

Developing a Rating Scale for Classroom Assessment of the Argumentative  
Writing of Chinese EFL College Students Majoring in English

Thesis submitted for the degree of Doctor of Philosophy  
at the University of Leicester

by  
Keke Zhang

2019

## **Abstract**

### **Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English**

Keke Zhang

The argumentative writing ability of Chinese EFL college students majoring in English is frequently assessed with both formative and summative assessments in the classroom. However, rating scales used in these assessments fail to provide precise and clear scoring criteria and fail to adequately evaluate the quality of arguments/argumentation.

By adopting a theoretically-based data-driven approach proposed by Knoch (2009), this study aims to develop a rating system appropriate for classroom assessment of the argumentative writing ability of Chinese EFL college students majoring in English and to investigate its usability by raters/Chinese EFL writing teachers. This study was undertaken in two phases. In phase one, 258 writing scripts were selected from a corpus of argumentative writing of Chinese EFL college students built for this study, and were analyzed using discourse analytic measures. The results of the main study were used to create descriptors of the new rating scale. In phase two, 30 writing scripts, representing a wide range of writing proficiency levels, were rated by three writing teachers (or raters). Their ratings using the new rating scale were analyzed using inter-rater reliability analysis, and follow-up questionnaire feedback from the raters was analyzed using content analysis.

The new rating scale was shown to be generally adequate to represent the construct of argumentative writing ability, useful for providing detailed feedback for the teaching and assessment of argumentative writing ability in the classroom, and reliable for three of five trait scales – mechanics, fluency and argumentation. The new rating scale was also shown to be not practical, not authentic as a whole and not reliable for the scales of accuracy and coherence, although no exact reason for the unreliability was found. Findings are discussed in terms of their implications for assessment of argumentation ability, rating scale development, and score reporting in classroom-based assessment in the Chinese EFL context.

## **Acknowledgements**

I would like to express my sincere gratitude and appreciation to all the people who have helped me with my thesis for their encouragement, inspiration and support.

I owe my truest gratitude to Prof. Glenn Fulcher, who offered me the opportunity to pursue PhD study at University of Leicester. Professor Fulcher has been a patient and caring supervisor. He provided me with expert guidance and constant support throughout this research. I am especially grateful to him for reading my dissertation and providing me with inspiring advice during the writing process.

I would like to extend my deep gratitude to the professors at the University of Leicester for their mind-opening lectures. I am equally grateful to Dr. Agneta Svalberg and Dr. Nick Smith at the School of Arts who gave me valuable suggestions and advice on data analysis.

I would also like to extend my gratitude to Prof. Wu Xianyou, the Dean of the Foreign Language School at Chongqing Normal University in China, who provided me with access to teacher and student participants for this study. I would also like to extend my sincere thanks to teacher and student participants for their time and cooperation. They actively took part in writing tests, interviews, and questionnaires and honestly shared their thoughts with me. I would also like to express my gratitude to colleagues at the Foreign Languages School for their participation in this study as raters, error coders, and double coders and to students who did the tedious typing task.

My special thanks go to my husband, my parents, my parents-in-law and my daughter, who was born healthy and lovely during my PhD study. Their constant love, understanding and support enabled me to finish this thesis. I especially want to say thank you to my husband, who offered great comfort and support when I was frustrated, and great encouragement when I made progress.

Finally, I would like to express my gratitude to the China Scholarship Council who funded my PhD study, and Educational Testing Service who provided me with an additional Doctoral Grant. Without their financial support, this thesis would not have been possible.

## Table of Contents

Abstract .....	ii
Acknowledgements.....	iii
List of Tables.....	x
List of Figures.....	xiv
List of Abbreviations .....	xvi
Chapter 1 Introduction.....	1
1.1 Background to the study .....	1
1.2 Statement of the problem.....	2
1.3 Purpose and nature of the study.....	4
1.4 Research questions .....	5
1.5 Outline of the thesis .....	5
Chapter 2 Rating Scales.....	7
2.1 Definition of a rating scale .....	7
2.2 Types of rating scales .....	8
2.2.1 Holistic scales .....	9
2.2.2 Analytic scales .....	11
2.2.3 Primary trait scales .....	12
2.2.4 Multiple trait scales .....	15
2.2.5 Implications for the new rating scale.....	17
2.3 Important features of rating scales .....	18
2.3.1 Functions .....	18
2.3.2 Categories/traits .....	19
2.3.3 Distinguishing between levels.....	20
2.3.4 Descriptor formulation .....	21
2.3.5 Number of bands .....	22
2.3.6 Weighting of categories .....	23
2.3.7 Score reporting .....	24
2.4 Design method of rating scales.....	24
2.4.1 Intuitive method.....	25
2.4.2 Theoretically-based method .....	26

