explaining the class topic to a fictional classmate who was absent during the explanation of the topic. In contrast, the students in this study were asked to write four opinion essays about personal experiences. Due to the task-related differences, this could be an explanation for the varying results. The language used in writing a short e-mail to a friend may be more repetitive and less advanced than the language used in writing an academic opinion essay, and thus there was a greater likelihood for improvements to be made.

Bonilla López et al. (2018) found that direct corrections improved accuracy on new texts compared to the other treatment groups, but the difference between their study and the present one is the interaction between the research and the graded coursework. They note that "all

learners had been informed from the start of the study that because the topics and tasks were part of the curriculum, the texts could eventually become drafts of future graded compositions at the end of the study if this was deemed desirable by their instructor” (p.826). This may have increased extrinsic motivation, which was different to the present study as the essays students wrote were not graded and would not be part of their grade in any form – which could thus explain the variance. Therefore again, the combined effect of increased motivation and direct WCF was strong enough to improve accuracy.

The study by Nicolás–Conesa et al., (2019), also found that on new texts there were positive significant gains in accuracy from the combined effect of WCF and written languaging. These gains were larger in the direct WCF group compared the indirect group, but both groups had significant gains compared to the control group. However, importantly, all their groups made use of written languaging, which the present study did not. Therefore there is a possibility (as explained above regarding text revisions) that the combined effect of written languaging and WCF was the reason for the significant gains, and thus the comparability between their study and the present one is questionable. Nicolás–Conesa et al. (2019) argue that the languaging session made possible the noticing and retention of corrections needed for language learning and thus could cause the positive significant gains on the new texts.

The only studies on the effect of WCF on new texts regarding complexity and lexical diversity that have been published are that of Fazilatfar et al. (2014) and Van Beuningen et al. (2012) that have looked at both issues, and that of Hartshorn et al. (2010) which only looked at complexity. Fazilatfar et al. (2014) examined the effect of unfocused WCF on complexity and found increases in complexity and lexical diversity on the final essay written by the treatment groups receiving WCF. Fazilatfar et al.’s (2014) study had a major methodological difference regarding the amount of exposure to WCF compared to this study, and that has possibly contributed to the difference in results. In Fazilatfar’s (2014) study, the treatment

group receiving unfocused corrective feedback were given feedback on ten compositions; all new texts and not revisions, which is a far greater amount of times than in this study. This extra amount of exposure to WCF may have increased the effect of WCF compared to the three times received in the current study. Although repeated tasks and multiple treatment sessions of WCF can lower motivation, which is what may have happened in the present study. In the study of Fazilatfar et al. (2014) it may be the case that the motivation of the students was higher than the motivation of the students in the present study. Thus, the negative effect on motivation of repeated tasks and multiple WCF treatment sessions was not strong enough to negate effect of the effects of multiple treatments of WCF in increasing complexity and lexical diversity. The reason why it is possible that the students in Fazilatfar et al.'s (2014) study had greater motivation than in the present study is that the students in their study came from students enrolled in conversation classes in an English institute. The authors note that the students had participated in these classes to improve their communicative competence ability (but also completed writing assignments at the end of the lessons), because they found it very difficult to do so in their formal classes in schools and universities. Thus, if these students were attending supplemental English classes, it is more likely that they would have higher motivation than the sample in the present study, who were in a required class. Van Beuningen et al.'s (2012) study did not find any significant between-groups differences on measures of structural complexity and lexical diversity on new texts produced by the participants, and they conclude that their data did not confirm Truscott's (2007) assumption that WCF would lead to simplified writing. Their study is similar to the present study in that complexity and lexical diversity did not improve in the groups given WCF. In Hartshorn et al's (2010) study, however, the treatment group wrote less complex writing than the control group. Hartshorn et al. (2010) suggest that this is due to the increased monitoring of accuracy, which would cause a trade-off

with complexity to occur. At present, there are no comparable studies that look at the effects of WCF on fluency on writing.

In conclusion, there are studies of unfocused WCF that similar to this study, did not find effects of the WCF on CAF on new texts. This could possibly be due to either a reduction in motivation, increased monitoring, or cognitive overload. However, there are a few studies of unfocused WCF that did find positive effects on CAF, but they exhibited differences in the research design. For example, using e-mails with informal language, and including a motivational effect of allowing the texts to eventually become drafts of future graded compositions. This includes languaging or multiple treatment sessions, and thus including elements from these studies could generate effects on CAF from unfocused WCF.

5.6 General Considerations Regarding Unfocused Feedback

Several researchers (Sheen, 2007; Bitchener, 2008; Ellis et al., 2008; Van Beuningen, 2011; and Van Beuningen, et al., 2012) have argued that since some studies of unfocused WCF did not show significant improvements in performance on revised or new tasks, using focused feedback - which almost always show effects, targeting specific error type or limited error categories - would be more effective. Since in the present study, on both revised texts and new texts, no significant gains in CAF were made, it could be that because the present study provided WCF on wide range of linguistic features at the same time, which might have created a cognitive overload for the participants. This cognitive overload may have possibly prohibited them from processing the WCF, and applying it to the new texts and even the text revisions they wrote. Sheen et al. (2009) argue that when learners are exposed to the correction of a large variety of grammar features, they might have greater difficulty in processing different error types while also retaining the feedback effectively. They argue that the focused approach may

be superior to unfocused feedback, since unfocused feedback can become overburdening for the learner because of the need to attend to various error types. Those who favour focused feedback have suggested that targeting a specific error type is more effective than unfocused error correction as it fails to facilitate acquisition because L2 learners have a limited processing capacity (Bitchener, 2008; Ellis et al., 2008). However, research on unfocused WCF is still important, since unfocused feedback is the form of WCF that most writing teachers give to their students (Ellis et al. 2008). The studies of unfocused WCF that found effects on CAF exhibited differences in their research designs compared with the present study - for example by using e-mails with informal language, including a motivational effect of allowing the texts to eventually become drafts of future graded compositions and including languaging, or multiple treatment sessions. These modifications could create effects that are stronger than the negative effect of cognitive overload and thus lead to positive effects of unfocused WCF.

Having negative gains in CAF measures with groups receiving unfocused feedback can be further understood by looking at the studies that examined the differences between focused and unfocused feedback. A study by Farrokhi and Sattarpour (2012), for example, that looked at the difference between 60 students in direct focused and direct unfocused WCF groups and a control group, yielded superior effects for focused feedback over unfocused feedback. The authors found that focused corrective feedback increased learners' grammatical accuracy in L2 writing more effectively.

The lack of significant gains in CAF in new texts in this study could also be due to the lack of learners' successful intake of WCF during revision, which can also happen with focused feedback, but would be further amplified by the information overload of unfocused feedback. Long (2007) and Van Beuningen, (2012) all noted that learners' successful uptake of feedback does not guarantee long-term acquisition and they point out that errors might not be able to be corrected due to fossilisation.

Various pedagogical implications arise from the findings that will be covered in the following chapter.

Chapter 6: Conclusions, Theoretical and Pedagogical Implications

The present study aimed to find the effects of unfocused direct, indirect, and metalinguistic written corrective feedback on the development of complexity, accuracy, and fluency in EFL students' academic writing. The study also attempted to discover if the moderating variables of aptitude, attitudes, and proficiency had a role to play in the uptake of WCF. To achieve this, aptitude was operationalised by the LLAMA_B and F test, and attitudes by way of a questionnaire and proficiency by the Oxford Quick Placement test. The study also examined the interrelationship of CAF measures. The following CAF measures were included: fluency, accuracy, and complexity, which was made up of general syntactic complexity, complexity by phrasal elaboration, sentential syntactic complexity and lexical diversity. The participants were instructed to write new texts and text re-writes and were given four rounds of WCF. The texts the participants wrote were argumentative essays and the feedback types given were direct, indirect, and metalinguistic.

6.1 Summary of Findings

First, the study aimed to explore the interrelationship of CAF measures and correlations were found (for the re-test to the post-test), whereby fluency and complexity showed a weak positive correlation, thus showing that it is possible for students to improve in both fluency and complexity without a trade-off occurring. A weak positive correlation between accuracy and complexity was also found, which revealed that students can both improve in accuracy and complexity without trade-offs occurring.

Second, the study examined the effect of WCF on CAF with text revisions, and discovered that unfocused WCF did not lead to an increase in the accuracy, complexity, and

fluency of student writing on revised tasks, compared to no feedback. The results showed that there were no significant gains in CAF on revised tasks compared to the control group. The study also looked at the effects of the different types of WCF on CAF on text revisions, and found the direct group and the metalinguistic group had negative gains in fluency compared to the control group and indirect group. Thus, it was concluded that direct and metalinguistic feedback have the effect of reducing the amount students write on revised tasks. This could be due to indirect feedback being potentially easier to ignore and the fact that direct feedback and metalinguistic feedback may increase monitoring, and thus could reduce fluency. Regarding lexical diversity, there was also a significant difference between the indirect, direct and metalinguistic group, showing that indirect feedback lowered the students' lexical diversity on text revisions. This could possibly be because direct and metalinguistic feedback would provide students with the correct word or in the case of metalinguistic feedback, let the students know the error was due to the word chosen, whereas indirect feedback would not. Therefore, on text revisions a form of trade-off that is dependent on feedback type was found.

Third, when looking at the effects of WCF on new texts generally, as well as the effects of the different types of WCF on new texts, the results showed that unfocused corrective feedback did not lead to an increase in CAF compared to no feedback and that none of the feedback types had a greater effect than each other or no feedback. Therefore, on new texts, if improvements in CAF are the goal of unfocused WCF, then differences in the approach must be necessary, such as increasing the amount of exposure to the WCF or including other interventions, such as written languaging.

Finally, the correlations between IDs and CAF showed a weak negative correlation with complexity and students' attitudes toward corrective feedback (for the re-test to post-test), showing that students who had positive attitudes towards corrective feedback wrote less complex writing. There was also a weak negative correlation found with fluency (for the pre-

test to post-test) and thus students who had positive attitudes toward corrective feedback wrote less than those that had negative attitudes towards it. This can be explained by the fact that positive attitudes toward WCF may suggest particularly careful attention to WCF and thus excessive caution on the part of the students making them write less and choose simpler language.

6.2 Theoretical Implications

The theoretical implications of this study are interesting. The results of the CAF correlations in this study did not show trade-off effects, but did show connected growth, for example the positive correlation with accuracy and complexity that was found. Dynamic Systems Theory, also known as Complexity Theory has been applied to second language acquisition (Larsen-Freeman, 1997; De Bot, Lowie, & Verspoor, 2007) and has a hypothesis that allows for both trade-offs and connected growth. This theory assumes that cognitive resources are limited; however, they are connected and may be compensatory. Since all variables in the system are interrelated, changes will affect all the other parts of the system. The results of the CAF correlations in this study did not show trade-off effects, but did show connected growth, thus the results of the study support Dynamic Systems Theory, which allows for connected growth.

The positive correlation with accuracy and complexity that was mentioned above can also be linked with Robinson's Cognition Hypothesis (2011). Robinson points out that tasks will either promote fluency, or complexity and accuracy. According to Robinson (2011), simple tasks are expected to promote fluency, but not complexity or accuracy, and complex tasks promote accuracy and complexity, but not fluency. This study does not actually test

Robinson's Cognition Hypothesis directly because it does not include manipulation of task complexity; however, the results, to some extent, support Robinson's (2011) Cognition Hypothesis since accuracy and complexity both increased. However, Robinson's Cognition Hypothesis also expects a trade-off between fluency and complexity, which was not found in the results.

The results of the CAF correlations did not show trade-off effects, but displayed connective growth, thus supporting Complexity and Dynamic Systems theory. However, in this study, when looking at the relationship between ID variables and CAF, the results showed that students with positive attitudes toward WCF (for the pre-test to post-test data) had a weak negative correlation with fluency; therefore, students who had positive attitudes toward corrective feedback produced fewer words. These results could support Krashen's (1982) Monitor hypothesis. The Monitor hypothesis (Skehan, 1998) states that corrective feedback could reduce fluency if it caused the learners to monitor their production more carefully.

In conclusion, this study demonstrates theoretical implications, such as lending support to Dynamic Systems Theory (Complexity Theory) as well as a more specific hypothesis: Krashen's (1982) Monitor Hypothesis. As well as theoretical implications, the study also contains some noteworthy pedagogical implications.

6.3 Pedagogical Implications

As well as theoretical implications, this study also contains pedagogical implications that will be relevant to teachers of academic writing classes at universities. The recommendations apply to the context in which this study was conducted where the participants came from freshmen

academic writing classes at a university in the Middle East. The main L1 of the participants was Arabic, and their L2 English level according to the ALTE levels ranged from lower intermediate to very advanced. Connected growers, the correlations between attitudes and corrective feedback, the effects of the different types of WCF on text revisions and new texts, and the effects of WCF generally on text revisions and new texts - are findings from the study that lead to pedagogical implications.

6.3.1 Connected Growers

The study showed that it is possible for students to improve in both fluency and complexity without a trade-off occurring, and that students can both improve in accuracy and complexity without trade-offs occurring. This means that when teaching academic writing, teachers should try not to worry about trying to improve students' complexity at a detriment to fluency or accuracy when writing new texts. Therefore if they are not working on text revisions, educators do not need to worry about what aspect they wish to focus on in their teaching of academic writing.

6.3.2 Correlations between Attitudes and Corrective Feedback

The study also showed that students who had positive attitudes towards corrective feedback wrote less complex writing, and that students who had positive attitudes toward corrective feedback wrote less. The implication of this in the classroom is that teachers may wish to consider a selective approach to WCF rather than a one-size-fits-all policy for the class. When a teacher is working on improving students' written fluency, if students have positive attitudes towards WCF, the teacher may decide to give less WCF to these students to improve their

fluency and complexity. They could then give more WCF to the students with negative attitudes toward it, as it will be less likely to impede their fluency and complexity development. Thus, the first action the teacher should take at the beginning of the semester would be to determine what their students' attitudes towards WCF are, and then proceed from there by identifying which students will benefit the most from the WCF, and then how much to give them. However, there is an issue regarding the validity of an attitude assessment regarding the length of time it would be valid, due to the fact that student attitudes may change over time, but this approach could be practical if the time frame is short. Furthermore, there is also an issue of practicality and motivation of the teacher, if they are willing to go to such lengths for their students.

6.3.3 The Effects of the Different Types of WCF on Text Revisions and New Texts

When looking at the effects of the different types of WCF on text revisions, the study shows that indirect feedback lowered lexical diversity, but direct and metalinguistic feedback did not. This is possibly due to the fact that direct feedback would provide the student with the correct word form and in the case of metalinguistic feedback, let the students know the error was due to the word chosen. Indirect feedback would not suggest ways to improve lexis and thus students would have less chance to improve their lexical diversity. The possible effects indirect WCF have on lowering lexical diversity would mean that in the classroom, when giving unfocused WCF on text revisions, teachers may wish to make use of direct and metalinguistic feedback and not give indirect feedback to make sure students' lexical diversity is not negatively affected. However, caution should also be advised as the study also shows that on text revisions, unfocused direct and metalinguistic WCF lower fluency and thus a form of trade-off dependent on feedback type was found. Therefore, if the teacher is working on an activity where the students write new texts, as mentioned in 6.1, they do not need to worry about what

aspect they wish to focus on in their teaching of academic writing. However, if the teacher is giving WCF on an activity or assignment that is a text re-write, they can then decide if they wish to prioritise fluency or lexical diversity and decide on the type of WCF to give the students based on that.

In conclusion, when working on improving lexical diversity in students, using unfocused direct and metalinguistic WCF on text re-writes would be the preferable option. However, due to the trade-off dependent on feedback type on text re-writes the teacher must decide what they are focusing on improving, and adjust the feedback type in the appropriate way to achieve the desired goal. Regarding new texts, there would be no difference between effects dependent on feedback type, thus it would not matter what the teacher uses.

6.3.4 The Effects of WCF Generally on Text Revisions and New Texts

The study also examined the effect of WCF generally on CAF on text revisions and found that unfocused WCF did not lead to an increase in the accuracy, complexity, and fluency of student writing on revised tasks or new tasks, compared to no feedback. Although some may argue that the concept of giving unfocused WCF should be questioned and propose giving focused rather than unfocused feedback, which is also controversial. Lee (2019), notes that focused WCF is much less overwhelming for students, both emotionally and cognitively and thus at first, may appear to be the better way of providing WCF; however, the main issue is that if focused WCF is possibly more beneficial, it is not as easy to implement as unfocused WCF. To make focused WCF gradual and contingent on students' developmental level, Lee (2019) points out that teachers need to make numerous decisions in advance and furthermore, if in a coordinated writing programme, this should be in collaboration with colleagues, who may or may not agree

with what to focus on. Purposefully planning focused WCF could be much more demanding and time-consuming for teachers. Teachers would need to decide what to include, not just based on personal preferences, but based on sound research when choosing what they will and will not correct when providing focused WCF, and the justifications for their choices.

It may also be beneficial to look at this study compared to studies that showed improvements in students' writing on text revisions and see if elements of the studies could be adapted for in-class use. Some studies on unfocused WCF did show improvement in student writing, but they were different methodologically to what happens in a regular classroom, and dissimilar to this study. The present study utilises unfocused feedback on all errors, but the study of Ferris and Roberts (2001), also classed as a study of unfocused WCF, only looked at five grammatical error categories and found that WCF improved the overall accuracy of text revisions. The present study used unfocused WCF, but did not find improvements, so only using less error categories than the present study in the classroom could be a possible starting point as it could lead to increases in accuracy. The current study did not use written languaging in combination with WCF, but the study of Nicolás–Conesa et al. (2019) did, and found that the groups receiving unfocused WCF outperformed the control group (that also engaged in languaging) regarding accuracy on text revisions. Therefore, although the current study did not include languaging, making use of it in the classroom with unfocused WCF may increase accuracy and could be adopted by practitioners to help their students.

Another important pedagogical implication is that the approach to giving WCF on new texts and text revisions should also be dependent on the writing task. In the present study, the students were asked to write four opinion essays about personal experiences, but in Van Beuningen et al.'s (2012) study, participants wrote a short e-mail to a friend and found that the WCF led to positive effects on accuracy. These task-related differences could be an explanation for the varying results. The language used in writing an academic opinion essay is less

repetitive and more advanced and thus, dependent on the task, the teacher could decide on which type of WCF to provide.

Teachers of academic writing could also increase the amount of unfocused WCF treatment for positive effects. In the present study, the students received four rounds of unfocused WCF during the experiment. However, another study by Fazilatfar et al. (2014) examined the effect of unfocused WCF on complexity and found increases in complexity on the final essay written by the treatment groups receiving ten unfocused WCF sessions. Repeated tasks and multiple treatment sessions of WCF can lower motivation, but it may be the case that the motivation of the students may be higher to offset these demotivating effects. Although the present study did not offer ten sessions of WCF, having more than the four rounds of unfocused WCF in an academic writing course is something that practitioners should consider. The issue with multiple rounds of WCF is that the course may not be long enough for this to be practical or feasible, however in a year-long course, this is a method that could be used.

In conclusion, the pedagogical implications of this study are interesting. Firstly, teachers should not worry about trying to improve students' complexity at a detriment to fluency or accuracy and therefore, they do not need to be overly concerned about what aspect to prioritise in their teaching of academic writing, if they are working with activities where students are writing new texts. Secondly, at the beginning of the semester, teachers should try to determine what their students' attitudes towards WCF are, and then will know who will benefit the most from the WCF. Thirdly, regarding which type of unfocused WCF should be given for text revisions; if the teacher is focusing on fluency or lexical diversity, they must choose which type to give. Thus, if the teacher is focusing on lexical diversity, they should not give indirect feedback, since it has the effect of lowering lexical diversity on text revisions, and should provide direct or metalinguistic feedback. If the teacher is prioritising fluency, then they should

not give direct or metalinguistic feedback, but should provide indirect feedback. Finally, regarding unfocused WCF on text revisions and new texts in general - although there are calls for practitioners to use focused WCF as it appears to work better in experimental studies - there are many reasons why in the real world, this may be more problematic than using unfocused WCF. Better still, a more blended approach between the focused and unfocused that also includes languaging and more treatment sessions than usual, could be adopted.