2.4.3	Empirically-developed method .....	27
2.4.4	A hybrid approach .....	30
2.4.5	Implications for the new rating scale.....	32
2.5	The current TEM4 rating scale and its issues .....	33
2.5.1	The considerations of the selection of the current TEM4 rating scale .....	33
2.5.2	The current TEM4 rating scale .....	33
2.5.3	The development of the current TEM4 rating scale of writing .....	35
2.5.4	The strengths and issues of the current TEM4 rating scale of writing .....	39
Chapter 3 A theoretical model of argumentative writing ability and its operational framework .....		42
3.1	Theoretical basis for scale design .....	42
3.2	Building the theoretical model of argumentative writing ability (the AWA model) .....	46
3.2.1	General components .....	48
3.2.2	Argumentation knowledge .....	53
3.2.3	Developing the AWA model .....	64
3.3	An operational framework of argumentative writing ability .....	66
3.3.1	Grammatical knowledge.....	67
3.3.2	Textual knowledge.....	76
3.3.3	Sociolinguistic knowledge.....	87
3.3.4	Argumentation knowledge .....	90
3.4	Conclusion .....	96
Chapter 4 Methodology- Analysis of writing scripts.....		99
4.1	Research design .....	99
4.2	Research questions .....	100
4.3	Research instruments .....	101
4.3.1	Writing tasks .....	101
4.3.2	Rating scale .....	106
4.4	Research participants .....	107
4.4.1	Raters .....	107
4.4.2	Students .....	108
4.5	The development of the Chinese EFL college students' argumentative	

writing corpus (the CEAW corpus).....	110
4.5.1 Administration of writing tests.....	110
4.5.2 Collecting writing samples and text digitization.....	111
4.5.3 Rater training exercise and rating.....	111
4.5.4 Data entry .....	114
4.6 Pilot study.....	115
4.6.1 Data selection .....	115
4.6.2 Data analysis and results .....	115
4.7 Main study .....	154
4.7.1 Data selection .....	154
4.7.2 Data analysis.....	159
4.7.3 Data analysis – Inferential statistics .....	162
Chapter 5 Results and discussion – Analysis of writing scripts .....	166
5.1 Mechanics.....	167
5.1.1 Results .....	168
5.1.2 Trial scale for mechanics .....	170
5.2 Fluency .....	170
5.2.1 Results .....	170
5.2.2 Trial scale for fluency .....	172
5.3 Accuracy.....	173
5.3.1 Results .....	173
5.3.2 Trial scale for accuracy.....	175
5.4 Cohesion.....	176
5.5 Coherence.....	178
5.5.1 Results .....	179
5.5.2 Trial scale for coherence.....	193
5.6 Argument structure .....	195
5.6.1 Results .....	195
5.6.2 Trial scale for argumentation.....	213
5.7 Conclusion.....	216
Chapter 6 Methodology – Analysis of ratings and questionnaire data .....	218
6.1 Research design .....	218
6.2 Participants .....	220
6.2.1 Students .....	220

6.2.2	Raters .....	222
6.3	Instruments .....	223
6.3.1	Writing scripts .....	223
6.3.2	New rating scale .....	224
6.3.3	Rating sheets.....	224
6.3.4	Training materials.....	224
6.3.5	Questionnaire.....	225
6.4	Procedures .....	226
6.4.1	Trial of new rating scale .....	226
6.4.2	Rater training .....	230
6.4.3	Data collection.....	230
6.4.4	Data analysis.....	231
Chapter 7	Results and discussion – Analysis of ratings and questionnaire data .....	235
7.1	Results of inter-rater reliability.....	235
7.2	Raters’ perceptions of the new rating scale .....	240
7.3	Discussion.....	246
7.3.1	Construct validity .....	246
7.3.2	Reliability .....	254
7.3.3	Authenticity .....	256
7.3.4	Impact.....	257
7.3.5	Practicality .....	258
7.3.6	Conclusion.....	259
Chapter 8	Conclusion .....	261
8.1	Introduction .....	261
8.2	Findings .....	262
8.3	Implications .....	267
8.3.1	Theoretical implications .....	267
8.3.2	Practical implications .....	271
8.4	Limitations.....	274
8.5	Suggestions for further research.....	277
8.6	Conclusion.....	279
Appendices.....		280
Appendix 1:	Interview guide: Developing argumentative writing tasks for Chinese	



EFL learners in writing courses .....	280
Appendix 2: Questionnaire on the difficulty of argumentative writing tasks....	281
Appendix 3: A taxonomy of errors in English writing adapted from the CEM error taxonomy (2008).....	284
Appendix 4: An exemplar essay coded for the TSA and a diagram of different topical progression types .....	293
Appendix 5: Argumentation elements coding manual (adapted from a brief version of Chase’s argumentative elements, 2011, p. 98–100) .....	296
Appendix 6: Rating Sheet: The TEM4 rating scale .....	304
Appendix 7: Rating sheet: The new rating scale .....	305
Appendix 8: Training manual .....	307
Appendix 9: Test paper 1 .....	312
Appendix 10: Test paper 2 .....	314
Appendix 11: Test paper 3 .....	316
Appendix 12: Student participant information sheet .....	318
Appendix 13: Student informed consent form.....	320
Appendix 14: Rater participant information sheet.....	321
Appendix 15: Rater informed consent form .....	324
Appendix 16: Teacher participant information sheet.....	325
Appendix 17: Teacher informed consent form .....	328
Appendix 18: Codes and calculations.....	329
Bibliography .....	330

## List of Tables

Table 2.1 Four types of rating scales (based on Weigle, 2002, p. 109; North, 2003, p. 64) .....	8
Table 3.1 Comparison between the CLA model, part of Grabe and Kaplan's (1996) taxonomy, and Connor and Mbaye's (2002) model of writing competence ...	50
Table 3.2 The definitions of rational appeals (adapted from Anthony & Gladkov, 2007, p.124–5) .....	62
Table 3.3 The definitions of credibility appeals (adapted from Anthony & Gladkov, 2007, p. 124–5) .....	63
Table 3.4 The definitions of affective appeals (adapted from Anthony & Gladkov, 2007, p. 124–5) .....	64
Table 3.5 The AWA model without strategic competence and knowledge of the world .....	65
Table 3.6 Moderate level of a rating scale of register knowledge (Bachman & Palmer, 1996, p. 288) .....	88
Table 3.7 Rational appeal scale.....	95
Table 3.8 The operational framework of argumentative writing ability .....	97
Table 4.1 Raters' background .....	107
Table 4.2 Gender distribution .....	108
Table 4.3 Major distribution .....	109
Table 4.4 Year distribution .....	109
Table 4.5 University background.....	110
Table 4.6 Distribution of final scores awarded to scripts in the CEAW corpus ...	113
Table 4.7 Descriptive analysis – Final scores awarded to scripts in the CEAW corpus .....	114
Table 4.8 Inter-rater reliability for the three raters .....	114
Table 4.9 Descriptive statistics – Accuracy .....	117
Table 4.10 Descriptive analysis – Fluency .....	119
Table 4.11 Descriptive analysis – grammatical complexity .....	120
Table 4.12 Descriptive statistics – Lexical complexity .....	122
Table 4.13 Descriptive statistics – Mechanics .....	123
Table 4.14 Descriptive statistics – Cohesion .....	128
Table 4.15 Definitions and examples of metadiscourse markers, adapted from Dafouz-Milne (2008, p. 98, italics added) .....	137
Table 4.16 Descriptive analysis – Coherence .....	139
Table 4.17 Descriptive statistics – Register .....	141
Table 4.18 Descriptive analysis – Argumentation structure elements .....	145