6.4 Limitations of the Study and Suggestions for Further Research

To begin with, the study had limitations that should be acknowledged. The study was limited in that it made use of only one type of task. The persuasive/argument essay and other studies using creative or various types of writing, such as writing e-mails to friends, could also be undertaken to see if certain forms of tasks or different rhetorical modes affect the uptake of WCF. Research with a similar methodology to the present study, but a different type of task may yield varying results, because the type of language used depends on the task type.

Second, in the present study a further limitation was that the WCF groups received corrections on their texts, which were given to them and then removed before they wrote the text revision and the new texts. This method was employed to prevent students copying their texts; however, this lacks ecological validity as it is not the normal practice in writing classes, because it is unusual in classrooms for teachers to take away students' feedback after distributing it. Students may not bring their feedback to class when writing assignments, but that is normally their decision to do so and not due to it being not available. Furthermore, although it may be considered an uncommon practice for students to completely re-write their text after having received WCF, in academic writing classes, students normally complete drafts and submit them for teacher review and then continue to work on the same text. In the

researcher's academic writing classes, the drafts are given indirect unfocused WCF whereby students are shown where the grammatical and lexical errors lie in their texts. Unfortunately, there is no easy way to rectify this issue and there are no suggestions for further research because if students are allowed to retain their corrected texts, they would simply copy the corrections onto their text re-write and thus it would not be an exercise in applying what they learned from the WCF, but would just be a simple exercise in copying.

The third major limitation was that the majority of unfocused WCF research to date - this study included - has investigated the effectiveness of error correction by examining group performances on global accuracy, complexity and fluency measures. However, sometimes group data that does not show improvements, in fact masks individual data where in some cases improvements have taken place. Thus, when looking at group data, it may appear that there are no improvements, but when looking at individual scores, there could be some students who have actually made improvements. Furthermore, adding a qualitative aspect to the study through focus group interviews could be beneficial. When looking at group differences, the surprising finding of negative attitudes toward WCF being associated with greater fluency, for instance, could be further investigated.

Another possibility for further research would involve trying to find out how many error categories of WCF would lead to cognitive overload. The study of Ferris and Roberts (2001) only used five error categories, and the WCF improved the overall accuracy of text revisions. This could be used as a possible starting point and more error categories could be used until positive effects diminish or are no longer found; to discover a certain "sweet-spot" whereby students are not overloaded cognitively, but are given more categories of WCF than in the study of Ferris and Roberts (2001). It would also be possible to measure students' cognitive overload threshold on an individual, as the threshold varies between students.

Another limitation of the study was that it only provided four rounds of WCF, which may be insufficient to establish whether regular feedback had an effect on students' accuracy and complexity and fluency. However, it was against the University's policy to work with the participants for more than one academic semester, and also taking into account the requirements of the course, students could not be given extra assignments just for the sake of research, and thus this research had to be carried out within the period of time allowed. Further research that provides more rounds of WCF, possibly for a full academic year would be interesting as it may show improvements in CAF from unfocused WCF. In the same way, discovering how many WCF treatment sessions may or may not have positive effects on CAF is another area researchers could explore. In the study of Fazilatfar et al. (2014), ten unfocused WCF sessions showed positive effects on complexity on the final composition. However, it could be that a greater number of feedback sessions than in the present study, though less than in the study by Fazilatfar et al. (2014), could still lead to positive effects on CAF, while still being feasible in the classroom with limited time. Thus yet again, finding the mid-point regarding how many treatment sessions, could be useful.

Furthermore, research would also need to take place in other contexts in both EFL and in other languages beyond English. Presently, the findings across contexts might not be readily comparable, or cannot be automatically transferred to any other contexts.

In conclusion, this study on the effects of unfocused WCF on students' CAF shows that there is still scope for research in this area as there are several unknowns. The purpose of the study was to contribute to the field of WCF and CAF research by looking at the scarcely studied unfocused, rather than focused feedback, of which many studies already exist. Furthermore, by looking at both text revisions and new texts and well as including the moderating variables of L2 proficiency, attitudes and aptitude as well as including complexity as a CAF measure, and showing that connected growers in CAF are possible, this study brings something new to the

field. The results of this study, however, still provide useful information for practitioners. It is hoped that future studies consider these limitations as well as the findings from this study and their interpretations when designing new studies that will be able to shed further light on the applied and theoretical dimensions of the effects of unfocused WCF.

References

- Abdollahzadeh, S. & FardKashani, K. (2012). The effect of task complexity on EFL learners' narrative writing task performance. *Journal of English Language Teaching and Learning*, 8(1), 1-28.
- Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, 30, 481–509. <http://dx.doi.org/10.1017/S027226310808073X>.
- Adel, S., Alwi, N. M., & Aloesnita, N. (2014). The Effects of an Increase in Task Complexity on Learners Written Productions Via Wiki. *International Journal of English and Education*, 3, 2278-4012.
- Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, 15(1), 35-59. <https://doi.org/10.1177%2F1362168810383329>.
- Ai, H., & Lu, X. (2010). A Web-based System for Automatic Measurement of Lexical Complexity. Paper presented at 27th Annual Symposium of the Computer Assisted Language Consortium (CALICO-10). http://www.personal.psu.edu/hua126/papers/calico10_ai_lu.pdf.
- Alimohammadi, B., & Nejadansari, D. (2014). Written Corrective Feedback: Focused and Unfocused. *Theory And Practice In Language Studies*, 4(3), 581-587. [doi:10.4304/tpls.4.3.581-587](https://doi.org/10.4304/tpls.4.3.581-587).
- Ashwell, T. (2000). Patterns of Teacher Response to Student Writing in a Multi Draft Composition Classroom: Is Content Feedback Followed by Form Feedback the Best method. *Journal of Second Language Writing*, 9(3), 227-257. [https://doi.org/10.1016/S1060-3743\(00\)00027-8](https://doi.org/10.1016/S1060-3743(00)00027-8).
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bacon, S.M., & Finnemann, M.D. (1990). A study of attitudes motives and strategies of university foreign language students and their disposition to authentic oral and written input. *Modern Language Journal*, 74(4), 459-473. <https://doi.org/10.1111/j.1540-4781.1990.tb05338.x>.
- Bakri, H. (2016) The Role of Individual Differences in Second Language Writing Corrective Feedback. *Arab World English Journal (AWEJ)*, 6(4). <https://dx.doi.org/10.2139/ssrn.2843943>.
- Barcelos, A. M. F. (2003). Researching beliefs about SLA: A critical review. In P. Kalaja and A. M. F. Barcelos (Eds.), *Beliefs about SLA: New research approaches* (pp. 7-33). Dordrecht: Kluwer Academic Publishers.

- Benson S., & DeKeyser R. M. (2018). Effects of written corrective feedback and language aptitude on verb tense accuracy. *Language Teaching Research*, 23(6), 702-726. <https://doi.org/10.1177%2F1362168818770921>.
- Benson, P., & Lor, W. (1999). Conceptions of language and language learning. *System*, 27(4), 459–472. [https://doi.org/10.1016/S0346-251X\(99\)00045-7](https://doi.org/10.1016/S0346-251X(99)00045-7).
- Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). New York: Oxford University Press.
- Bitchener, J. (2019). The intersection between SLA and feedback research. In K. Hyland & F. Hyland (Eds.), *Feedback in second language writing: Contexts and issues* (2nd ed.) (pp. 85–105). Cambridge University Press.
- Bitchener, J. and Knoch, U. (2010b). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing*, 19(4), 207-217. <https://doi.org/10.1016/j.jslw.2010.10.002>.
- Bitchener, J., & Ferris, D. (2012). *Written corrective feedback in second language acquisition and second language writing*. New York: Routledge.
- Bitchener, J., & Knoch, U. (2009). The value of a focused approach to written corrective feedback. *ELT Journal: English Language Teachers Journal*, 63(3), 204-211. <https://doi.org/10.1093/elt/ccn043>.
- Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten month investigation. *Applied Linguistics*, 31(2), 193-214. <https://doi.org/10.1093/applin/amp016>.
- Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development* (Second language acquisition, 96; Second language acquisition (Clevedon, England), 96). Bristol: Multilingual Matters.
- Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing*, 14(3), 191–205. <https://doi.org/10.1016/j.jslw.2005.08.001>.
- Bitchner, J. (2008). Evidence in Support of Written Corrective Feedback. *Journal of Second Language Writing* 17(2), 102-118. <https://doi.org/10.1016/j.jslw.2007.11.004>.
- Bitchner, J., & Knoch, U. (2008). The Value of Written Corrective Feedback for Migrant and International Students. *Language Teaching Research*, 12(3), 409-431. <https://doi.org/10.1177%2F1362168808089924>.
- Bokander, L. (2020). Language Aptitude and Crosslinguistic Influence in Initial L2 Learning. *Journal of the European Second Language Association*, 4(1), 35–44. DOI: <http://doi.org/10.22599/jesla.69>.

- Bokander, L., & Bylund, E. (2020). Probing the internal validity of the LLAMA language aptitude tests. *Language Learning*, 70(1), 11-47. <https://doi.org/10.1111/lang.12368>.
- Bonilla Lopez, M., Van Steendam, E., & Buyse, K. (2018). The differential effects of comprehensive feedback forms in the second language writing class. *Language Learning*, 68(3), 813-850. <http://dx.doi.org/10.1111/lang.12295>.
- Brezina, V. & Pallotti, G. (2016). Morphological complexity in written ESL texts. *Second Language Research*, 35(1), 99-109. <https://doi.org/10.1177/0267658316643125>
- Brown, D. H. (2000). *Principles of language learning and teaching* (4th Eds.). Pearson Education.
- Brown, H. D. (1994). *Principles of language learning and teaching*. Englewood Cliffs, NJ: Prentice Hall.
- Bulté, B., & Housen, A. (2012). Defining and Operationalising L2 Complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds), *Dimensions of L2 Performance and Proficiency - Investigating Complexity, Accuracy and Fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.
- Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, 26, 42–65. <https://doi.org/10.1016/j.jslw.2014.09.005>.
- Bygate, M. (1999). Quality of language and purpose of task: Patterns of learners' language on two oral communication tasks. *Language Teaching Research*, 3(3), 185–214. <https://doi.org/10.1177/136216889900300302>.
- Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Harlow, UK: Longman.
- Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2010). The role of language aptitude in first language attrition: The case of pre-pubescent attriters. *Applied Linguistics*, 31(3), 443–464. <http://dx.doi.org/10.1093/applin/amp059>.
- Canale, M. (1983). A communicative competence approach to language proficiency assessment in a minority setting. In C. Rivera (ed.), *Communicative competence approaches to language proficiency assessment: research and application*. Clevedon: Multilingual Matters, 107-22.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47. <https://doi.org/10.1093/applin/I.1.1>.
- Carroll, J. (1981). Twenty-five years of research on foreign language aptitude . In K. Diller, (Ed.), *Individual differences and universals in language learning aptitude* (pp. 83 – 118). Rowley, MA: Newbury House.

- Carroll, J. B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education*. Pittsburgh, PA: Univ. of Pittsburgh Press.
- Carroll, J. B. (1971). Implications of aptitude test research and psycholinguistic theory for foreign language teaching. Paper presented at XVIIth International Congress, *International Association of Applied Psychology*, Liège.
- Carroll, J., & Sapon, S. (1959). *Modern language aptitude test*. New York, NY: The Psychological Corporation/Harcourt Brace Jovanovich.
- Carroll, J., & Sapon, S. (2002). *Modern Language Aptitude Test: Manual 2002 Edition*. Bethesda, MD: Language Learning and Testing Foundation, Inc.
- Carroll, J., Sapon, S., Reed, D., & Stansfield, C. (2010). *Manual for the MLAT*. Second Language Testing. N. Bethesda, MD.
- Chambers, F. (1997). What do we mean by fluency. *System*, 25(4), 535-544. [https://doi.org/10.1016/S0346-251X\(97\)00046-8](https://doi.org/10.1016/S0346-251X(97)00046-8).
- Chan, H., Verspoor, M., & Vahtrick, L. (2015). Dynamic development in speaking versus writing in identical twins. *Language Learning*, 65(2), 298–325. <https://doi.org/10.1111/lang.12107>.
- Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing*, 12(3), 267-296. [https://doi.org/10.1016/S1060-3743\(03\)00038-9](https://doi.org/10.1016/S1060-3743(03)00038-9).
- Charge, N., & Taylor, L. (1997). Recent developments in IELTS. *English Language Teaching Journal*, 51(4), 374-380. <https://doi.org/10.1093/elt/51.4.374>.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ :Lawrence Erlbaum Associates.
- Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition*, 26(2), 227-248. [doi:10.1017/S0272263104262040](https://doi.org/10.1017/S0272263104262040).
- Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hyper-nymic relationships. *Language Learning*, 59, 307–334. <https://doi.org/10.1111/j.1467-9922.2009.00508.x>.
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In H. N. Hornberger (Ed.), *Encyclopedia of language and education*, 5(2) (pp. 487–499). New York: Springer.
- de Bot, K. (2008). Introduction: Second language development as a dynamic process. *The Modern Language Journal*, 92, 166-178. <https://doi.org/10.1111/j.1540-4781.2008.00712.x>.

- de Bot, K., Lowie, W., & Verspoor, M. (2007). A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and Cognition*, 10(1), 7-21. doi:10.1017/S1366728906002732.
- de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. *Studies in Second Language Acquisition*, 19(2), 249–276. <http://dx.doi.org/10.1017/S0272263197002064>.
- DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, 22(4), 499–533. doi:10.1017/S0272263100004022.
- DeKeyser, R. M. (2017). Age in learning and teaching grammar. *The TESOL Encyclopedia of English Language Teaching*, 1–6. <https://doi.org/10.1002/9781118784235.eelt0106>.
- DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age-effects in second language acquisition. *Applied Psycholinguistics*, 31(3), 413–438. doi:10.1017/S0142716410000056.
- DeKeyser, R., & Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 395 – 406). New York, NY: Routledge.
- Derwing, M., & Rossiter, J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, 13(1), 1-17.
- Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The Modern Language Journal*, 78(3), 273-284. <https://doi.org/10.2307/330107>.
- Dörnyei, Z. (2001). New themes and approaches in second language motivation research. *Annual Review of Applied Linguistics*, 21, 43-59. Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Dörnyei, Z. (2003). *Individual differences in foreign language learning: effects of aptitude, intelligence, and motivation*. Steve Cornwell and Peter Robinson (Eds.). Tokyo: Aoyama Gakuin University, 2000. Pp. ii + 199.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: individual differences from a dynamic systems perspective. In E. Macaro (Ed.), *Continuum companion to second language acquisition* (pp. 247-267) London: Continuum.
- Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 589-630). Malden, MA: Blackwell Publishing Ltd.

- Ehrman, M. E., & Oxford, R. L. (1995). Cognition plus: Correlates of language learning success. *The Modern Language Journal*, 79(1), 67–89. <https://doi.org/10.2307/329394>.
- Ellis, R. & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition*, 26(1), 59-84. [doi:10.1017/S0272263104026130](https://doi.org/10.1017/S0272263104026130).
- Ellis, R. (1994). A theory of instructed second language acquisition. In N. Ellis (ed.): *Implicit and Explicit Learning of Language*. Academic Press.
- Ellis, R. (2002). Does form-focused instruction affect the acquisition of implicit knowledge? A review of the research. *Studies in Second Language Acquisition*, 24(2), 223-236. <http://dx.doi.org/10.1017/S0272263102002073>.
- Ellis, R. (2004). *Individual differences in second language learning*. In eds. A. Davies, and C. Elder, 525–551.
- Ellis, R. (2005). Principles of instructed language learning. *System*, 33(2), 209-224. <https://doi.org/10.1016/j.system.2004.12.006>.
- Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.
- Ellis, R. (2009). A typology of written corrective feedback types. *ELT Journal*, 63(2), 97-107. <https://doi.org/10.1093/elt/ccn023>.
- Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. New York: Oxford University Press.
- Ellis, R., & Yuan, F. (2005). The effects of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167–192). Philadelphia: John Benjamins.
- Ellis, R., Sheen, Y., Morakami, M., & Takashima, H. (2008). The effect of focused and unfocused corrective feedback in an English as a foreign language context. *System*, 36(3), 353-371. <https://doi.org/10.1016/j.system.2008.02.001>.
- Eslami, E. (2014). The Effects of Direct and Indirect Corrective Feedback Techniques on EFL Students' Writing. *Procedia - Social And Behavioral Sciences*, 98(6), 445-452. <https://doi.org/10.1016/j.sbspro.2014.03.438>.
- Falhasiri, M. (2021). Is Less Really More? The Case for Comprehensive Written Corrective Feedback. *Canadian Journal of Applied Linguistics*, 24(3), 145–165. <https://doi.org/10.37213/cjal.2021.31242>.
- Farrokhi, F., & Sattarpour, S. (2012). The Effects of Direct Written Corrective Feedback on Improvement of Grammatical Accuracy of High- proficient L2 Learners. *World Journal of Education*, 2(2), 49-57. <http://dx.doi.org/10.5430/wje.v2n2p49>.

- Fazilatfar, A.M., Fallah, N., Hamavandi, M., & Rostamian, M. (2014). The Effect of Unfocused Written Corrective Feedback on Syntactic and Lexical Complexity of L2 Writing. *Procedia-Social and Behavioral Sciences*, 98, 482-488. <http://dx.doi.org/10.1016/j.sbspro.2014.03.443>.
- Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development in A. Housen, F. Kuiken, and I. Vedder (eds): *Dimensions of L2 Performance and Proficiency*. John Benjamins.
- Ferris, D. R. (2004). The “Grammar Correction” debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime . . .?). *Journal of Second Language Writing*, 13(1), 49–62. <https://doi.org/10.1016/j.jslw.2004.04.005>.
- Ferris, D. R. (2010). Second Language Writing Research and Written Corrective Feedback in SLA: Intersections and Practical Applications. *Studies In Second Language Acquisition*, 32(2), 181-201. doi:10.1017/S0272263109990490.
- Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be. *Journal of Second Language Writing*, 10(3), 161-184. [https://doi.org/10.1016/S1060-3743\(01\)00039-X](https://doi.org/10.1016/S1060-3743(01)00039-X).
- Ferris, D.R. (1999). The case for grammar correction in L2 writing classes: A response to Truscot (1996). *Journal of Second Language writing*. 8(1), 1-11. [https://doi.org/10.1016/S1060-3743\(99\)80110-6](https://doi.org/10.1016/S1060-3743(99)80110-6).
- Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Mahwah, NJ: Erlbaum.
- Flahive, D. (2010). A reconsideration of “pedagogical implications” and “further research needed” moves in the reporting of second language writing research and their roles in theory building. In T. Silva & P. K. Matsuda (Eds.), *Practicing theory in second language writing* (pp. 126– 158). West Lafayette: Parlor Press.
- Flavell, J. H. (1987). Speculations about the nature and development of metacognition. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, Motivation and Understanding* (pp. 21-29). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Forsberg Lundell, F., & Sandgren, M. (2013). High-level proficiency in late L2 acquisition – Relationships between collocational production, language aptitude and personality. In: Granena, G. and Long, M. (eds.), *Sensitive periods, aptitudes and ultimate attainment in L2*. Amsterdam: Benjamins. 231–256.
- Freed, B. (1995). What makes us think that students who study abroad become fluent? Benjamins. <https://doi.org/10.1075/sibil.9.09fre>.
- Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: Edward Arnold.
- Gee, T. C. (1972). Students’ responses to teacher comments. *Research in the Teaching of English*, 6(2), 212–221.