Table 4.19 Descriptive analysis – Acceptability of reasons.....	148
Table 4.20 Descriptive statistics – Appeals of argumentation .....	151
Table 4.21 Descriptive statistics – Number of paragraphs per element.....	153
Table 4.22 Measures to be used in the main analysis .....	153
Table 4.23 Levels of writing scripts in sample and in the CEAW corpus .....	156
Table 4.24 Descriptive statistics – Final scores awarded to writing scripts in sample and the CEAW corpus.....	156
Table 4.25 Writing task distribution used in sample.....	157
Table 4.26 Gender distribution in sample .....	157
Table 4.27 Major distribution in sample .....	158
Table 4.28 University background in sample .....	158
Table 4.29 Year distribution in sample.....	158
Table 4.30 Grade distribution over tasks .....	159
Table 5.1 Descriptive statistics – Number of paragraphs .....	168
Table 5.2 Trial scale – Paragraphing by level .....	170
Table 5.3 Descriptive statistics – Number of word tokens .....	171
Table 5.4 Trial scale –Fluency by level .....	172
Table 5.5 Descriptive statistics – Error-free T-unit ratio.....	174
Table 5.6 Trial scale – Accuracy by level .....	176
Table 5.7 Descriptive statistics – Number of conjunctive devices per 10 T-units	177
Table 5.8 Inter-rater reliability for topical progression types .....	179
Table 5.9 Descriptive statistics – Proportion of parallel progression (number of T- units whose topic is coded as belonging to parallel progression, divided by (total number of T-units minus one)) .....	181
Table 5.10 Descriptive statistics – Proportion of extended parallel progression (number of T-units whose topic is coded as belonging to extended parallel progression, divided by (total number of T-units minus one)) .....	183
Table 5.11 Descriptive statistics – Proportion of related sequential progression (number of T-units whose topic is coded as belonging to sequential progression, divided by (total number of T-units minus one)) .....	184
Table 5.12 Descriptive statistics – Proportion of SP1 sequential progression (number of T-units whose topic is coded as belonging to SP1 sequential progression, divided by (total number of T-units minus one)) .....	186
Table 5.13 Descriptive statistics – Proportion of SP2 sequential progression (number of T-units whose topic is coded as belonging to SP2 sequential progression, divided by (total number of T-units minus one)) .....	187
Table 5.14 Descriptive statistics – Proportion of SP3 discourse-related sequential progression (number of T-units whose topic is coded as belonging to SP3	

discourse-related sequential progression, divided by (total number of T-units minus one)) .....	188
Table 5.15 Descriptive statistics – Proportion of unrelated topical progression (number of T-units whose topic is coded as belonging to unrelated topical progression, divided by (total number of T-units minus one)) .....	190
Table 5.16 Descriptive statistics – Proportion of extended sequential progression (number of T-units whose topic is coded as belonging to extended sequential progression, divided by (total number of T-units minus one)) .....	191
Table 5.17 Summary table – Proportion of different topical progression types ...	193
Table 5.18 Trial scale – Coherence .....	193
Table 5.19 Inter-rater reliability for argument structural elements .....	195
Table 5.20 Descriptive statistics – Proportion of introduction (number of introduction elements divided by total number of argumentative structural elements) .....	197
Table 5.21 Descriptive statistics – Proportion of level-1 reasons (number of level-1 reason elements, divided by total number of argumentative structural elements) .....	199
Table 5.22 Descriptive statistics – Proportion of level-2 reasons and below (number of level-2 reason elements, divided by total number of argumentative structural elements) .....	200
Table 5.23 Descriptive statistics – Proportion of standpoints (number of standpoint elements, divided by total number of argumentative structural elements) ...	202
Table 5.24 Descriptive statistics – Proportion of functional markers (number of functional marker elements, divided by total number of argumentative structural elements) .....	203
Table 5.25 Descriptive statistics – Proportion of yourside arguments (number of yourside argument elements, divided by total number of argumentative structural elements) .....	204
Table 5.26 Descriptive statistics – Proportion of myside arguments (number of myside argument elements, divided by total number of argumentative structural elements) .....	207
Table 5.27 Descriptive statistics – Proportion of non-functional elements (number of non-functional elements, divided by total number of argumentative structural elements) .....	209
Table 5.28 Descriptive statistics – Proportion of conclusion (number of conclusion elements, divided by total number of argumentative structural elements) ...	211
Table 5.29 Summary table – The proportion of different argument structural component .....	213