- Gholaminia, I., Gholaminia, A., & Marzban, A. (2014). An investigation of meta-linguistic corrective feedback in writing performance. *Procedia – Social and Behavioural Sciences*, 116, 316-320.
- Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2), 371-398. <https://doi.org/10.1111/j.1756-8765.2010.01081.x>.
- Granena, G. (2012). Age differences and cognitive aptitudes for implicit and explicit learning in ultimate L2 attainment. Unpublished PhD thesis, University of Maryland, College Park, Maryland.
- Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. What aptitude does the LLAMA measure? In G. Granena, & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment*. Amsterdam: Benjamins.
- Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test In: Granena, G. and Long, M. H. eds. *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment*. Amsterdam: John Benjamins, 35pp. 105–130, DOI: <https://doi.org/10.1075/llt.35.04gra>.
- Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, 29(3), 311–343. <https://doi.org/10.1177/0267658312461497>.
- Gregersen, T., & Horwitz, E. (2002). Language learning and perfectionism: Anxious and non-anxious language learner's reactions to their own oral performance. *Modern Language Journal*, 86(4), 562-570. <http://dx.doi.org/10.1111/1540-4781.00161>.
- Grigorenko, E., Sternberg, R., & Ehrman, M. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal*, 84, 390–405. <https://psycnet.apa.org/doi/10.1111/0026-7902.00076>.
- Guénette, D. (2007). Is feedback pedagogically correct? Research design issues in studies of feedback on writing. *Journal of Second Language Writing*, 16(1), 40-53. <https://doi.org/10.1016/j.jslw.2007.01.001>.
- Guiraud, P.L. (1959). *Problèmes et méthodes de la statistique linguistique* (Dordrecht).
- Gunnarson, C. (2012). The development of complexity, accuracy and fluency in the written production of L2 French. In: Housen, A., Kuiken, F. & Vedder, I. (eds.) *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Halleck, G.B. (1995). Assessing oral proficiency: A comparison of holistic and objective measures. *Modern Language Journal*, 79(2), 223–234. <https://doi.org/10.1111/j.1540-4781.1995.tb05434.x>.

- Harsch, C. (2017). Proficiency. *ELT Journal*, 71(2), 250–253. <https://doi.org/10.1093/elt/ccw067>.
- Hartshorn, K.J., Evans, N.W., Merrill, P.F., Sudweeks, R.R., Strong-Krause, D., & Anderson, N.J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly* 44(1), 84-109. <https://doi.org/10.5054/tq.2010.213781>
- Hauptman, P. C. (1971). A structural approach vs. a situational approach to foreign language teaching. *Language Learning*, 21(2), 235–244. <https://doi.org/10.1111/j.1467-1770.1971.tb00062.x>.
- Havranek, G., & Cesnik, H. (2001). Factors affecting the success of corrective feedback. *EUROSLA Yearbook*, 1, 99. <https://doi.org/10.1075/eurosla.1.10hav>.
- Hedgcock, J., & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing. *Journal of Second Language Writing*, 3(2), 141-163. [https://doi.org/10.1016/1060-3743\(94\)90012-4](https://doi.org/10.1016/1060-3743(94)90012-4).
- Horwitz, E. K. (1990). Attending to the affective domain in the foreign language classroom. In S. Magnan (Ed.), *Shifting the instructional focus to the learner* (pp. 15–33). Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages.
- Horwitz, E.K. (1985). Using student beliefs about language learning and teaching in the foreign language methods course. *Foreign Language Annals*, 18(4), 333-340. <https://doi.org/10.1111/j.1944-9720.1985.tb01811.x>.
- Horwitz, E.K. (1987). Surveying student beliefs about language learning. In A. Wenden & J. Rubin (Eds.), *Learning strategies in language learning*. Englewood Cliffs NY: Prentice Hall.
- Horwitz, E.K. (1988). The beliefs about language learning of beginning university foreign language students. *Modern Language Journal*, 72(3), 283-294. <https://doi.org/10.2307/327506>
- Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4), 461-473. <https://doi.org/10.1093/applin/amp048>
http://www.ehow.com/way_5766640_studies-attitudes-towards-learning-english.html.
- Hyland, K., & Hyland, F. (2006). Contexts and issues in feedback on L2 writing: An introduction. In *Feedback in second language writing: Contexts and issues*. Eds. K. Hyland and F. Hyland. Cambridge: Cambridge University Press. 1-19.
- Ishikawa T. (2007). The effect of manipulating task complexity along the [+/- Here-and-Now] dimension on L2 written narrative discourse. In M. P. García Mayo (ed.) *Investigating Tasks in Formal Language Learning*. Clevedon: Multilingual Matters: 136–156.
- Ishikawa, T. (2006). The effects of task complexity and language proficiency on task-based language performance. *The Journal of Asia TEFL*, 3(4), 1-225.

- Ishikawa, M. (2013). Examining the effect of written languaging: the role of metanotes as a mediator of second language learning. *Language Awareness*, 22(3), 220–233. doi:10.1080/09658416.2012.683435.
- Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct. *Applied Linguistics*, 29(1), 24–49. <https://doi.org/10.1093/applin/amm017>.
- Jiang, W. (2013). Measurements of development in L2 written production: The case of L2 Chinese. *Applied Linguistics*, 34, 1-24. <https://doi.org/10.1093/applin/ams019>.
- Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, 56(2), 1–15. <https://psycnet.apa.org/doi/10.1037/h0093508>.
- Kang, E., & Han, Z. (2015). The Efficacy of Written Corrective Feedback in Improving L2 Written Accuracy: A Meta-Analysis. *Modern Language Journal*, 99(1), 1-18. <https://doi.org/10.1111/modl.12189>.
- Karim, K. & Nassaji, N. (2018) The revision and transfer effects of direct and indirect comprehensive corrective feedback on ESL students' writing, *Language Teaching Research*, 24(4), 519-539. <https://doi.org/10.1177%2F1362168818802469>.
- Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 143-164). Amsterdam: John Benjamins.
- Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second-language writing skills. *The Modern Language Journal*, 75(3), 305-313. <https://doi.org/10.2307/328724>.
- Kim, H.R., & Bowles, M. (2019). How deeply do second language learners process written corrective feedback? Insights gained from think-alouds. *TESOL Quarterly*, 53(4), 913-938. <https://doi.org/10.1002/tesq.522>.
- Kormos, J. (2013). New conceptualisations of language aptitude in second language attainment. In G. Granena & M. Long (Eds.), *Sensitive periods, language aptitude and ultimate attainment* (pp. 131-152). Amsterdam: John Benjamins.
- Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning*, 62(2), 439-472. <https://doi.org/10.1111/j.1467-9922.2012.00695.x>.
- Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164. <https://doi.org/10.1016/j.system.2004.01.001>.
- Krashen, S (1982). *Principles and Practice in Second Language Acquisition* Oxford: Pergamon Press.
- Krashen, S. (1985). *The Input Hypothesis: Issues and Implications*. Harlow: Longman.

- Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M. P. Garcia-Mayo (Ed.), *Investigating Tasks Informal Language Settings* (pp. 117-135). Clevedon, England: Multilingual Matters.
- Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48-60. <https://doi.org/10.1016/j.jslw.2007.08.003>.
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In *Eurosla Yearbook*. Vol. 5, S. Foster-Cohen, M.P. García-Mayo, and J. Cenoz (eds.), 195-222. Amsterdam: John Benjamins.
- Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in 2 writing. *International Review of Applied Linguistics in Language Teaching*, 45(3), 261-284. <http://dx.doi.org/10.1515/iral.2007.012>.
- Laerd Statistics. (2018). MANOVA in SPSS Statistics. <https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php>.
- Lalande, J.F. (1982). Reducing composition errors: An experiment. *Modern Language Journal*, 66(2), 140-149. <https://doi.org/10.1111/j.1540-4781.1982.tb06973.x>.
- Larsen-Freeman, D. (1978). An ESL Index of Development. *TESOL Quarterly*, 12(4), 439-448. <https://doi.org/10.2307/3586142>.
- Larsen-Freeman, D. (1997). Chaos / complexity science and second language acquisition. *Applied Linguistics*, 18(2), 141-165. <https://doi.org/10.1093/applin/18.2.141>.
- Larsen-Freeman, D. (2001). *Teaching Language: From Grammar to Gramming*. Thomson/Heinle.
- Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, 27(4), 590-619. <https://doi.org/10.1093/applin/aml029>.
- Larsen-Freeman, D. (2009). Adjusting Expectations: The Study of Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4), 579-589. <https://doi.org/10.1093/applin/amp043>.
- Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics. *Language Teaching*, 45(2), 202-214. doi:10.1017/S0261444811000061.
- Larsen-Freeman, D., & Long, M. H. (1991). *An Introduction to Second Language Acquisition Research*. Harlow: Longman Group.
- Latchanna, G., & Dagnew, A. (2009). Attitude of teachers towards the use of active learning methods. *E-journal of All India Association for Educational Research*, 21(1). <http://www.ejournal.aiaer.net/vol21109/12.%20Latchana%20&%20Dagnew.pdf>.

- Lee, I., Yu, S., & Liu, Y. (2018). Hong Kong Secondary Students' Motivation in EFL Writing: A Survey Study. *TESOL Quarterly*, 52(1), 176-187. <https://doi.org/10.1002/tesq.364>.
- Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching*, 52(4), 524–536. doi:10.1017/S0261444819000247.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40 (3), 387–417. <https://doi.org/10.1111/j.1467-1770.1990.tb00669.x>.
- Li, S. (2015). The associations between language aptitude and second language grammar acquisition: A meta-analytic review of five decades of research. *Applied Linguistics*, 36(3), 385-408. <https://doi.org/10.1093/applin/amu054>.
- Li, S. (2016). The Construct Validity of Language Aptitude. *Studies in Second Language Acquisition*, 38(4), 801-842. doi:10.1017/S027226311500042X.
- Linck, J.A., Hughes, M.M., Campbell, S.G., Silbert, N.H., Tare, M., Jackson, S.R., Smith, B.K., Bunting, M.F., & Doughty, C.J. (2013). Hi-LAB: A New Measure of Aptitude for High-Level Language Proficiency. *Language Learning*, 63: 530-566. <https://doi.org/10.1111/lang.12011>.
- Loewen, S., Li, S., Fei, F., Thompson, A., Nakatsukasa, K., Ahn, S., & Chen, X. (2009). Second Language Learners' Beliefs About Grammar Instruction and Error Correction. *The Modern Language Journal*. 93, 91-104. <https://doi.org/10.1111/j.1540-4781.2009.00830.x>.
- Long, M. H. (2007). *Problems in Sla*. Mahwah, New Jersey: Lawrence Erlbaum Associates
- Lorenzo, F., & Rodríguez, L. (2014). Onset and expansion of L2 cognitive academic language proficiency in bilingual settings: CALP in CLIL. *System*. 47, 64-72. <https://doi.org/10.1016/j.system.2014.09.016>.
- Lu, X. (2010). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1), 3–28. <https://doi.org/10.1075/ijcl.14.1.02lu>.
- Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96(2), 190–208. <http://dx.doi.org/10.2307/41684069>.
- Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29(1), 16–27. <https://doi.org/10.1016/j.jslw.2015.06.003>.
- Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: quantification and assessment*. New York: Palgrave Macmillan.

- Mantle-Bromley, C. (1995). Positive attitudes and realistic beliefs: Links to proficiency. *Modern Language Journal*, 79(3), 372-386. <https://psycnet.apa.org/doi/10.2307/329352>.
- McCarthy, P. M., & Jarvis, S. (2010). MTL D, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>.
- McCarthy, P.M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International*, 66, 12.
- McDonald, J.H. (2014). *Handbook of Biological Statistics (3rd ed.)*. Sparky House Publishing, Baltimore, Maryland.
- McKee, G., Malvern D. D., & Richards, B. J. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15(3), 323-338. <https://doi.org/10.1093/lc/15.3.323>.
- McLaughlin, B. (1990). Restructuring. *Applied Linguistics*, 11, 113-128.
- McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.
- McNamara, D., Louwse, M., McCarthy, P., & Graesser, A. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion, *Discourse Processes*, 47(4), 292-330. DOI: 10.1080/01638530902959943.
- McNamara, P., McLaren, D., Smith, D., Brown, A., & Stickgold, R. (2005). A “Jekyll and Hyde” within: Aggressive versus friendly interactions in REM and non-REM dreams. *Psychological Science*, 16(2), 130-136. <https://doi.org/10.1111/j.0956-7976.2005.00793.x>.
- Meara, P. (2005). LLAMA Language Aptitude Tests. Swansea: Lognostics.
- Meara P, and Rogers, V. (2020). The LLAMA Tests v3. Cardiff: Lognostics.
- Meihami, H. (2013). Truscott’s claims in giving corrective feedback: Does it matter in EFL writing context. *International letters of social and humanistic sciences*, 8, 8-23. <https://doi.org/10.18052/www.scipress.com/ILSHS.8.8>.
- Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, 45(3), 241-259. <http://dx.doi.org/10.1515/iral.2007.011>.
- Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Florence: Taylor & Francis.

- Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339 – 364).
- Mizera, G. J. (2006). Working memory and L2 fluency. *Unpublished Doctoral Dissertation*. University of Pittsburgh.
- Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning level foreign language classrooms: A study of guided planning and relativisation. *Language Teaching Research*, 12(1), 11–37. <https://doi.org/10.1177%2F1362168807084492>.
- Mori, Y. (1999). Epistemological beliefs and language learning beliefs: What do language learners believe about their learning. *Language Learning* 49, 377–415.
- Muñoz, C. (2006). *Age and the Rate of Foreign Language Learning*. Multilingual Matters.
- Nicolás–Conesa, F., Manchón, R. M., & Cerezo, L. (2019). The effect of unfocused direct and indirect written corrective feedback on rewritten texts and new texts: Looking into feedback for accuracy and feedback for acquisition. *The Modern Language Journal*, 103(4), 848-873. <https://doi.org/10.1111/modl.12592>.
- Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15(5), 625-632. <http://dx.doi.org/10.1007/s10459-010-9222-y>.
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555-578. <https://doi.org/10.1093/applin/amp044>.
- Oladejo, J. A. (1993). Error correction in ESL: Learners’ preference. *TESL Canada Journal*, 10(2), 71-89. <https://doi.org/10.18806/tesl.v10i2.619>.
- Olson, C. L. (1974). Comparative Robustness of Six Tests in Multivariate Analysis of Variance. *Journal of the American Statistical Association*, 69 (348), 894-908. <https://doi.org/10.1080/01621459.1974.10480224>.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college- level L2 writing. *Applied Linguistics*, 24(4), 492–518. <https://doi.org/10.1093/applin/24.4.492>.
- Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, 25, 26-45. [doi:10.1017/S0267190505000024](https://doi.org/10.1017/S0267190505000024).
- Oxford, R. (2003). Language learning styles and strategies: Concepts and relationships. *International Review of Applied Linguistics in Language Teaching*, 41(4), 271-278. <https://doi.org/10.1515/iral.2003.012>.

- Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590-601. <https://doi.org/10.1093/applin/amp045>.
- Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research* 31 (1), pp. 117-134.
- Pallotti, G. (2020). *Measuring Complexity, Accuracy and Fluency (CAF)*. London: Routledge.
- Park, G.-P. (1995). *Language learning strategies and beliefs about language learning of university students learning English in Korea*. Austin, TX: University of Texas, Department of Curriculum and Instruction. [Ph.D. Dissertation].
- Parry, T., & Child, J. (1990). Preliminary investigation of the relationship between VORD, MLAT and Language Proficiency. In P. Thomas & C. Stansfield (Eds.), *Language aptitude reconsidered* (pp. 30 – 66). Englewood Cliffs, NJ: Prentice Hall Regents.
- Peregoy, S. F., & Boyle, O. F. (2005). *Reading, writing, and learning in ESL: A resource book for K-12 teachers*. New York, NY: Pearson Education.
- Pienemann, M. (1989). Is language teachable. *Applied Linguistics* 10(1), 52-79. <https://doi.org/10.1093/applin/10.1.52>.
- Pienemann, M. (1998). Language processing and second language development—Processability theory. *Studies in Bilingualism*, 15. Amsterdam: John Benjamins.
- Pimsleur, P. (1966). *Pimsleur Language Aptitude Battery (PLAB)*. San Diego: Harcourt Brace Jovanovich.
- Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47(1), 101–143. <https://doi.org/10.1111/0023-8333.31997003>.
- Polio, C. (2001). Research methodology in second language writing: The case of text-based studies. In T. Silva, & P. Matsuda (Eds.). *On second language writing* (pp. 91–116). Mahwah, NJ: Lawrence Erlbaum Associates.
- Polio, C. (2012). The relevance of second language acquisition theory to the written error correction debate. *Journal of Second Language Writing*, 21(4), 375-389. <https://doi.org/10.1016/j.jslw.2012.09.004>.
- Polio, C., & Yoon, H. (2018). The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*. 29(2), 171-172. <https://doi.org/10.1111/ijal.12200>.
- Polio, C., Fleck, C., & Leder, N. (1998). If I only had more time: ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, 7(1), 43-68. [https://doi.org/10.1016/S1060-3743\(98\)90005-4](https://doi.org/10.1016/S1060-3743(98)90005-4).

- Polyà, G. (1920). Ueber den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das /Momentproblem. *MathematischeZeitschrift* 8, 171–181.
<https://doi.org/10.1007/BF01206525>.
- Pourdana, N., Payam, N., Yousefi, F. (2021) Investigating metalinguistic written corrective feedback focused on EFL learners' discourse markers accuracy in mobile-mediated context. *Asian-Pacific Journal of Second and Foreign Language Eduction* 2021;6(1):7. doi: 10.1186/s40862-021-00111-8. Epub 2021 Jan 25. PMID: PMC7829097.
- Puchta, J. (1999). *Beyond materials, techniques, and linguistic analysis: The role of motivation, beliefs, and identity*. Plenary paper, 33rd International IATEFL Annual Conference, Edinburgh.
- Rahimpour, M. (1999). Task complexity and variation in interlanguage. In N. O. Jungheim & P. Robinson (Eds.), *Pragmatic and pedagogy: proceeding of the 3rd pacific Second Language Research Forum* (pp.115-134). Tokyo, Japan: Pac LRF.
- Rahimpour, M. (2008). Implementation of task-based approaches to language teaching. *Pazhuhesh-e-Zabanha-ye Khareji Journal*, 41, 45-61.
- Robb, T., Ross, S., & Shortreed, I. (1986). Saliency of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, 20(1), 83-95. <https://doi.org/10.2307/3586390>.
- Robinson, B. F., & Mervis, C. B. (1998). Disentangling early language development: Modeling lexical and grammatical acquisition using and extension of case-study methodology. *Developmental Psychology*, 34(2), 363–375. <https://doi.org/10.1037/0012-1649.34.2.363>.
- Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning*, 47(1), 45–99.
- Robinson, P. (2001). Individual differences, cognitive abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research*, 17(4), 368 – 392. <https://doi.org/10.1177/026765830101700405>.
- Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge, UK: Cambridge University.
- Robinson, P. (2002). Effects of individual differences in intelligence, aptitude and working memory on adult incidental SLA. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 212 – 266). Amsterdam and Philadelphia: John Benjamins Publishing Company.
- Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 631-678). Malden, MA: Blackwell Publishing.

- Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfield, and Hernstad (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, 27(2), 235 – 268. <http://dx.doi.org/10.1017/S0272263105050126>.
- Robinson, P. (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (Task-based language teaching, v. 2; Task-based language teaching, v. 2). Amsterdam: John Benjamins Pub.
- Robinson, P. and Ellis, N. (2008) a. *Conclusion: Cognitive Linguistics, Second Language Acquisition and L2 Instruction—Issues for Research*. In *Handbook of cognitive linguistics and second language acquisition*, edited by P. Robinson and N. C. Ellis. London: Routledge.
- Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL - International Review of Applied Linguistics in Language Teaching*, 45(3), 161-176. <http://dx.doi.org/10.1515/IRAL.2007.007>.
- Roehr-Brackin (2020) in Winke, P., & Brunfaut, T. eds. (2020). *The Routledge Handbook of Second Language Acquisition and Language Testing (1st ed.)*. Routledge. <https://doi-org.serlib0.essex.ac.uk/10.4324/9781351034784>.
- Roehr-Brackin, K. (2020). Measuring Aptitude. *The Routledge Handbook of Second Language Acquisition and Language Testing*.
- Rogers, V., Meara, P., Aspinall, R., Fallon, L., Goss, T., Keey, E., & Thomas, R. (2016). Testing aptitude: Investigating Meara's (2005) LLAMA tests. *EuroSLA Yearbook*, 16, 179-210.
- Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association*, 1(1), 49–60. DOI: <http://doi.org/10.22599/jesla.24>.
- Rosenthal, B. D. (2007). Improving elementary-age children's writing fluency: A comparison of improvement based on performance feedback frequency. *DAI-B*, 67(11).
- Rosmawati. (2013). A case study in complexity and accuracy development in ESL academic writing: A dynamic perspective. In *The Asian Conference on Education Conference 2013 Official Conference Proceedings* (pp. 54–71). Osaka, Japan, 23-27 October 2013.
- Ruegg, R. (2010). Interlanguage development: The effect of unfocused feedback on L2 writing. *Intercultural Communication Studies*, 19(1), 247-254.
- Sachs, R., & Polio, C. (2007). Learners' uses of two types of written feedback on an L2 writing revision task. *Studies in Second Language Acquisition*, 29(1), 67–100. doi:10.1017/S0272263107070039.

- Sadeghi, K., & Dilmaghani, K.S. (2013). The relationship between lexical diversity and genre in Iranian EFL learners' writings. *Journal of Language Teaching and Research*, 4(2), 328-334. DOI: 10.4304/jltr.4.2.328-334.
- Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, 46, 113 – 136. <https://doi.org/10.1515/IRAL.2008.005>.
- Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111-142). PA: John Benjamins.
- Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York, NY: Peter Lang.
- Sawyer, M., & Ranta, L. (2001). Aptitude, individual differences, and instructional design. In P. Robinson (Ed.), *Cognition and second language acquisition* (pp. 319- 353). New York: Cambridge University Press.
- Schmidt, R. (1995). Attention and awareness in foreign.
- Schmid, M., & Jarvis, S. (2014). Lexical access and lexical diversity in first language attrition. *Bilingualism: Language and Cognition*, 17(4), 729-748. doi:10.1017/S1366728913000771.
- Schmidt, R. (1990). The Role of Consciousness in Second Language Learning. *Applied Linguistics*, 11(2), 129-158. <https://doi.org/10.1093/applin/11.2.129>.
- Schmitt, N., Dörnyei, Z., Adolphs, S. , & Durow, V. (2004). Knowledge and acquisition of formulaic Sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, Processing, and Use* (pp. 55-70). Philadelphia, USA: John Benjamins Publishing Company.
- Schommer, M. (1990). Effects of beliefs about the nature of knowledge and learning among post-secondary students. *Journal of Educational Psychology*, 82(3), 498–504. <https://doi.org/10.1037/0022-0663.82.3.498>.
- Schommer, M. (1994). Synthesizing epistemological belief research: Tentative understandings and provocative confusions. *Educational Psychology Review*, 6, 293–319. <https://doi.org/10.1007/BF02213418>.
- Schulz, R. A. (1996). Focus on form in the foreign language classroom: Students' and teachers' views on error correction and the role of grammar. *Foreign Language Annals*, 29(3), 343-364. <https://doi.org/10.1111/j.1944-9720.1996.tb01247.x>.
- Schulz, R.A. (2001). Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA-Columbia. *The Modern Language Journal*, 85(2), 244–258. <https://doi.org/10.1111/0026-7902.00107>.
- Semke. H. D. (1984). Effects of the red pen. *Foreign Language Annals*, 17(3), 195-202. <https://doi.org/10.1111/j.1944-9720.1984.tb01727.x>.

- Sercu, L., De Wachter, L., Peters, E., Kuiken, F., & Vedder, I. (2006). The effect of task complexity and task conditions on foreign language development and performance. Three empirical studies. *International Journal of Applied Linguistics*, 152(1), 55–84. <https://doi.org/10.2143/ITL.152.0.2017863>.
- Shams, M. (2008). Students' attitudes, motivation and anxiety towards English language learning. *Journal of Research*, 2, 121–144.
- Shaofeng, L. (2015). The Associations Between Language Aptitude and Second Language Grammar Acquisition: A Meta-Analytic Review of Five Decades of Research. *Applied Linguistics*, 36 (3), 385–408. <https://doi.org/10.1093/applin/amu054>.
- Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), *Conversational interaction in second language acquisition* (pp. 301 – 322). New York, NY: Oxford University Press.
- Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System*, 37(4), 556-569. doi:10.1016/j.system.2009.09.002.
- Sheppard, K. (1992). Two feedback types: Do they make a difference. *RELC Journal*, 23(1), 103-110. <https://doi.org/10.1177/003368829202300107>.
- Šišková, Z. (2012). Lexical Richness in EFL Students' Narratives. *Language Studies Working Papers*, 4, 26-36.
- Skehan, P. (1989). *Individual differences in second language learning*. London: Edward Arnold.
- Skehan, P. (1996). Second language acquisition and task- based instruction. In J. Willis, & D. Willis (Eds.). *Challenge and change in language teaching* (pp. 17–30). Oxford: Heinemann.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (ed.). *Individual differences and instructed language learning*. Amsterdam/ Philadelphia: John Benjamins.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, lexis. *Applied Linguistics*, 30(4), 510–532. <https://doi.org/10.1093/applin/amp047>.

- Skehan, P. (2009a). 'Lexical performance by native and non-native speakers on language learning tasks' in B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds): *Vocabulary Studies in First and Second Language Acquisition: The Interface between Theory and Application*. Palgrave Macmillan.
- Skehan, P. (2009b). 'Models of speaking and the assessment of second language proficiency' in A. Benati (ed.): *Issues in Second Language Proficiency*. Continuum.
- Skehan, P. (2009b). Lexical performance by native and non-native speakers on language-learning tasks. In Richards, B., Daller, H., Malvern, D.D., Meara, P. (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (107–24). London: Palgrave Macmillan.
- Skehan, P. and P. Foster. (2007). 'Complexity, accuracy, fluency and lexis in task-based performance: A meta-analysis of the Ealing Research' in S. Van Daele et al. (eds): *Complexity, Accuracy and Fluency in Second Language Use, Learning and Teaching*.
- Skehan, P., & Foster, P. (1996). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, 3(3), 215-247. <http://dx.doi.org/10.1191/136216899672186140>.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49(1), 93-120. <https://doi.org/10.1111/1467-9922.00071>.
- Smitt, N., Zoltan, D., Adolph, S., & Durow, V. (2004). Knowledge and acquisition of formulaic Sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, Processing, and Use* (pp. 55-70). Philadelphia, USA: John Benjamins Publishing Company.
- Spada, N., Tomita, Y. Daele Svan. (2007). *The complexities of selecting complex (& simple) forms in instructed SLA research*. Complexity, Accuracy and Fluency in second Language Use, Learning and Teaching.
- Sparks, R. L., Ganschow, L., & Patton, J. (1995). Prediction of performance in first year foreign language courses: Connections between native and foreign language learning. *Journal of Educational Psychology*, 87(4), 638–655. <https://psycnet.apa.org/doi/10.1037/0022-0663.87.4.638>.
- Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, 31(4), 532-553. <https://psycnet.apa.org/doi/10.1093/applin/amq001>.
- Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.
- Stansfield, C. W. (1989). *Language aptitude reconsidered*. ERIC Digest.
- Stern, H. (1976). Optimal age: myth or reality. *Canadian Modern Language Review* 32(3), 283–294. <https://doi.org/10.3138/cmlr.32.3.283>.

- Stern, H.H. (1983). *Fundamental Concepts of Language Teaching*. Oxford: Oxford University Press.
- Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*(2), 103-118. <https://doi.org/10.1016/j.jslw.2009.02.003>.
- Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake and retention of corrective feedback on Writing. *Studies in Second Language Acquisition, 32*(2), 303-334. doi:10.1017/S0272263109990532.
- Suzuki, W., Nassaji, H., & Sato, K. (2019). The effects of feedback explicitness and type of target structure on accuracy in revision and new pieces of writing. *System, 81*, 135-145. <https://doi.org/10.1016/j.system.2018.12.017>.
- Swain, M. (2006). Languaging, agency and collaboration in advanced second language learning. In H. Byrnes (Ed.), *Advanced language learning: The contributions of Halliday and Vygotsky* (pp. 95–108). London, England: Continuum.
- Tai, H.-Y. (2015). Writing development in syntactic complexity, accuracy and fluency in a content and language integrated learning class. *International Journal of Language and Linguistics* 2(3), 149–156.
- Tanaka, K., & R. Ellis. (2003). Study abroad, language proficiency, and learner beliefs about language learning. *JALT Journal, (25.1)*, 81–102. <https://doi.org/10.37546/JALTJJ25.1-3>.
- Tarone, E., Downing, B., Cohen, A., Gillette, S., Murie, R., & Dailey, B. (1993). The writing of Southeast Asian-American students in secondary school and university. *Journal of Second Language Writing, 2*(2), 149-172. [https://doi.org/10.1016/1060-3743\(93\)90015-U](https://doi.org/10.1016/1060-3743(93)90015-U).
- Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239–277). Amsterdam: John Benjamins.
- Thomas, M. (1994). Assessment of L2 proficiency in second language research. *Language Learning, 44*(2), 307-337. <http://dx.doi.org/10.1111/j.1467-1770.1994.tb01104.x>.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84–119. <https://doi.org/10.1093/applin/17.1.84>.
- Treffers-Daller J. (2013) Measuring lexical diversity among L2 learners of French: An exploration of the validity of D, MTLD and HD-D as measures of language ability. In: Jarvis S, Daller M, editors. *Vocabulary knowledge: Human ratings and automated measures*. 28. Amsterdam: Benjamins, 79–104.

- Tremblay, A. (2011). Proficiency assessment standard in second language acquisition research “Clozing” the gap. *Studies in Second Language Acquisition*, 33(3), 339-372. doi:10.1017/S0272263111000015.
- Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). *How effective are recasts? The role of attention, memory and analytical ability*. In ed. A. Mackey, 171–195.
- Truscot, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning*, 46(2), 327-369. <https://doi.org/10.1111/j.1467-1770.1996.tb01238.x>.
- Truscott, J. (2007). The effect of error correction on learners’ ability to write accurately. *Journal of Second Language Writing*, 16(4), 255-272. <https://doi.org/10.1016/j.jslw.2007.06.003>.
- Truscott, J., & Hsu, A. Y. P. (2008). Error correction, revision, and learning. *Journal of Second Language Writing*, 17(4), 292-305. <https://doi.org/10.1016/j.jslw.2008.05.003>.
- Tumposky, N.R. (1991). Student beliefs about language learning: A cross-cultural study. *Carleton Papers in Applied Language Studies*, 8, 50-65.
- Turner, M.L., & Engle, R. W. (1989). Is working memory capacity task dependent. *Journal of Memory & Language*, 28(2), 127–154. [https://doi.org/10.1016/0749-596X\(89\)90040-5](https://doi.org/10.1016/0749-596X(89)90040-5).
- UCLES. (2004). *Quick Placement Test*. Oxford University Press.
- Van Beuningen, C. (2010). Corrective Feedback in L2 Writing: Theoretical Perspectives, Empirical Insights, and Future Directions. *International Journal of English Studies*, 10(2), 1-27. <http://dx.doi.org/10.6018/ijes.10.2.119171>.
- Van Beuningen, C., de Jong, N. H., & Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL International Journal of Applied Linguistics*, 156, 279-296. <http://dx.doi.org/10.2143/ITL.156.0.2034439>.
- Van Beuningen, C., de Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in Dutch multilingual classrooms. *Language Learning*, 62(1), 1-41. <https://doi.org/10.1111/j.1467-9922.2011.00674.x>.
- Van Geert, P. and H. Steenbeek. (2005a.) ‘A complexity and dynamic systems approach to development assessment, modeling and research’ in K. W. Fischer, A. Battro, and P. Lena (eds): *Mind, Brain, and Education*. Cambridge: Cambridge University Press.
- Van Geert, P., & Steenbeek, H. (2005). Explaining After by Before: Basic Aspects of a Dynamic Systems Approach to the Study of Development. *Developmental Review*, 25(3-4), 408-442. <https://doi.org/10.1016/j.dr.2005.10.003>.
- Vercellotti, M. L. (2012). *Complexity, accuracy, and fluency as properties of language performance: The development of the multiple subsystems over time and in relation to each other*. Ph. D. dissertation. University of Pittsburgh.

- Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *Modern Language Journal*, 92(2), 214-231. <https://doi.org/10.1111/j.1540-4781.2008.00715.x>.
- Victori, M., & Lockhart, W. (1995). Enhancing metacognition in self-directed language learning. *System*, 23(2), 223-234. [https://doi.org/10.1016/0346-251X\(95\)00010-H](https://doi.org/10.1016/0346-251X(95)00010-H).
- Wang, S., & Slater, T. (2016). Syntactic Complexity of EFL Chinese Students' Writing. *English Language and Literature Studies*, 6(1), 81-86. <https://doi.org/10.5539/ells.v6n1p81>.
- Wechsler, D. (1972). *Wechsler Memory Scale, Form I*. New York: The Psychological Corporation.
- Weinburgh, M. H. (1998). *Gender, ethnicity, and grade level as predictors of middle school students' attitudes towards science*. http://www.Ed.Psu.Edu/Ci/Journals/1998aets/S5_1_Weinburgh.Rtf.
- Wen, Z. (2007). Working memory as foreign language aptitude: theory and practice. *Xiandai Waiyu Modern Foreign Languages*, 1, 85-97.
- Wen, Z., Biedroń, A., & Skehan, P. (2017). Foreign language aptitude theory: Yesterday, today and tomorrow. *Language Teaching*, 50(1), 1-31. doi:10.1017/S0261444816000276.
- Wenden, A. (1986a). Helping language learners think about learning. *English Language Journal*, 40(1), 3-12.
- Wenden, A. (1987). How to be a successful language learner: Insights and prescriptions from L2 learners. In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 103-117). London: Prentice Hall.
- Wenden, A. (1999). An introduction to metacognitive knowledge and beliefs in language learning: Beyond the basics [Special Issue]. *System*, 27(4), 435-441.
- Wenden, A. (2001). Metacognitive knowledge. In Breen, M.P. (Ed.), *Learner contributions to language learning. New Directions in Research* (pp. 44-64). Harlow, Essex: Pearson Education Limited.
- White, C. (1999). Expectations and emergent beliefs of self-instructed language learners. *System*, 27(4), 443-457.
- Williams, J. N. (1999). Learner-generated attention to form. *Language Learning*, 49(4), 583-625.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawai'i, Second Language Teaching and Curriculum Center.

- WorldCat (2001). *Quick placement test: Paper and pen test: user manual*. Oxford: Oxford University Press.
- Xu, C. (2009). Overgeneralisation from a narrow focus: A response to Ellis et al. (2008) and Bitchener (2008). *Journal of Second Language Writing*, 18(4), 270-275. doi:10.1016/j.jslw.2009.05.005.
- Yang, N. D. (1999). The relationship between EFL learners' beliefs and learning strategy use. *System*, 27(4), 515-535. [https://doi.org/10.1016/S0346-251X\(99\)00048-2](https://doi.org/10.1016/S0346-251X(99)00048-2).
- Yang, N.-D. (1992). *Second language learners' beliefs about language learning and their use of learning strategies: A study of college students of English in Taiwan*. Austin, TX: University of Texas, Department of Curriculum and Instruction. [Ph.D. Dissertation].
- Yilmaz, Y. (2013). The role of working memory capacity and language analytic ability in the effectiveness of explicit correction and recasts. *Applied Linguistics*, 34(2), 344-368. <http://dx.doi.org/10.1093/applin/ams044>.
- Yu, B. (2008). *Cross-cultural adaptation and second language acquisition: a study of international students in universities of the People's Republic of China*. Unpublished doctoral dissertation, University of Hong Kong, Hong Kong, China. Retrieved on March 15, 2017 from <http://hub.hku.hk/handle/123456789/51421>.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral production. *Applied Linguistics*, 24(1), 1-27. <https://doi.org/10.1093/applin/24.1.1>.
- Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing*, 37, 13–24. <https://doi.org/10.1016/j.asw.2018.03.001>.

Appendix A: Metalinguistic Error Codes Used

Metalinguistic Error Codes Used

Wt = wrong tense

Ww = wrong word

Gr = grammar error

Sp = spelling

Art = article

Prep = preposition

S/V = Subject/Verb agreement

^ = word missing

/ = start a new sentence

// = start a new paragraph

??? = I do not understand

P = punctuation error

Wo = wrong word order

Appendix B : Essay Topic Questionnaire

1. Homework should be banned.
2. School uniforms should be required.
3. Year round education is not a good idea for student learning.
4. Schools should block Youtube.
5. All parents should be required to attend parenting classes before having a child.
6. High stakes testing, such as final exams, should be abolished.
7. The driving age should be raised.
8. Animal testing should be banned.
9. Technology makes us more alone.
10. Cheating is getting worse.
11. Students should be able to grade their teachers.
12. Teachers assign too much homework.
13. Technology gets in the way of learning.
14. 12 is too young for an iPhone.
15. Schools should offer cash bonuses for good test scores.
16. Enough is not done to stop cyberbullying.
17. Talent is more important than hard work.
18. Facebook is no longer as popular as it used to be.
19. Tablet computers should become the primary way students learn in class.
20. A college education is necessary for financial success.

Likert scale: I am interested in this topic / neutral/ I am not interested in this topic.