Table 5.30 Trial scale – Argumentation .....	214
Table 6.1 Gender distribution .....	220
Table 6.2 Major distribution .....	221
Table 6.3 Year distribution .....	221
Table 6.4 University background .....	221
Table 6.5 Raters’ background .....	222
Table 6.6 Questionnaire questions .....	225
Table 6.7 Ratings results of a good writing script (ID 184).....	228
Table 6.8 Rating results of a poor writing script (ID517).....	229
Table 7.1 Descriptive statistics – Scale of Mechanics .....	235
Table 7.2 Descriptive statistics – Scale of Fluency .....	236
Table 7.3 Descriptive statistics – Scale of Accuracy .....	236
Table 7.4 Descriptive statistics – Scale of Coherence .....	237
Table 7.5 Descriptive statistics – Scale of Argumentation .....	238
Table 8.1 Discourse analytic measures included in the rating scale .....	263
Table 8.2 Discourse analytic measures not included in the rating scale .....	264
Table 8.3 Extended scale classification by including the theory-based and data- driven multiple trait scale .....	268
Table 8.4 A communicative model of argumentative writing ability.....	269

## List of Figures

Figure 2.1 Holistic scoring rubric for TOEFL CBT writing prompts – Band 6 (Lee, Gentile & Kantor, 2008) .....	10
Figure 2.2 ESL Composition Profile: content (Jacobs et al., 1981) .....	12
Figure 2.3 Primary trait scale (Lloyd-Jones, 1977) .....	14
Figure 2.4 Michigan Writing Assessment Scoring Guide – Level 6 (Hamp-Lyons, 1991) .....	16
Figure 2.5 The Diagnostic English Language Needs Assessment (DELNA) rating scale – Lexical complexity (Knoch, 2009) .....	22
Figure 2.6 The rating scale of the TEM4 writing section (Li, 2010) .....	35
Figure 2.7 The questionnaire on the writing abilities of Chinese EFL learners majoring in English - Accuracy (Li, 2010) .....	36
Figure 3.1 Toulmin's argument elements (Toulmin, 1958/2003, p. 92) .....	54
Figure 4.1 An excerpt writing prompt from the TEM4 writing task .....	102
Figure 4.2 Writing task I – Weibo .....	104
Figure 4.3 Writing task II – Higher education .....	105
Figure 4.4 Writing task III – Air pollution .....	106
Figure 4.5 Sample text for error analysis .....	117
Figure 4.6 Sample text with revisions .....	119
Figure 4.7 Sample text with complex nominals .....	120
Figure 4.8 Sample text with reference .....	124
Figure 4.9 Sample text with conjunctions .....	125
Figure 4.10 Sample text with lexical chains .....	127
Figure 4.11 Sample text with errors of cohesive devices .....	128
Figure 4.12 Parallel progression .....	130
Figure 4.13 Sequential progression I .....	132
Figure 4.14 Sequential progression II .....	133
Figure 4.15 Unrelated topic progression .....	134
Figure 4.16 Extended sequential progression .....	135
Figure 4.17 Sample text with argumentative elements .....	144
Figure 4.18 Sample text with acceptable and unacceptable arguments (correction of errors is presented in round brackets after the error) .....	147
Figure 4.19 Sample text with weak and unacceptable arguments (correction of errors is presented in round brackets after the error) .....	147
Figure 4.20 Three level-1 reasons in convergent arguments operating as three appeals .....	149
Figure 4.21 Level-1 reasons and reasons below in subordination operating as one	