Appendix C: Language Background Questionnaire

Name: _____

ID:

1. Age: _____

2. Gender: _____

3. Native Language(s):

4. Have you spent any time in English speaking countries (apart from the UAE)? If yes, please state which country/countries and how many months/years you have spent in each:

5. What language(s) do you speak at home? :

6. What language(s) do you use with friends? :

7. What was the language of instruction in your previous school(s) (please mention the years) e.g. English 12 years, or Arabic 5 years, English 6 years etc..)? :

8. Do you know any other languages than English and your native language and what is your level

If yes go to next question

Language:

Level:

Language:

Level:

Language:

Level:

(i.e. advanced, intermediate, beginner):

9. How many years have you been learning English? : _____

10. TOEFL score : _____

11. IELTS score: _____

Appendix D: Attitudes Questionnaire

1. I consider error correction useful:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

2. I find indirect feedback (the teacher underlines my mistakes) useful:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

3. When my teacher corrects my errors, I feel less motivated to continue learning:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

4. I find direct feedback (the teacher writes the correct form where the error was)

useful : Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree

5. I am more likely to repeat errors if they are not corrected :

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

6. In my writing, I like the teacher to correct all my errors:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

7. I find metalinguistic feedback (the teacher tells me what kind of error I made i.e. tense error, pronoun error) useful:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

8. Corrective feedback does not help me:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
9. Even if an error does not impede the meaning it should be corrected:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
10. I find reviewing the corrective feedback with the teacher important:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
11. I feel I learn a lot from corrective feedback:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
12. I always look at the corrective feedback given by the teacher:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
13. The feedback I receive on my essays helps me improve my English writing:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
14. The feedback I receive on my essays increases my motivation to revise my essays:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
15. The feedback I receive on my essays helps me improve my overall English:
Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /
16. I feel anxious about receiving corrective feedback:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

17. Reviewing feedback by myself helps me improve my English grammar:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

18. I feel a lack of self-confidence when I receive corrective feedback:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

19. Corrective feedback is an important part of the academic writing class:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

20. Corrective feedback given by peers useful:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

21. Corrective feedback improves my understanding of English grammar:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

22. Reviewing feedback by myself helps me improve my English writing:

Strongly agree / agree / neither agree nor disagree / disagree / strongly disagree /

23. What do you like about corrective feedback?

24. What do you dislike about corrective feedback?

25. How do you feel when you receive corrective feedback?

Appendix E: Correlations of CAF Measures and PCA

Table 32. Correlations of CAF Measures at pre-test

Spearman Correlations	(Proficiency) Oxford Quick Placement Test	Fluency	Accuracy	Syntactic complexity Overall (mean length of t-units)	Syntactic complexity Sentenial (clauses per t-unit)	Syntactic complexity Sentenial (dependant clause ratio)	Syntactic complexity Subsentenial (mean length of clauses)	Lexical Diversity (TTR)	Lexical Diversity (mean segmental TTR)
<hr/>									
Proficiency (Oxford Quick Placement Test)									
<hr/>									
Fluency	Rho=.347 Sig. = .006 N= 58								
<hr/>									
Accuracy	Rho=.209 Sig. = .109 N= 58	Rho=.259 Sig. = .046 N= 58							
<hr/>									
Syntactic complexity	Rho=.162 Sig. = .226	Rho=.073 Sig. = .584	Rho=-.313 Sig. = .017						
<hr/>									

Spearman Correlations	(Proficiency) Oxford Quick Placement Test	Fluency	Accuracy	Syntactic complexity Overall (mean length of t-units)	Syntactic complexity Sentenial (clauses per t-unit)	Syntactic complexity Sentenial (dependant clause ratio)	Syntactic complexity Subsentenial (mean length of clauses)	Lexical Diversity (TTR)	Lexical Diversity (mean segmental TTR)
Overall (mean length of t-units)	N= 58	N= 58	N= 58						
Syntactic complexity	Rho=.084 Sig. = .553	Rho=-.071 Sig. = .559	Rho=-.204 Sig. = .124	Rho = .810 Sig. = .000					
Sentenial (clauses per t-unit)	N= 58	N= 58	N= 58	N= 58					
Syntactic complexity	Rho=.703 Sig. = .585	Rho=-.010 Sig. = .940	Rho=-.122 Sig. = .362	Rho = .709 Sig. = .000	Rho = .895 Sig. = .000				
Sentenial (dependant clause ratio)	N= 58	N= 58	N= 58	N= 58	N= 58				
Syntactic complexity	Rho=.152 Sig. = .255	Rho=.208 Sig. = .118	Rho=-.172 Sig. = .197	Rho = .269 Sig. = .041	Rho= -.264 Sig. = .046	Rho= -.280 Sig. = .033			
	N= 58	N= 58	N= 58	N= 58	N= 58	N= 58			

Spearman Correlations	(Proficiency) Oxford Quick Placement Test	Fluency	Accuracy	Syntactic complexity Overall (mean length of t-units)	Syntactic complexity Sentenial (clauses per t-unit)	Syntactic complexity Sentenial (dependant clause ratio)	Syntactic complexity Subsentenial (mean length of clauses)	Lexical Diversity (TTR)	Lexical Diversity (mean segmental TTR)
Subsentenial (mean length of clauses)									
Lexical Diversity (TTR)	Rho=-.260 Sig. = .049 N= 58	Rho=-.716 Sig. = .000 N= 58	Rho=-.128 Sig. = .337 N= 58	Rho= -.059 Sig. = .658 N= 58	Rho= -.058 Sig. = .664 N= 58	Rho= -.189 Sig. = .156 N= 58	Rho = -.089 Sig. = .033 N= 58		
Lexical Diversity (mean segmental TTR)	Rho=.038 Sig. = .080 N= 58	Rho=.116 Sig. = .385 N= 58	Rho=.144 Sig. = .281 N= 58	Rho = .093 Sig. = .488 N= 58	Rho= -.155 Sig. = .338 N= 58	Rho= -.213 Sig. = .108 N= 58	Rho = .249 Sig. = .059 N= 58	Rho=.615 Sig. = .033 N= 58	

Table 33 Principal Component Analysis 4 Factor Solution

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of		Total	% of		Total	% of	
		Variance	Cumulative %		Variance	Cumulative %		Variance	Cumulative %
1	2.795	34.935	34.935	2.795	34.935	34.935	2.507	31.341	31.341
2	1.997	24.957	59.892	1.997	24.957	59.892	1.822	22.772	54.112
3	1.388	17.356	77.248	1.388	17.356	77.248	1.500	18.745	72.858
4	.937	11.711	88.959	.937	11.711	88.959	1.288	16.101	88.959
5	.536	6.694	95.653						
6	.185	2.308	97.961						
7	.140	1.755	99.715						
8	.023	.285	100.000						

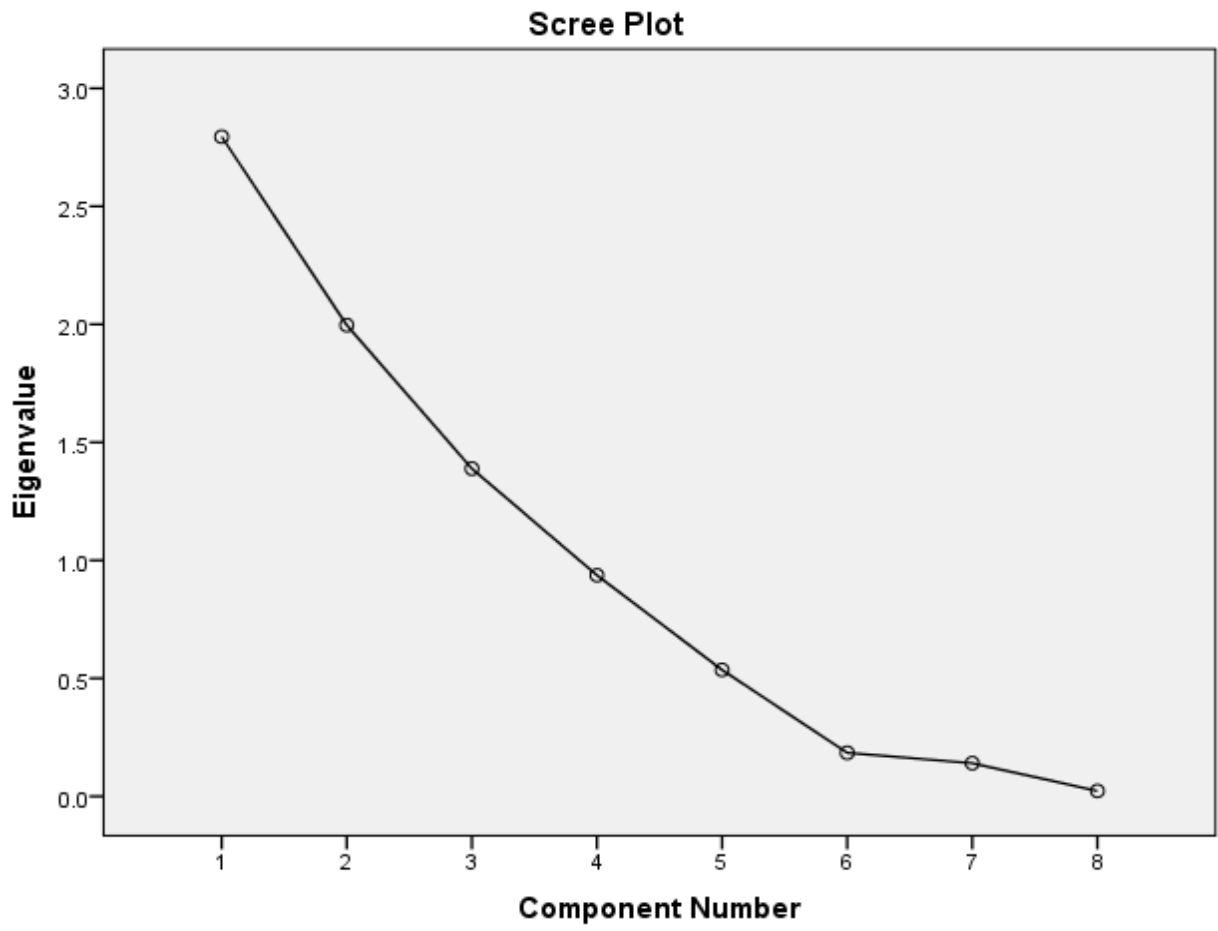


Figure 20. Scree Plot

Table 34. Rotated Component Matrix^a

	Component			
	1	2	3	4
Fluency pre test number of words written	-.076	-.439	.486	.614
Accuracy pre test measured as Error free t- unit per t-unit	-.080	.006	-.201	.900
Syntatic complexity overall pre test measured as mean length of t-unit	.811	-.011	.524	-.153
Syntatic complexity sentential pre test measured as clause per t- unit	.954	-.130	-.062	-.094
Syntatic complexity sentential pre test measured as dependant clause ratio	.947	-.084	.005	.025
Syntatic complexity subsential pre test measured as mean length per clause	.082	.130	.937	-.106

Lexical complexity diversity pre test measured as TTR	-.135	.905	-.108	-.219
Lexical complexity diversity pre test measured as mean segmental TTR	-.082	.876	.235	.092

Appendix F: Tests of Normality

Table 35. Tests of Normality

Tests of Normality						
Variable	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
LLAMAB aptitude test	.102	58	.200	.966	58	.106
LLAMAF aptitude test	.139	58	.007	.951	58	.019
Oxford Quick Placement Test	.142	58	.005*	.884	58	.000*
I consider error correction useful	.425	58	.000*	.628	58	.000*
I find indirect feedback useful	.311	58	.000*	.840	58	.000*
I find direct feedback useful	.354	58	.000*	.717	58	.000*
I find metalinguistic feedback useful	.261	58	.000*	.735	58	.000*
I am more likely to repeat errors if they are not corrected	.310	58	.000*	.779	58	.000*
Making errors in English is necessary for improving my English	.222	58	.000*	.845	58	.000*

I am worried about making errors in English when I write in English	.289	58	.000*	.865	58	.000*
When my teacher corrects my errors, I feel more motivated to continue learning	.208	58	.000*	.862	58	.000*
I would accept an English teacher who did not correct my writing	.234	58	.000*	.832	58	.000*
In my writing, I like the teacher to correct all my errors	.254	58	.000*	.808	58	.000*
In my writing, I like the teacher to correct some of my errors	.242	58	.000*	.882	58	.000*
Even if an error does not impede meaning it should be corrected	.288	58	.000*	.851	58	.000*
If an error impedes the meaning it should be corrected	.269	58	.000*	.798	58	.000*
Fluency pre-test number of words written	.116	58	.050	.967	58	.111

Fluency post-test number of words written	.097	58	.200	.970	58	.162
Gain in fluency	.078	58	.200	.981	58	.499
Accuracy pre-test measured as Error free t- unit per t-units	.115	58	.054	.927	58	.002*
Accuracy post-test measured as Error free t- unit per t-units	.123	58	.029	.930	58	.003*
Gain in accuracy	.079	58	.200	.984	58	.650
Syntactic complexity overall pre-test measured as mean length of t-units	.188	58	.000*	.655	58	.000*
Syntactic complexity overall post-test measured as mean length of t-units	.169	58	.000*	.902	58	.000*
Gain in syntactic complexity overall post- test measured as mean length of t-units	.173	58	.000*	.885	58	.000*
Syntactic complexity sentential pre-test measured as clauses per t-unit	.182	58	.000*	.885	58	.000*

Syntactic complexity sentential post-test measured as clauses per t-unit	.156	58	.001*	.856	58	.000*
Gain in Syntactic complexity sentential post-test measured as clauses per t-unit	.134	58	.011	.949	58	.017
Syntactic complexity subsential pre-test measured as mean length per clause	.123	58	.029	.832	58	.000*
Syntactic complexity subsential post-test measured as mean length per clause	.182	58	.000*	.784	58	.000*
Gain in syntactic complexity subsential post-test measured as mean length per clause	.140	58	.006	.848	58	.000*
Lexical diversity pre-test measured as mean segmental TTR	.348	58	.000*	.612	58	.000*

Lexical diversity post-test measured as mean segmental TTR	.381	58	.000*	.569	58	.000*
Gain in Lexical diversity post-test measured as mean segmental TTR	.223	58	.000*	.612	58	.000*

Appendix G: Kruskal-Wallis Test

Kruskal-Wallis test used to test for any initial between-group differences in language proficiency and CAF measures

Table 36. Kruskal-Wallis test

CAF Measure	P value
Fluency	.608
Accuracy	.211
Syntactic Complexity Overall	.239
Syntactic Complexity Sentential (measured as clauses per t-unit)	.721
Syntactic Complexity Subsentential	.671
Lexical Complexity Diversity (measured as mean segmental TTR)	.092

Appendix H: Friedman Test for Effect of Corrective Feedback on Complexity Measures

Table 37. Friedman test

Complexity measure	Friedman test statistics
Fluency	Direct N = 13, $\chi^2(1) = .692$, $p = .405$
	Indirect N = 17 $\chi^2(1) = .059$, $p = .808$
	Metalinguistic N = 16, $\chi^2(1) = 2.25$, $p = .134$
	Control N = 16, $\chi^2(1) = .088$, $p = .302$
Accuracy	Direct N = 12, $\chi^2(1) = .111$, $p = .739$
	Indirect N = 16 $\chi^2(1) = 1.923$, $p = .166$
	Metalinguistic N = 16, $\chi^2(1) = .067$, $p = .796$
	Control N = 16, $\chi^2(1) = .734$, $p = .683$
Syntactic complexity overall measured as mean length of t-unit	Direct N = 11, $\chi^2(1) = .091$, $p = .763$
	Indirect N = 15, $\chi^2(1) = 1.667$, $p = .197$
	Metalinguistic N = 16, $\chi^2(1) = 1.000$, $p = .317$
	Control N = 16, $\chi^2(1) = 1.158$, $p = .183$
Syntactic complexity sentential measured as clause per t-unit	Direct N = 11, $\chi^2(1) = .810$, $p = .366$
	Indirect N = 15, $\chi^2(1) = 1.657$, $p = .573$
	Metalinguistic N = 16, $\chi^2(1) = .000$, $p = 1.000$
	Control N = 16, $\chi^2(1) = 1.277$, $p = .324$

Complexity measure	Friedman test statistics
Syntactic complexity subsentential measured as mean length per clause	Direct N = 11, $\chi^2(1) = .818$, $p = .366$
	Indirect N = 15, $\chi^2(1) = 1.667$, $p = .275$
	Metalinguistic N = 16, $\chi^2(1) = 1.000$, $p = .317$
	Control N = 16, $\chi^2(1) = 1.342$, $p = .676$
Lexical diversity measured as mean segmental TTR	Direct N = 11, $\chi^2(1) = .091$, $p = .763$
	Indirect N=15, $\chi^2(1) = .067$, $p = .796$
	Metalinguistic N = 16, $\chi^2(1) = 2.571$, $p = .109$
	Control N = 16, $\chi^2(1) = 1.953$, $p = .472$

Appendix I: Kolmogorov-Smirnov Test

Table 37. Kolmogorov-Smirnov

	Statistic	Df	Sig.
Oxford Quick Placement Test	.086	134	.017
LLAMA F Aptitude test	.092	134	.007
LLAMA B Aptitude test	.090	134	.010
Total number of words in the text pre test	.089	134	.011
Total number of words in the text re test	.070	134	.200
Total number of words in the text post test	.083	134	.024
Proportion of error free t-units pre test	.069	134	.200
Proportion of error free t-units re test	.131	134	.000
Proportion of error free t-units post test	.425	134	.000
Errors per 100 words pre test	.082	134	.028
Errors per 100 words re test	.071	134	.098
Errors per 100 words post test	.058	134	.200*
Lexical errors per 100 words pre test	.132	134	.000

	Statistic	Df	Sig.
Lexical errors per 100 words retest	.132	134	.000
Lexical errors per 100 words post test	.123	134	.000
Grammatical errors per 100 words pre test	.181	134	.000
Grammatical errors per 100 words retest	.116	134	.000
Grammatical errors per 100 words post test	.176	134	.000
Spelling and punctuation errors per 100 words pre test	.096	134	.004
Spelling and punctuation errors per 100 words retest	.102	134	.002
Spelling and punctuation errors per 100 words post test	.085	134	.020
Mean length of T-units pre test	.216	134	.000
Mean length of T-units retest	.257	134	.000
Mean length of T-units post test	.274	134	.000
Complex nominals per clause pre test	.069	134	.200
Complex nominals per clause retest	.082	134	.029
Complex nominals per clause post test	.061	134	.200*

	Statistic	Df	Sig.
Clauses per t-unit pre test	.102	134	.002
Clauses per t-unit retest	.110	134	.000
Clauses per t-unit post test	.117	134	.000
voc-D pre test	.062	134	.200*
voc-D retest	.086	134	.017
voc-D post test	.039	134	.200*
MTLD pre test	.100	134	.002
MTLD re test	.073	134	.074
MTLD post test	.051	134	.200*

Appendix J: Tests of Normality

Table 39. Tests of Normality

Tests of Normality						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Zscore: Total number of words in the text pre test	.094	138	.005	.945	138	.000
CompositeZAccuracy	.055	138	.200*	.991	138	.566
ComplexityZComposite	.164	138	.000	.832	138	.000
Zscore: Complex nominals per clause pre test	.068	138	.200*	.974	138	.010
LexicaldiversityZcomposi te	.089	138	.009	.972	138	.006

Table 40. Tests of Normality

Tests of Normality						
	Treatment groups	Kolmogorov-Smirnov ^a			Shapiro-Wilk	
		Statistic	df	Sig.	Statistic	df
Zscore: Total number of words in the text pre test	Direct	.155	40	.016	.880	40
	Indirect	.079	35	.200*	.983	35
	Metalinguistic	.147	31	.087	.814	31
	Control	.073	32	.200*	.987	32
CompositeZAccuracy	Direct	.104	40	.200*	.981	40
	Indirect	.079	35	.200*	.982	35
	Metalinguistic	.109	31	.200*	.984	31
	Control	.119	32	.200*	.955	32
ComplexityZComposit e	Direct	.186	40	.001	.839	40
	Indirect	.172	35	.010	.815	35
	Metalinguistic	.170	31	.023	.921	31
	Control	.147	32	.078	.859	32
Zscore: Complex nominals per clause pre test	Direct	.149	40	.026	.968	40
	Indirect	.121	35	.200*	.952	35
	Metalinguistic	.099	31	.200*	.970	31
	Control	.176	32	.013	.913	32
LexicaldiversityZcomp osite	Direct	.123	40	.128	.943	40
	Indirect	.100	35	.200*	.970	35
	Metalinguistic	.093	31	.200*	.975	31
	Control	.131	32	.178	.974	32