appeal .....	150
Figure 5.1 Distribution of paragraphs over the AWA levels .....	168
Figure 5.2 Distribution of word tokens over the AWA levels .....	171
Figure 5.3 Distribution of error-free T-unit ratio over the AWA levels .....	174
Figure 5.4 Distribution of conjunctive devices over the AWA levels .....	177
Figure 5.5 Distribution of parallel progression over the AWA levels .....	181
Figure 5.6 Distribution of extended parallel progression over the AWA levels....	182
Figure 5.7 Distribution of related sequential progression over the AWA levels...	184
Figure 5.8 Distribution of the SP1 sequential progression over the AWA levels .	185
Figure 5.9 Distribution of the SP2 sequential progression over the AWA levels .	187
Figure 5.10 Distribution of SP3 discourse-related sequential progression over the AWA levels.....	188
Figure 5.11 Distribution of unrelated topical progression over the AWA levels ..	190
Figure 5.12 Distribution of extended sequential progression over the AWA levels .....	191
Figure 5.13 Distribution of introduction.....	196
Figure 5.14 Distribution of level-1 reasons .....	198
Figure 5.15 Sample text at level 1 with a disproportionate introduction.....	199
Figure 5.16 Distribution of level-2 reasons and below.....	200
Figure 5.17 Distribution of standpoints .....	202
Figure 5.18 Distribution of functional markers .....	203
Figure 5.19 Distribution of yourside arguments .....	204
Figure 5.20 Sample text at level 3, with underdeveloped yourside arguments ....	206
Figure 5.21 Distribution of myside arguments .....	207
Figure 5.22 Distribution of non-functional elements.....	208
Figure 5.23 Sample text with non-functional elements .....	210
Figure 5.24 Distribution of conclusion .....	211
Figure 5.25 Sample text with weak reasoning due to poor language use.....	214
Figure 8.1 Feedback on argumentation for test takers.....	272

## List of Abbreviations

Abbreviation	Meaning
ACTFL	American Council on the Teaching of Foreign Languages
ANOVA	Analysis of Variance
AWA	Argumentative Writing Ability
ASLPR	Australian Second Language Proficiency Ratings
AWL	Academic Word List
BNC	British National Corpus
BNF	Backus-Nauer form
CEM	Corpus for English Majors
CEAW	Chinese EFL argumentative writing
CLA	Communicative Language Ability
DELNA	Diagnostic English Language Needs Assessment
EBBs	Empirically derived, binary-choice, boundary definition scales
EFL	English as a Foreign Language
ESP	Extended sequential progression
ESL	English as a Second Language
FSI	Foreign Service Institute
ILR	Interagency Language Roundtable
L1	Primary language (s)
NAEP	National Assessment of Educational Progress
PP	Parallel progression
SP1	Sequential progression I
SP2	Sequential progression II
SP3	Sequential progression III
SPSS	Statistical Package for the Social Sciences
TEM 4	Test for English Majors – Band 4
TEM 8	Test for English Majors – Band 8
TOEFL	Test of English as a Foreign Language need to be changed in body



TOEFL-CBT	TOEFL computer-based test
TSA	Topical structure analysis
UTP	Unrelated topic progression
AWL	Academic Word List

---

## **Chapter 1 Introduction**

### **1.1 Background to the study**

The ability to write an argumentative essay is an essential skill for college EFL learners worldwide. According to the English Syllabus for English Majors of Higher Education (Foreign Language Teaching Advisory Committee of Higher Education English Group, 2000, p. 28), argumentation is one type of writing that Chinese college English majors need to master, alongside narration, description, and exposition. The task-based assessment of argumentative writing ability has therefore been one of the foci of the large-scale and classroom-based English as a Foreign Language (EFL) or English as a Second Language (ESL) assessment in China.

The Test for English Majors-Band 4 (TEM4) writing subtest is a criterion-referenced test, which is designed to assess the ability of writing notes, exposition and argumentation of Chinese EFL college students majoring in English at foundation stage (Revision Group of Syllabus for Test for English Majors-Band 4 of Higher Education, 2004). The writing subtest at band 4 incorporates two writing tasks: one for note-writing and the other for short essay writing. A typical TEM4 short essay task requires test-takers to respond to a topic and an outline with a 200 or so word short expository or argumentative essay. In the recent statement of adjustments of the items of Test for English Majors-Band 4 (TEM4) (Wang, 2016), TEM4 designers abandoned the note-taking task and replaced the topic and outline with a table, a diagram and reading material.