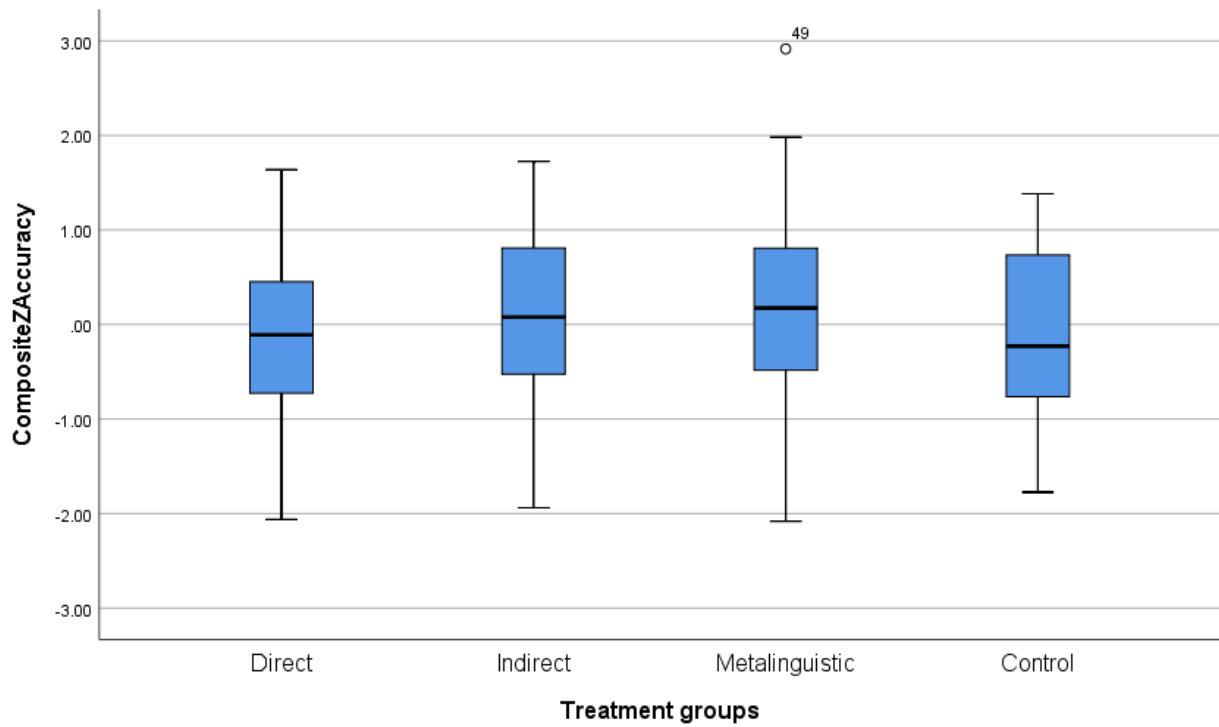
Table 41. Tests of Normality

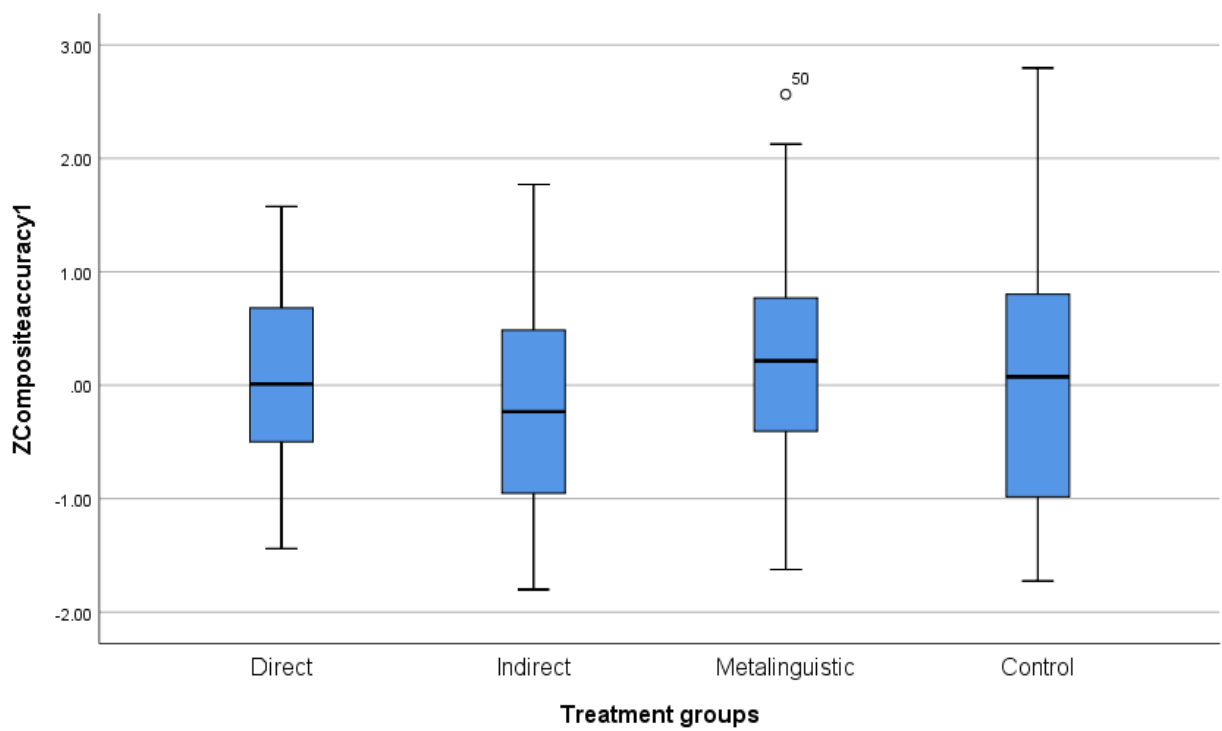
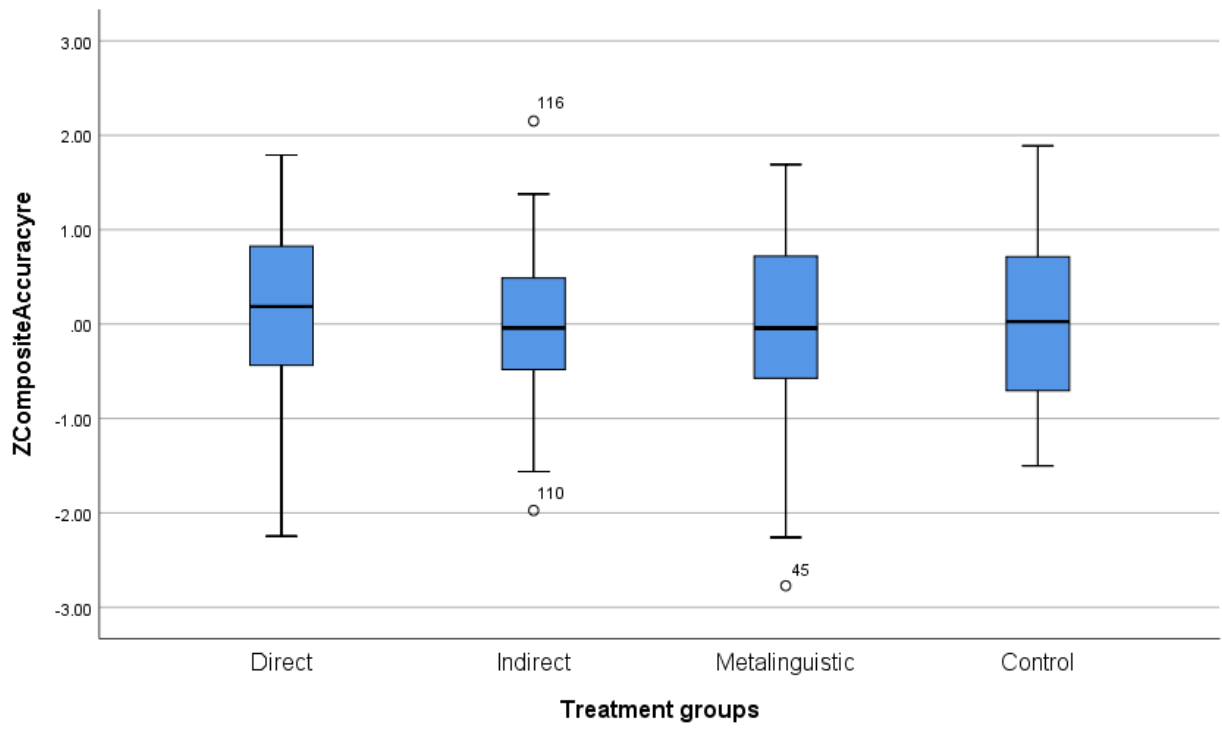
Tests of Normality		
	Treatment groups	Shapiro-Wilk ^a
		Sig.
Zscore: Total number of words in the text pre test	Direct	.001
	Indirect	.848
	Metalinguistic	.000
	Control	.955
CompositeZAccuracy	Direct	.735
	Indirect	.812
	Metalinguistic	.915
	Control	.200
ComplexityZComposite	Direct	.000
	Indirect	.000
	Metalinguistic	.026
	Control	.001
Zscore: Complex nominals per clause pre test	Direct	.307
	Indirect	.129
	Metalinguistic	.520
	Control	.014
LexicaldiversityZcomposite	Direct	.042
	Indirect	.456
	Metalinguistic	.670
	Control	.602

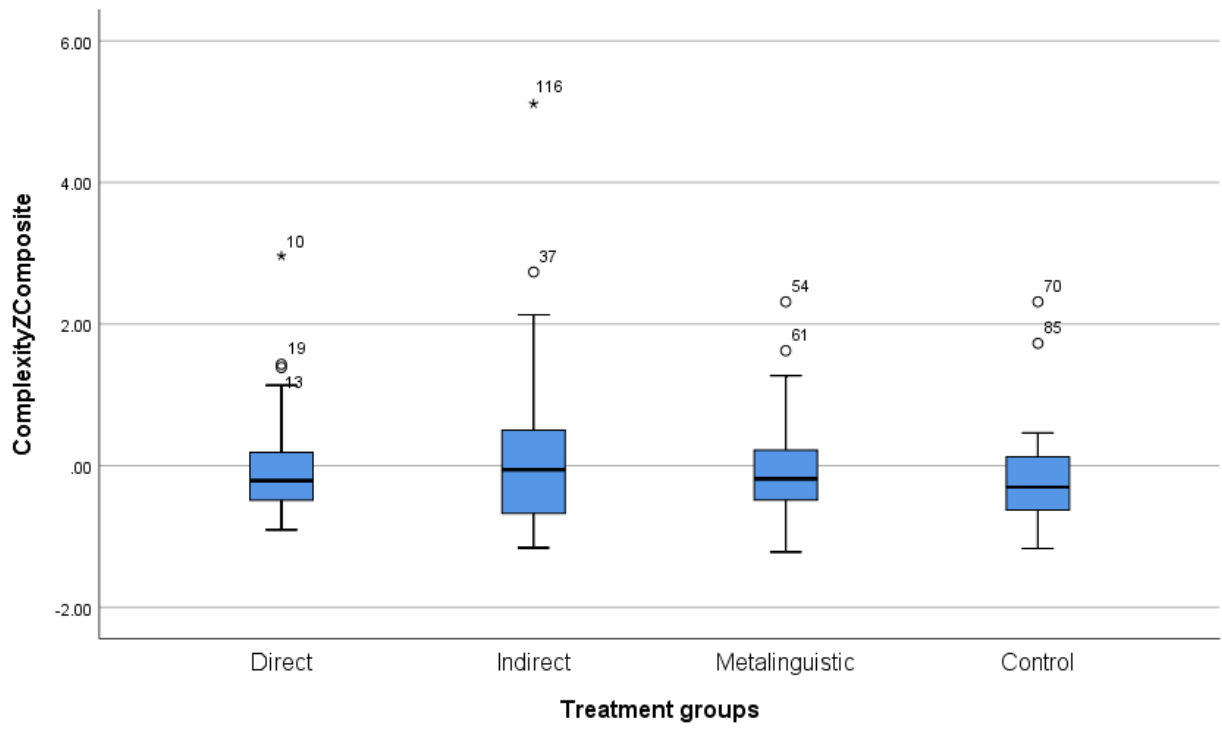
*. This is a lower bound of the true significance.

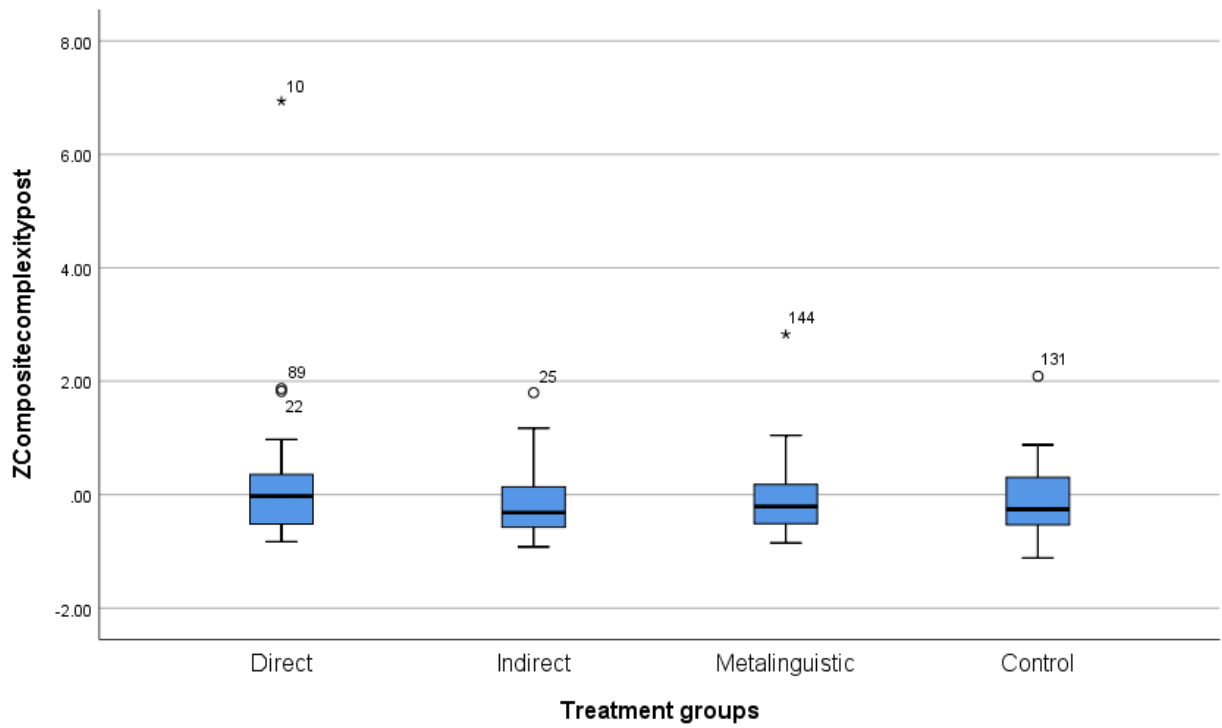
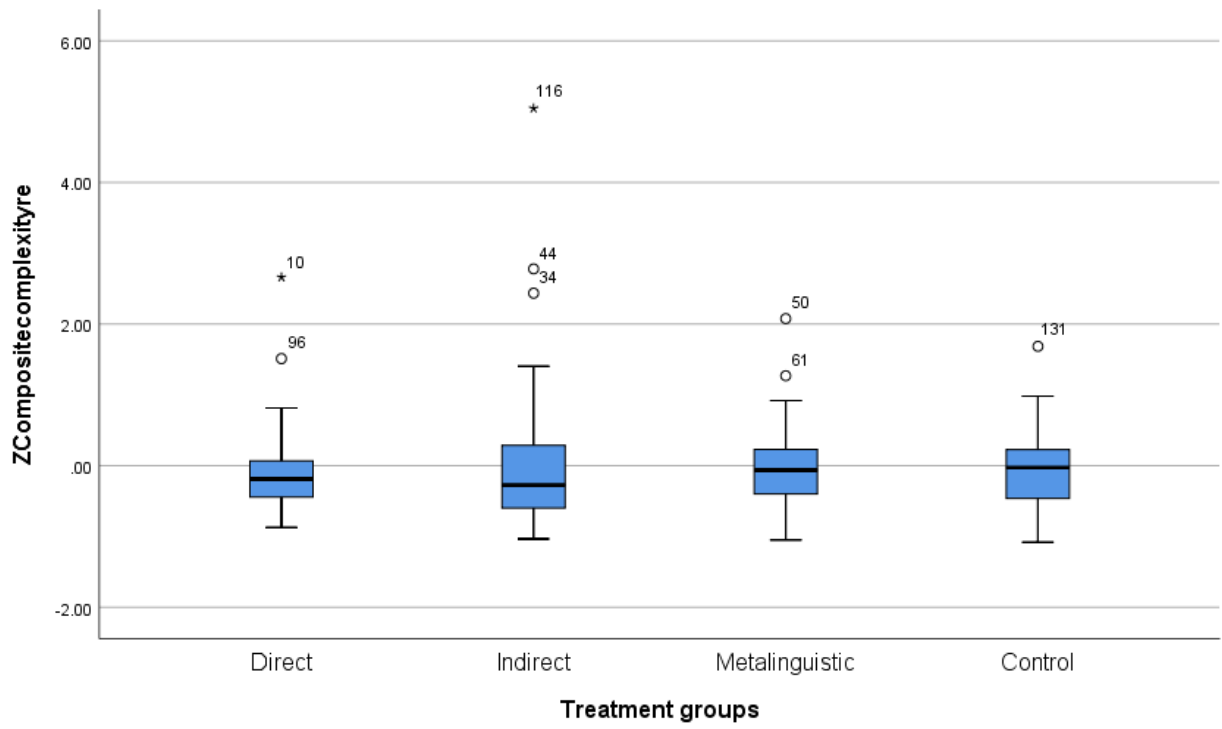
a. Lilliefors Significance Correction

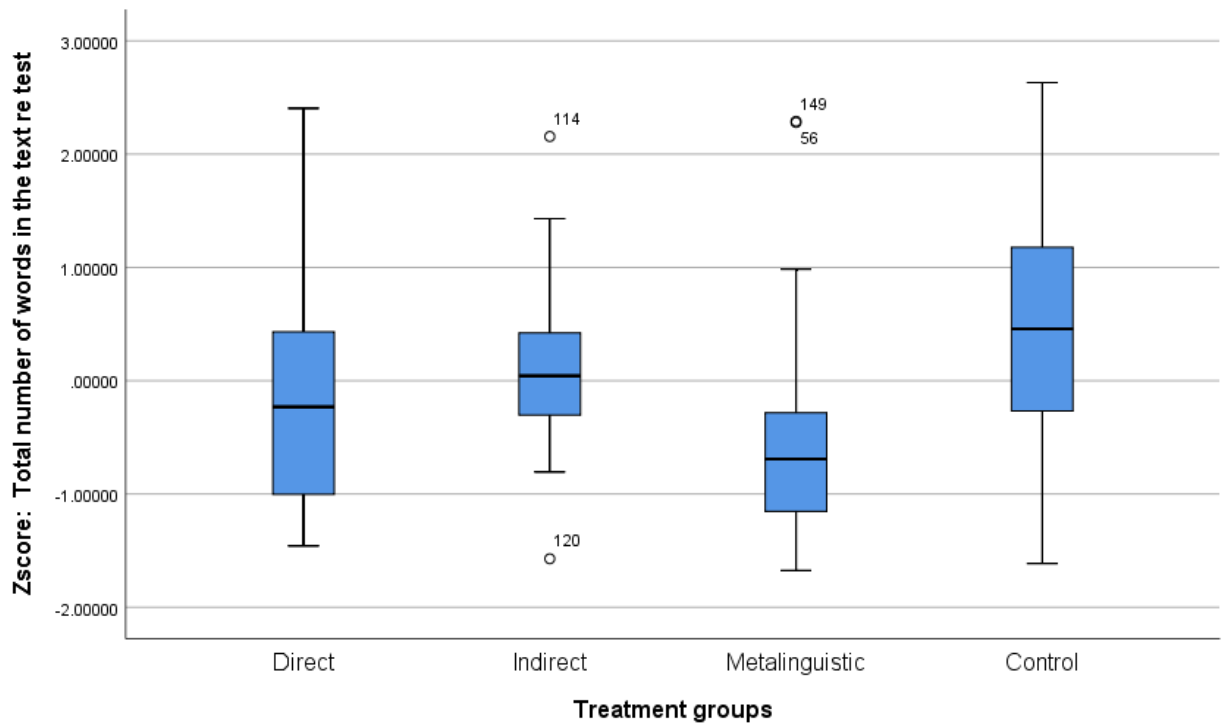
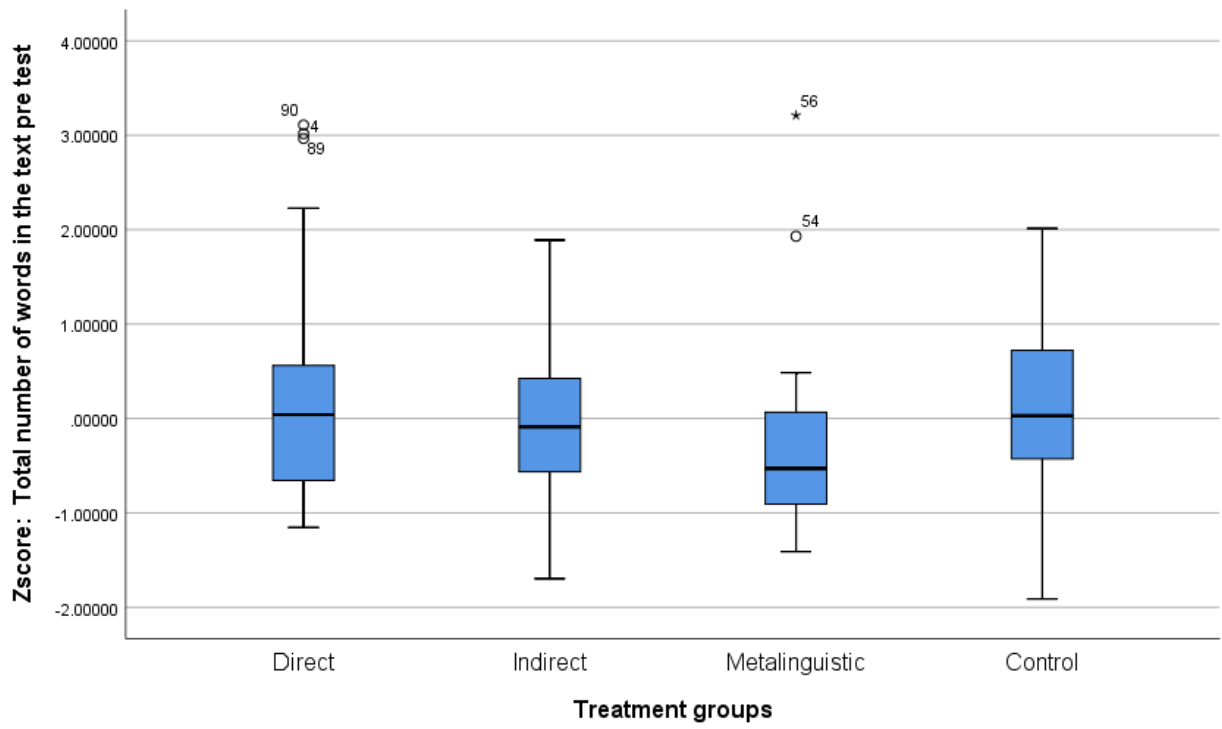
Appendix K: Descriptives Statistics by test for the Composite Variables

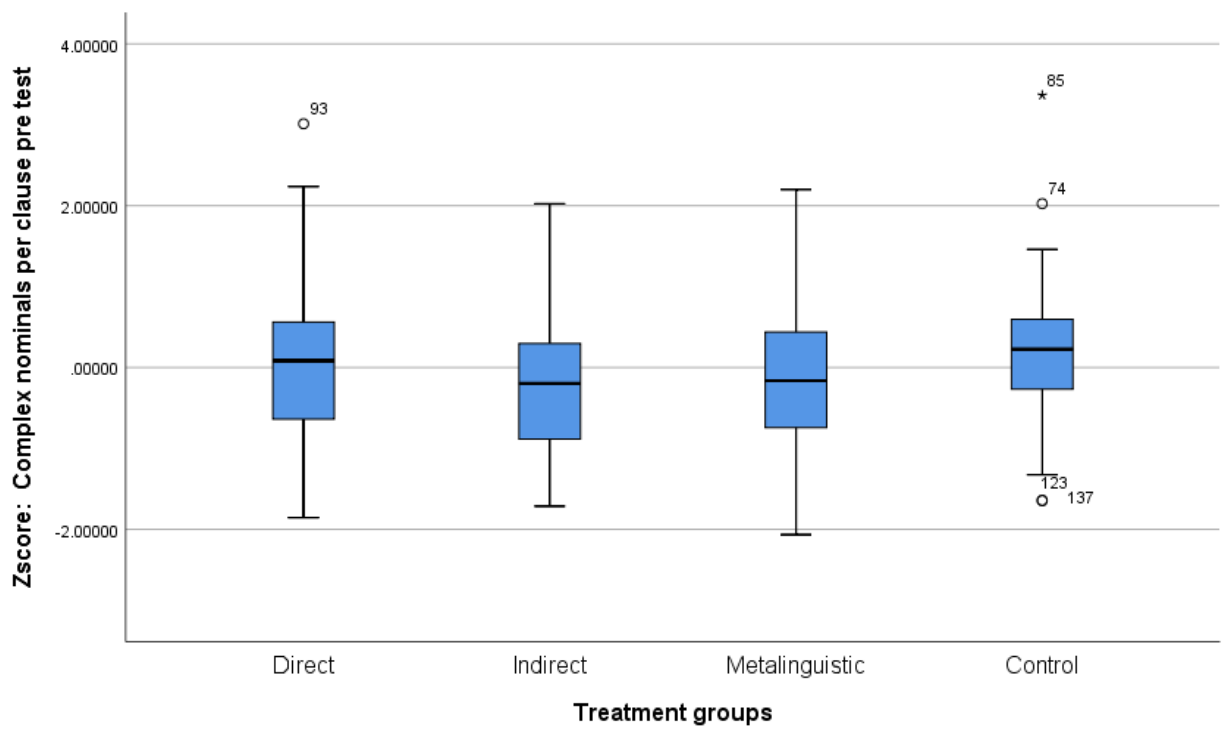
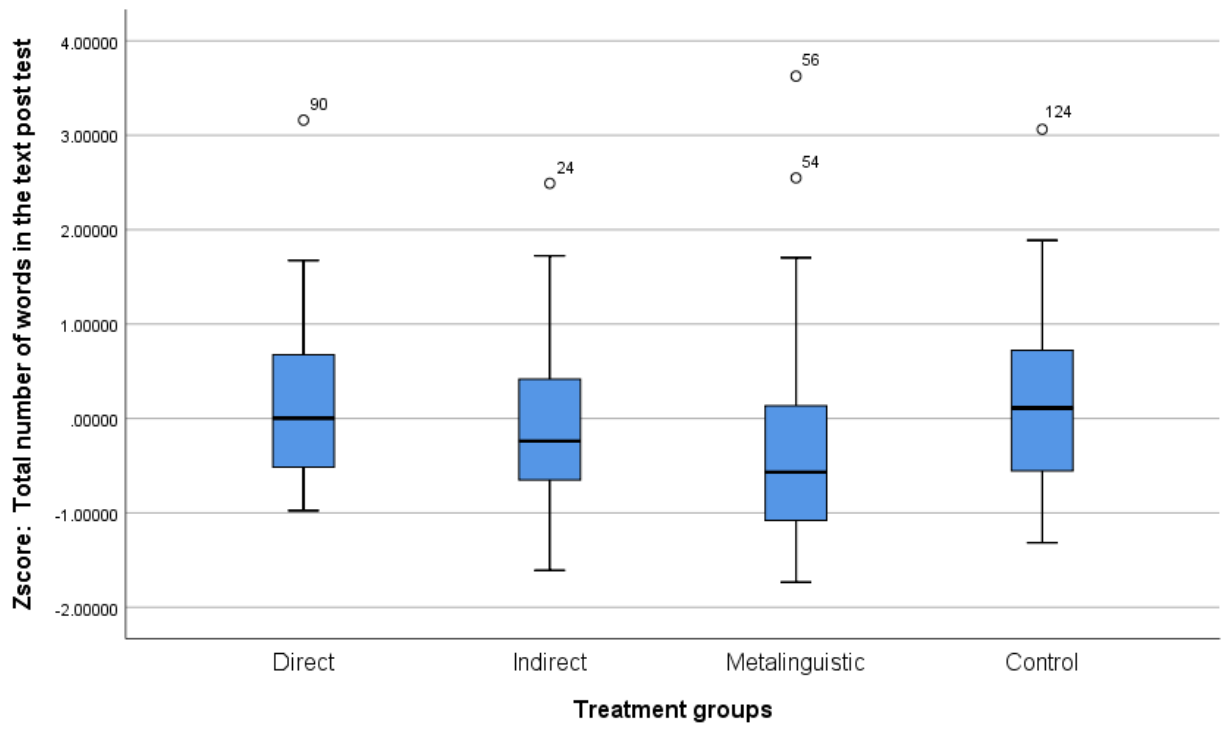


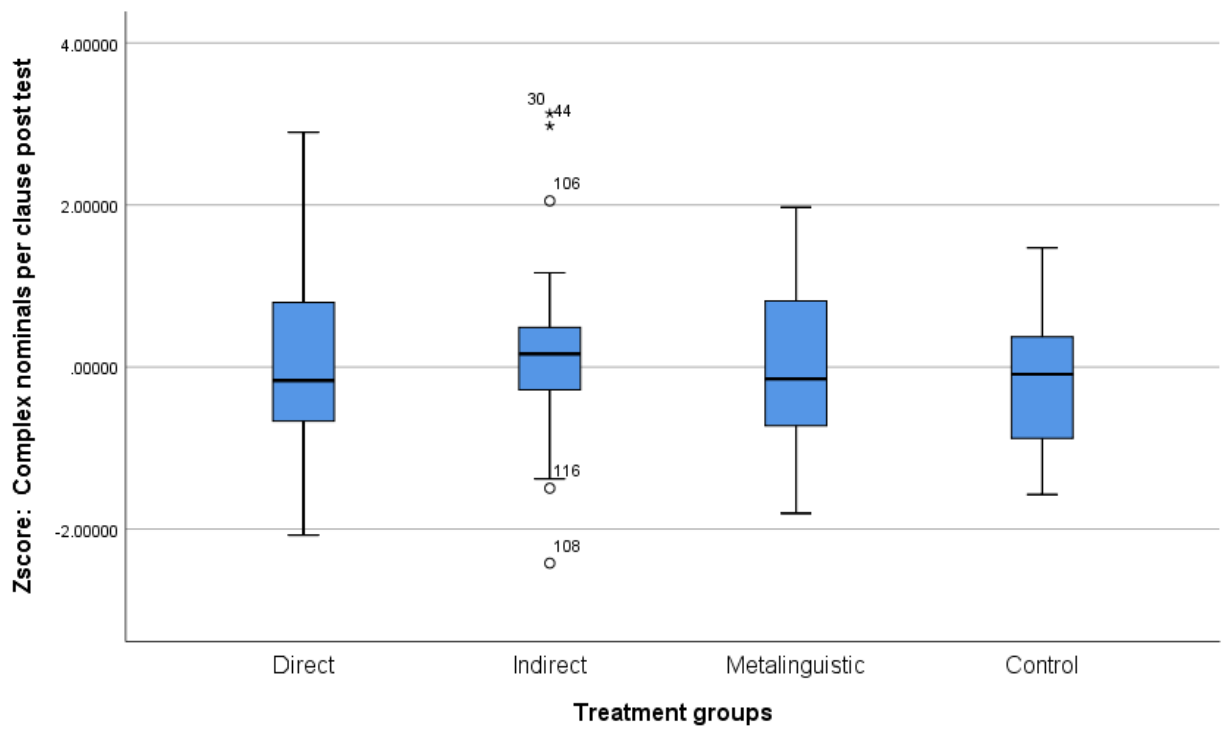
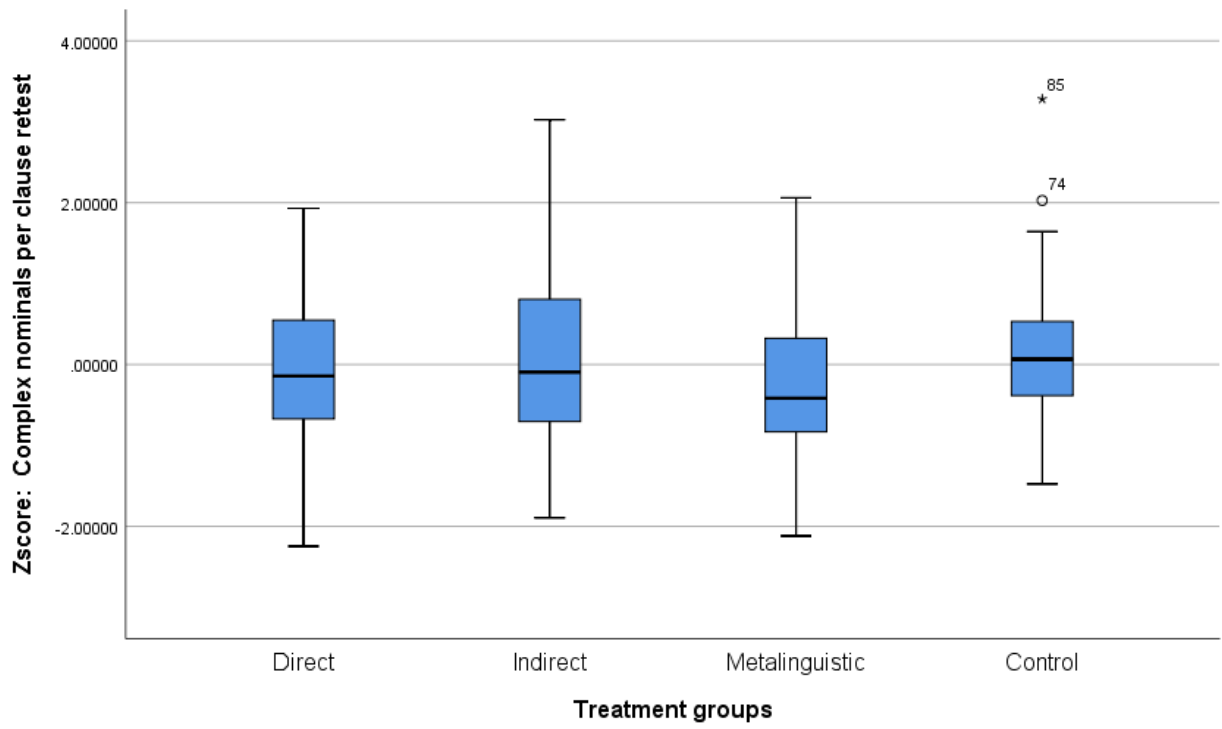


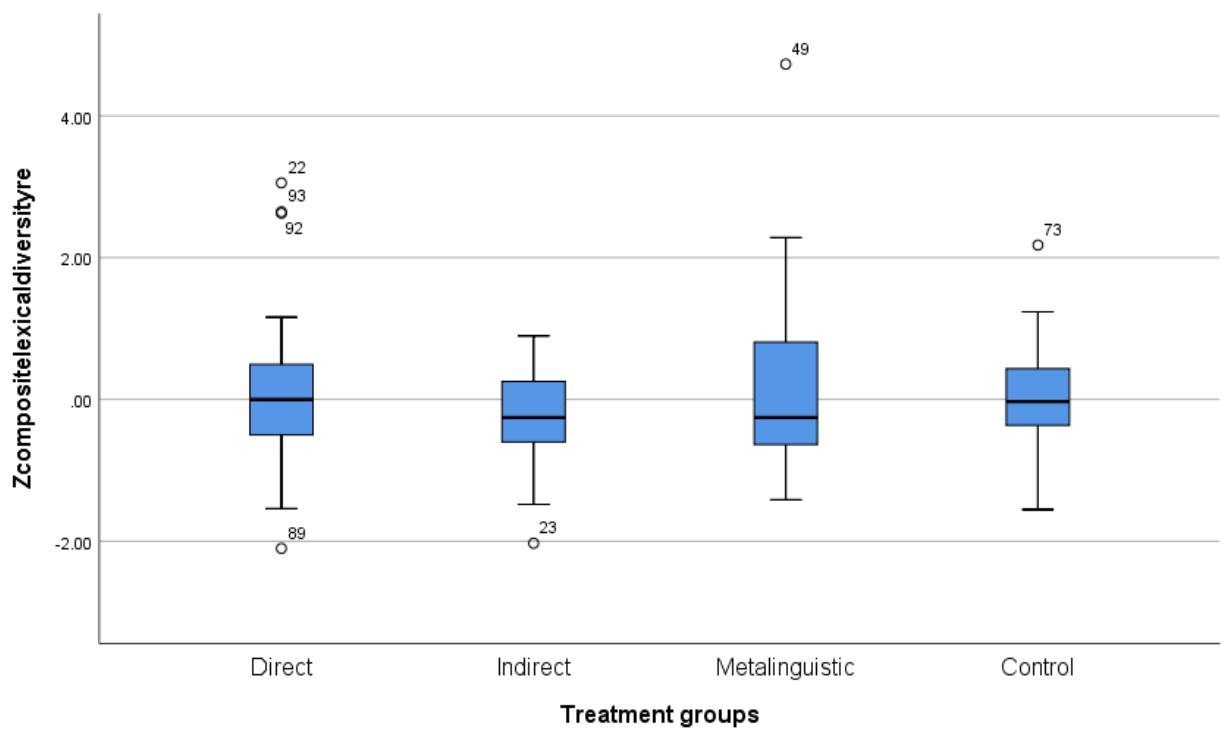
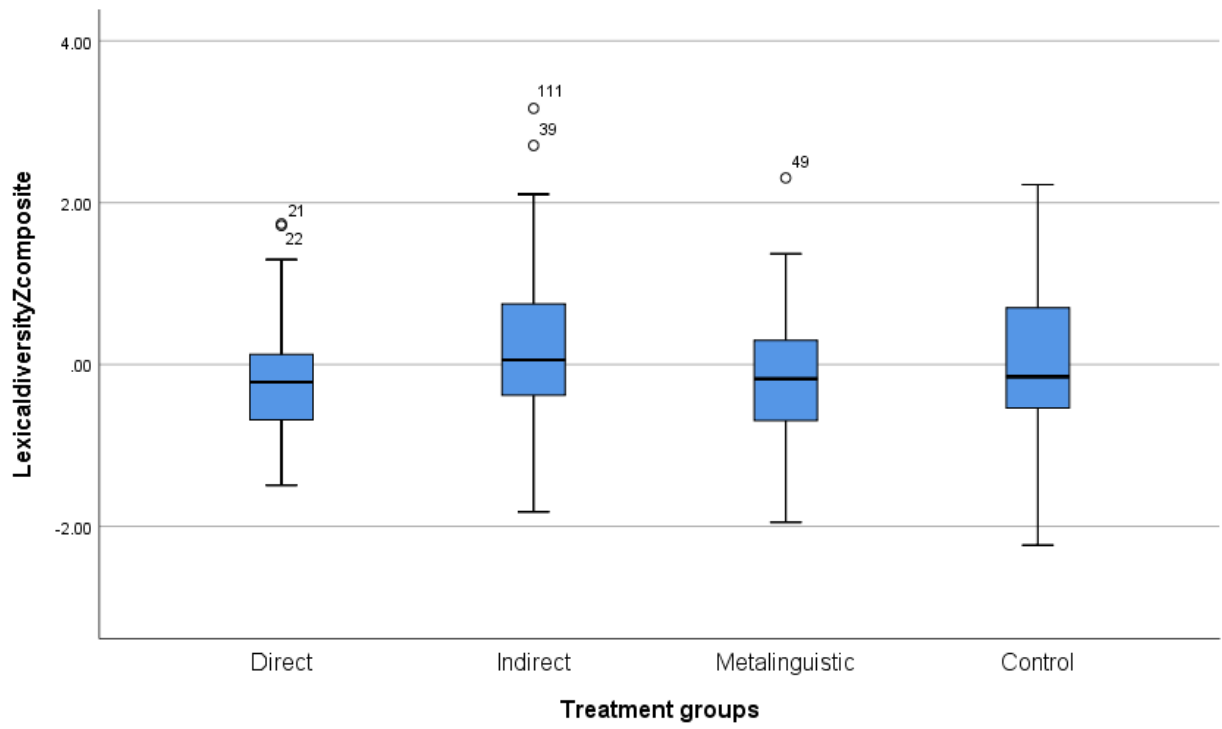