In addition to the large-scale high-stakes TEM4 test, assessing argumentative writing ability also takes place during the course. Classroom-based assessments mainly include mid-term tests, after-class assignments and in-class exercises. Writing tasks in classroom assessments are often adopted or adapted from those administered in the TEM4 test. In classroom-based assessments, students' responses to argumentative writing tasks are

collected and rated by teachers. During the teachers' rating, the TEM4 rating scale is sometimes used or no rating scale is used at all. Feedback on the strengths and weaknesses is provided.

The focus of this study is the rating scale for the TEM4 writing subtest. The existing TEM4 rating scale is an analytic rating scale – an upgraded version of the first-generation five-level holistic rating scale (see Figure 2.6). It is aimed at improving the reliability and validity of the first-generation holistic scale developed for the TEM4 writing subtest in 1997 (Li, 2010). It includes eight aspects and three dimensions. The eight aspects are: idea content, effectiveness, rhetorical organization, correctness, richness, appropriateness, fluency, and mechanics/orthographic conventions. Idea content, effectiveness, and rhetorical organization fall into the general dimension of ideas and arguments, while correctness, richness, appropriateness, and fluency fall into the dimension of language use. Mechanics forms a third dimension (Li, 2010, p.111-112). Each dimension is provided with a number of positive statements to describe expected responses, a numerical scale and a verbal scale, e.g., '1-----2-----3-----4-----5-----6-----7', and 'poor--good---very good---excellent', to indicate the difference. The numerical scales not only reflect the levels of writing performance but also the scores that are assigned to writing performance. Benchmark samples of typical performance at each level of the scale are provided. Raters are required to score essays from three dimensions using three analytic scales and report the total score for that text.

## **1.2 Statement of the problem**

There seems to be a dilemma over how to use the TEM4 rating scale. On the one hand, the large-scale TEM4 test requires the rating scale to be efficient in rating, both time-wise and cost-wise. On the other hand, there is always a need for the scale to provide more diagnostic information to writing teachers and learners in the classroom context. Since the purpose of the TEM4 test is to assess the teaching and learning of skills and

knowledge of English as a foreign language by English majors (or students in other English programs), writing teachers and learners expect to be able to use the scale in classroom-based assessment. Although using or adapting a rating scale from a large-scale test for classroom-based assessment is not a valid practice, it is still common at present (Weigle, 2002). Furthermore, the low stakes of classroom-based assessment also encourages this practice. Rating scales developed in the classroom assessment context have long been neglected in the Chinese EFL context, and there is a lack of systematic and scientific research. However, classroom-based assessment has begun to attract attention from researchers (e.g., Turner and Upshur, 2002; Fulcher, Davidson and Kemp, 2011; Fulcher, 2017). They argue that more academic effort goes into developing a task or program-specific rating scale.

The current TEM4 rating scale is not useful for assessing Chinese EFL writers' argumentative ability in the classroom context. Firstly, according to the test specifications for TEM4 writing subtests, writing tasks are developed for eliciting exposition and argumentation (Zou, 1997, p. 137, p. 166). The TEM4 rating scale is therefore developed for the rating of two types of writing tasks. Rating scales developed for more than one type of writing are usually not sensitive to features that are unique to each type of writing. Therefore, such rating scales have problems in helping raters to identify useful information specific to each kind of writing.

Secondly, the empirically-developed TEM4 rating scale in use, although proving more useful than the first-generation holistic scale, is inefficient in providing adequate diagnostic information. Firstly, there is no separate scale for each aspect of writing. When a rater assigns a point of five using the ideas and arguments scale, a number of interpretations may occur as it may refer to a good idea and content but poor rhetorical organization, or a good rhetorical organization but poor idea and content. This problem has also been mentioned by raters at the validation stage (Li, 2010). Therefore, even a

score obtained from the analytic scale is subject to multiple interpretations. Secondly, the use of ‘poor’, ‘fair’, ‘very good’, and ‘excellent’ may be convenient to differentiate between levels of performance in large-scale tests where a single score is needed, but these qualifiers are unable to provide a more specific description of students’ performance than a score. The problems with the existing rating scale described above might also affect the raters’ ability to make fine-grained distinctions between different aspects on a rating scale, because an average performance on the scale of ideas and arguments may create an image in one rater’s mind different from that in another’s.