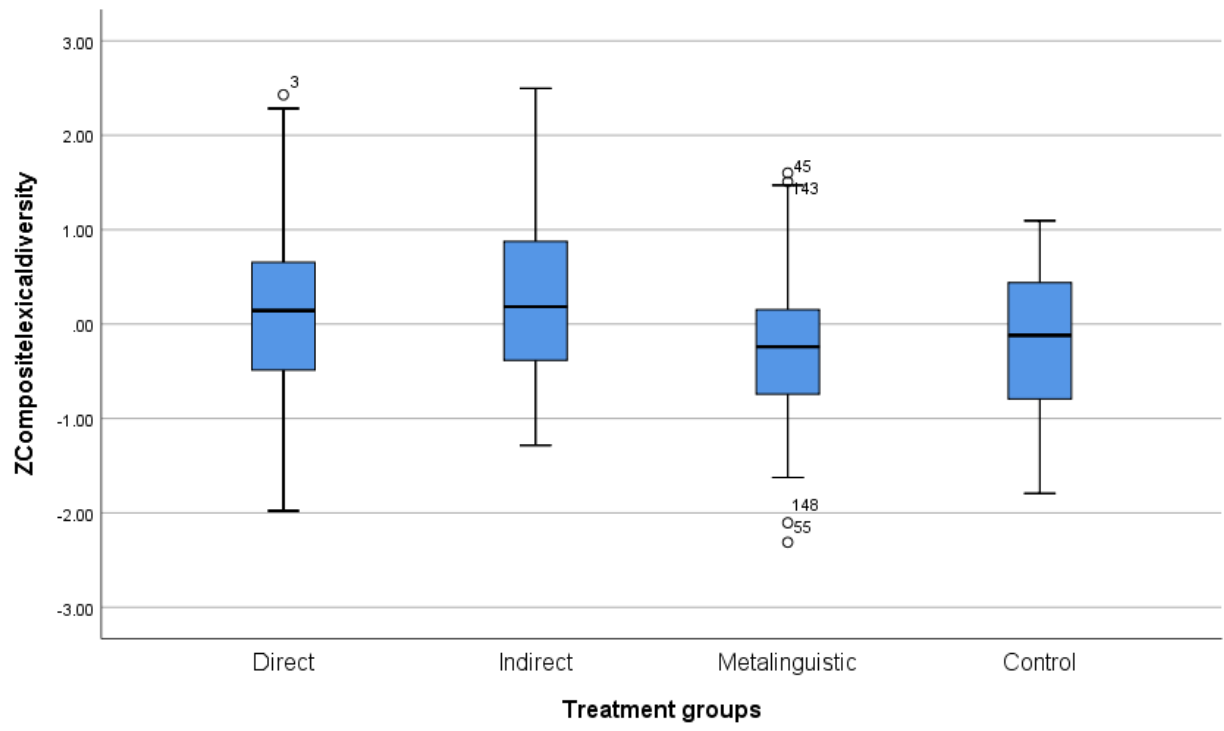












Appendix L : MANCOVA Re-test to Post-test Gains

A MANCOVA was run for re-test to post-test gains. The MANCOVA for re to post-test gains showed that Box's M test was significant, but since this test also checks normality and since some non-normal distributions in the variables would make this result significant and the data used did have some non-normal distribution, proceeding with the MANCOVA in this case was possible. But, Pillai's trace was used due to the significance of Box's M test. Levene's test was significant for complexity gain and lexical diversity gain; therefore, those variables were ignored.

Table 42. Box's Test of Equality of Covariance Matrices^a

Box's Test of Equality of Covariance Matrices^a	
Box's M	101.698
F	2.106
df1	45
df2	41580.098
Sig.	.000
Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.	
a. Design: Intercept + OQPT + LLAMAF + LLAMAB + Group	

Table 43. Levene's Test of Equality of Error Variances

Levene's Test of Equality of Error Variances^a				
	F	df1	df2	Sig.
Fluency gains re post	1.425	3	134	.238
Accuracy gains re post	1.040	3	134	.377
Complex nominal gains	.639	3	134	.591
repost				
Complexity gains re post	3.103	3	134	.029
Lexical diversity gains re	3.893	3	134	.011
post				
Tests the null hypothesis that the error variance of the dependent variable is equal across groups.				
a. Design: Intercept + OQPT + LLAMAF + LLAMAB + Group				

Table 44. Multivariate Tests

Multivariate Tests^a							
	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.029	.765 ^b	5.000	127.000	.576	.029
	Wilks' Lambda	.971	.765 ^b	5.000	127.000	.576	.029
	Hotelling's Trace	.030	.765 ^b	5.000	127.000	.576	.029
	Roy's Largest Root	.030	.765 ^b	5.000	127.000	.576	.029
OQPT	Pillai's Trace	.027	.700 ^b	5.000	127.000	.625	.027
	Wilks' Lambda	.973	.700 ^b	5.000	127.000	.625	.027
	Hotelling's Trace	.028	.700 ^b	5.000	127.000	.625	.027
	Roy's Largest Root	.028	.700 ^b	5.000	127.000	.625	.027
LLAMAF	Pillai's Trace	.054	1.452 ^b	5.000	127.000	.210	.054
	Wilks' Lambda	.946	1.452 ^b	5.000	127.000	.210	.054
	Hotelling's Trace	.057	1.452 ^b	5.000	127.000	.210	.054
	Roy's Largest Root	.057	1.452 ^b	5.000	127.000	.210	.054
LLAMAB	Pillai's Trace	.021	.543 ^b	5.000	127.000	.744	.021
	Wilks' Lambda	.979	.543 ^b	5.000	127.000	.744	.021

	Hotelling's Trace	.021	.543 ^b	5.000	127.000	.744	.021
	Roy's Largest Root	.021	.543 ^b	5.000	127.000	.744	.021
Group	Pillai's Trace	.244	2.280	15.000	387.000	.138	.081
	Wilks' Lambda	.773	2.286	15.000	350.992	.138	.082
	Hotelling's Trace	.272	2.281	15.000	377.000	.138	.083
	Roy's Largest Root	.157	4.056 ^c	5.000	129.000	.138	.136

a. Design: Intercept + OQPT + LLAMAF + LLAMAB + Group

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Appendix M: MANCOVA Pre-test to Re-test Gains

Table 8 (Full Version). MANCOVA Pre-test to Re-test

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.012	.300 ^b	5.000	122.000	.912	.012
	Wilks' Lambda	.988	.300 ^b	5.000	122.000	.912	.012
	Hotelling's Trace	.012	.300 ^b	5.000	122.000	.912	.012
	Roy's Largest Root	.012	.300 ^b	5.000	122.000	.912	.012
OQPT	Pillai's Trace	.019	.470 ^b	5.000	122.000	.798	.019
	Wilks' Lambda	.981	.470 ^b	5.000	122.000	.798	.019
	Hotelling's Trace	.019	.470 ^b	5.000	122.000	.798	.019
	Roy's Largest Root	.019	.470 ^b	5.000	122.000	.798	.019
LLAMAB	Pillai's Trace	.017	.433 ^b	5.000	122.000	.825	.017
	Wilks' Lambda	.983	.433 ^b	5.000	122.000	.825	.017
	Hotelling's Trace	.018	.433 ^b	5.000	122.000	.825	.017
	Roy's Largest Root	.018	.433 ^b	5.000	122.000	.825	.017
LLAMAF	Pillai's Trace	.040	1.025 ^b	5.000	122.000	.406	.040
	Wilks' Lambda	.960	1.025 ^b	5.000	122.000	.406	.040
	Hotelling's Trace	.042	1.025 ^b	5.000	122.000	.406	.040
	Roy's Largest Root	.042	1.025 ^b	5.000	122.000	.406	.040
Attitudes summary	Pillai's Trace	.019	.477 ^b	5.000	122.000	.793	.019

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
	Wilks' Lambda	.981	.477 ^b	5.000	122.000	.793	.019
	Hotelling's Trace	.020	.477 ^b	5.000	122.000	.793	.019
	Roy's Largest Root	.020	.477 ^b	5.000	122.000	.793	.019
Group	Pillai's Trace	.310	2.863	15.000	372.000	.000	.103
	Wilks' Lambda	.708	2.996	15.000	337.190	.000	.109
	Hotelling's Trace	.387	3.112	15.000	362.000	.000	.114
	Roy's Largest Root	.310	7.699 ^c	5.000	124.000	.000	.237

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Table 23 (Full Version). Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Fluency gains pre re	1347.949 ^a	7	192.564	2.541	.018	.124
	Accuracy gains pre re	664.531 ^b	7	94.933	1.192	.312	.062
	Complexity gains pre re	267.069 ^c	7	38.153	.564	.784	.030
	Lexical diversity gains pre re	1621.830 ^d	7	231.690	2.488	.020	.121
	Complex nominal gains pre re	650.647 ^e	7	92.950	1.487	.178	.076
Intercept	Fluency gains pre re	.010	1	.010	.000	.991	.000
	Accuracy gains pre re	.535	1	.535	.007	.935	.000
	Complexity gains pre re	37.134	1	37.134	.549	.460	.004
	Lexical diversity gains pre re	12.707	1	12.707	.136	.712	.001
	Complex nominal gains pre re	43.026	1	43.026	.688	.408	.005
OQPT	Fluency gains pre re	10.599	1	10.599	.140	.709	.001
	Accuracy gains pre re	14.695	1	14.695	.184	.668	.001
	Complexity gains pre re	92.650	1	92.650	1.370	.244	.011
	Lexical diversity gains pre re	22.658	1	22.658	.243	.623	.002
	Complex nominal gains pre re	22.787	1	22.787	.365	.547	.003
LLAMAB	Fluency gains pre re	2.662	1	2.662	.035	.852	.000
	Accuracy gains pre re	9.673	1	9.673	.121	.728	.001
	Complexity gains pre re	31.202	1	31.202	.462	.498	.004
	Lexical diversity gains pre re	2.237	1	2.237	.024	.877	.000

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
	Complex nominal gains pre re	91.700	1	91.700	1.467	.228	.012
LLAMAF	Fluency gains pre re	.669	1	.669	.009	.925	.000
	Accuracy gains pre re	130.000	1	130.000	1.632	.204	.013
	Complexity gains pre re	.370	1	.370	.005	.941	.000
	Lexical diversity gains pre re	240.529	1	240.529	2.583	.111	.020
	Complex nominal gains pre re	76.768	1	76.768	1.228	.270	.010
Attitudes summary	Fluency gains pre re	27.365	1	27.365	.361	.549	.003
	Accuracy gains pre re	9.720	1	9.720	.122	.727	.001
	Complexity gains pre re	9.993	1	9.993	.148	.701	.001
	Lexical diversity gains pre re	103.080	1	103.080	1.107	.295	.009
	Complex nominal gains pre re	22.375	1	22.375	.358	.551	.003
Group	Fluency gains pre re	1237.291	3	412.430	5.442	.001	.115
	Accuracy gains pre re	400.303	3	133.434	1.675	.176	.038
	Complexity gains pre re	114.181	3	38.060	.563	.640	.013
	Lexical diversity gains pre re	1355.065	3	451.688	4.850	.003	.104
	Complex nominal gains pre re	346.957	3	115.652	1.850	.141	.042
Error	Fluency gains pre re	9548.872	126	75.785			
	Accuracy gains pre re	10038.063	126	79.667			
	Complexity gains pre re	8518.106	126	67.604			

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
	Lexical diversity gains pre re	11734.801	126	93.133			
	Complex nominal gains pre re	7875.204	126	62.502			
Total	Fluency gains pre re	10896.822	134				
	Accuracy gains pre re	10703.481	134				
	Complexity gains pre re	8796.212	134				
	Lexical diversity gains pre re	13360.262	134				
	Complex nominal gains pre re	8526.129	134				
Corrected Total	Fluency gains pre re	10896.821	133				
	Accuracy gains pre re	10702.593	133				
	Complexity gains pre re	8785.175	133				
	Lexical diversity gains pre re	13356.631	133				
	Complex nominal gains pre re	8525.851	133				
a. R Squared = .124 (Adjusted R Squared = .075)							
b. R Squared = .062 (Adjusted R Squared = .010)							
c. R Squared = .030 (Adjusted R Squared = -.023)							
d. R Squared = .121 (Adjusted R Squared = .073)							

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
--------	--------------------	-------------------------	----	-------------	---	------	---------------------

c. R Squared = .076 (Adjusted R Squared = .025)

Appendix N: MANCOVA Pre-test to Post-test Gains

Table 27 (Full Version). Multivariate Tests

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.017	.421 ^b	5.000	122.000	.834	.017
	Wilks' Lambda	.983	.421 ^b	5.000	122.000	.834	.017
	Hotelling's Trace	.017	.421 ^b	5.000	122.000	.834	.017
	Roy's Largest Root	.017	.421 ^b	5.000	122.000	.834	.017
OQPT	Pillai's Trace	.026	.640 ^b	5.000	122.000	.670	.026
	Wilks' Lambda	.974	.640 ^b	5.000	122.000	.670	.026
	Hotelling's Trace	.026	.640 ^b	5.000	122.000	.670	.026
	Roy's Largest Root	.026	.640 ^b	5.000	122.000	.670	.026
LLAMAB	Pillai's Trace	.021	.532 ^b	5.000	122.000	.752	.021
	Wilks' Lambda	.979	.532 ^b	5.000	122.000	.752	.021
	Hotelling's Trace	.022	.532 ^b	5.000	122.000	.752	.021
	Roy's Largest Root	.022	.532 ^b	5.000	122.000	.752	.021
LLAMAF	Pillai's Trace	.020	.492 ^b	5.000	122.000	.781	.020
	Wilks' Lambda	.980	.492 ^b	5.000	122.000	.781	.020
	Hotelling's Trace	.020	.492 ^b	5.000	122.000	.781	.020
	Roy's Largest Root	.020	.492 ^b	5.000	122.000	.781	.020
Attitudes summary	Pillai's Trace	.065	1.698 ^b	5.000	122.000	.140	.065
	Wilks' Lambda	.935	1.698 ^b	5.000	122.000	.140	.065

	Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
	Hotelling's Trace	.070	1.698 ^b	5.000	122.000	.140	.065
	Roy's Largest Root	.070	1.698 ^b	5.000	122.000	.140	.065
Group	Pillai's Trace	.184	1.623	15.000	372.000	.065	.061
	Wilks' Lambda	.824	1.631	15.000	337.190	.064	.062
	Hotelling's Trace	.203	1.634	15.000	362.000	.063	.063
	Roy's Largest Root	.127	3.151 ^c	5.000	124.000	.010	.113

a. Design: Intercept + OQPT + LLAMAB + LLAMAF + Attitudes summary + Group

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Table 98 (Full Version). Between-Subjects Effects

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Fluency gain pre post	328.879 ^a	7	46.983	.749	.631
	Accuracy gain pre post	1065.100 ^b	7	152.157	1.320	.246
	Complexity gain pre post	818.610 ^c	7	116.944	1.530	.163
	Complex nominal gain pre post	1239.538 ^d	7	177.077	1.461	.187
	Lexical diversity gain pre post	794.498 ^e	7	113.500	1.093	.372
Intercept	Fluency gain pre post	13.013	1	13.013	.207	.650
	Accuracy gain pre post	98.605	1	98.605	.856	.357
	Complexity gain pre post	75.447	1	75.447	.987	.322
	Complex nominal gain pre post	12.794	1	12.794	.106	.746
	Lexical diversity gain pre post	.438	1	.438	.004	.948
OQPT	Fluency gain pre post	5.036	1	5.036	.080	.777
	Accuracy gain pre post	342.149	1	342.149	2.969	.087
	Complexity gain pre post	5.962	1	5.962	.078	.780
	Complex nominal gain pre post	33.181	1	33.181	.274	.602
	Lexical diversity gain pre post	4.458	1	4.458	.043	.836
LLAMAF	Fluency gain pre post	.045	1	.045	.001	.979
	Accuracy gain pre post	25.833	1	25.833	.224	.637
	Complexity gain pre post	150.979	1	150.979	1.975	.162
	Complex nominal gain pre post	4.281	1	4.281	.035	.851
	Lexical diversity gain pre post	10.055	1	10.055	.097	.756

LLAMAB	Fluency gain pre post	32.906	1	32.906	.525	.470
	Accuracy gain pre post	150.570	1	150.570	1.306	.255
	Complexity gain pre post	26.696	1	26.696	.349	.556
	Complex nominal gain pre post	28.526	1	28.526	.235	.628
	Lexical diversity gain pre post	7.976	1	7.976	.077	.782
Attitudes summary	Fluency gain pre post	238.558	1	238.558	3.804	.053
	Accuracy gain pre post	114.941	1	114.941	.997	.320
	Complexity gain pre post	305.145	1	305.145	3.992	.048
	Complex nominal gain pre post	37.779	1	37.779	.312	.578
	Lexical diversity gain pre post	102.039	1	102.039	.982	.324
Group	Fluency gain pre post	35.399	3	11.800	.188	.904
	Accuracy gain pre post	387.661	3	129.220	1.121	.343
	Complexity gain pre post	305.659	3	101.886	1.333	.267
	Complex nominal gain pre post	1008.649	3	336.216	2.774	.044
	Lexical diversity gain pre post	664.593	3	221.531	2.133	.099
Error	Fluency gain pre post	7902.508	126	62.718		
	Accuracy gain pre post	14521.849	126	115.253		
	Complexity gain pre post	9631.190	126	76.438		
	Complex nominal gain pre post	15269.462	126	121.186		

	Lexical diversity gain pre post	13086.954	126	103.865
Total	Fluency gain pre post	8232.344	134	
	Accuracy gain pre post	15596.539	134	
	Complexity gain pre post	10478.123	134	
	Complex nominal gain pre post	16517.169	134	
	Lexical diversity gain pre post	13897.903	134	
Corrected Total	Fluency gain pre post	8231.387	133	
	Accuracy gain pre post	15586.949	133	
	Complexity gain pre post	10449.800	133	
	Complex nominal gain pre post	16509.000	133	
	Lexical diversity gain pre post	13881.452	133	
a. R Squared = .040 (Adjusted R Squared = -.013)				
b. R Squared = .068 (Adjusted R Squared = .017)				
c. R Squared = .078 (Adjusted R Squared = .027)				
d. R Squared = .075 (Adjusted R Squared = .024)				
e. R Squared = .057 (Adjusted R Squared = .005)				

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Run pairwise comparisons approaching significance it is a trend.

Appendix O: Pearson Correlations Pre-test to Post-test

Table 31. Correlations Pre-test to Post-test Gains

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains pre post	Accuracy gains pre post	Complexity gains pre post	Complex nominal gains pre post
LLAMA F Aptitude test	Pearson	.117							
	Correlation								
	Sig. (2-tailed)	.179							
LLAMA B Aptitude test	Pearson	-.052	.189*						
	Correlation								
	Sig. (2-tailed)	.551	.028						
Attitudes summary score	Pearson	-.029	.056	.050					
	Correlation								
	Sig. (2-tailed)	.739	.523	.566					
Fluency gains pre post	Pearson	.036	-.003	.061	-.172*				
	Correlation								
	Sig. (2-tailed)	.682	.974	.482	.047				
Accuracy gains pre post	Pearson	.142	.022	-.127	-.098	.090			
	Correlation								

		Oxford Quick Placement Test	LLAMA F Aptitude test	LLAMA B Aptitude test	Attitudes summary score	Fluency gains pre post	Accuracy gains pre post	Complexity gains pre post	Complex nominal gains pre post
	Sig. (2-tailed)	.101	.805	.143	.262	.302			
Complexity gains pre post	Pearson Correlation	-.036	.102	-.063	-.165	.140	.086		
	Sig. (2-tailed)	.683	.242	.467	.057	.107	.325		
Complex nominal gains pre post	Pearson Correlation	.076	.072	.064	-.029	.055	.063	.001	
	Sig. (2-tailed)	.385	.408	.466	.739	.529	.470	.990	
Lexical diversity gains pre post	Pearson Correlation	-.012	.050	.002	-.077	.136	-.024	-.089	-.049
	Sig. (2-tailed)	.894	.564	.984	.377	.117	.779	.305	.577

ENDS