### **1.3 Purpose and nature of the study**

The purpose of this study is therefore to develop a rating scale specific to Chinese college EFL learners’ argumentative writing in a classroom-assessment context, using a theoretically-based data-driven approach to the scale design, and evaluate its usability by Chinese EFL writing teachers. The theoretical approach provides a theoretical model on the basis of which categories are selected. The data-driven approach involves the analysis of written texts using discourse analytic measures, and the development of scale descriptors based on the statistical analysis of these measures. The hybrid approach which combines the theoretical approach and the data-driven approach will provide more detailed band descriptors to different aspects of writing than the current empirically-developed rating scale for classroom-based assessment.

The study comprises two phases: a development phase and an evaluation phase. In the development phase, a pilot study and a main study were conducted. In the pilot study, 15 writing scripts were analyzed using discourse measures selected from the literature. In the main study, 258 writing scripts at different proficiency levels were analyzed using discourse measures derived from the pilot data. Statistical analysis was conducted to investigate the distribution of the discourse measures at different proficiency levels and differences between the proficiency levels. Detailed descriptors and the number of levels

were then formulated based on the results of the statistical analysis. In the evaluation phase, the inter-rater reliability of ratings produced using the new rating scale was investigated, and raters' perceptions on the usability of the new rating scale were elicited and analyzed using content analysis.

#### **1.4 Research questions**

The two research questions addressed in the study are:

- 1) Which discourse analytical measures are successful in distinguishing between argumentative writing samples from Chinese EFL college students majoring in English at different proficiency levels?
- 2) Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing?

#### **1.5 Outline of the thesis**

The thesis is comprised of eight chapters. Chapter 1 has introduced the research problem, the purpose of the study and an overview of the research. Chapter 2 first reviews the literature on the rating scales, including definition, types, important features and design processes of rating scales. This chapter then provides a brief overview of the development of the current TEM4 rating scale. Through reviews of constructs of writing ability and their related discourse analytic measures in the current literature, Chapter 3 develops a theoretical model of argumentative writing ability and an operational model of its measures. Chapters 4 to 5 describe the methodology, results and discussion of the development phase of the study – the development of the rating scale. Chapter 4 provides an outline of the methodology used, and a detailed description of methods used in the pilot study and the main study. Chapter 5 presents the results of statistical analysis in the main study, and the main findings and discussion in relation to previous research; rating scales are developed at the end of Chapter 5. Chapters 6 and 7 describe the methodology, results and discussion of the second phase – the usability of the new rating scale. Chapter

6 presents a detailed description of methods used in the inter-rater reliability study and the study of raters' perceptions of the new rating scale. Chapter 7 discusses the findings of the second phase under Bachman and Palmer's (1996) framework of test usefulness. At the end of Chapter 7, a usability argument of the new rating scale is given. Chapter 8, the concluding chapter, summarizes the study and discusses the implications of the study at both theoretical and practical levels. Suggestions for further research are offered and limitations of the study are identified.

## **Chapter 2     Rating Scales**

This chapter first reviews the definition, types, important features, and design methods of rating scales in ESL/EFL writing assessments. It aims to provide insights into the development of a rating scale for writing assessment. This chapter also gives a detailed introduction to the current TEM4 rating scale, mainly including its development process, strengths and issues. This chapter provides the rationale for the development of the new rating scale and the significance of using a development method that has never been used in the Chinese EFL context to address the issues that are known to exist.

### **2.1     Definition of a rating scale**

A well-recognized definition of a rating scale (also known as scoring rubric or proficiency scale) is given by Davies et al. (1999, p. 153) as:

A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged. Like a test, a proficiency (rating) scale provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterized in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency, and cohesion). Proficiency scales typically consist of subscales for the skills of speaking, reading, writing and listening... Scales are descriptions of groups of typically occurring behaviors; they are not in themselves test instruments and need to be used in conjunction with tests appropriate to the test population and test purpose. Raters or judges are normally trained in the use of proficiency scales so as to ensure the measure's reliability.



## 2.2 Types of rating scales

There are four major types of rating scales in language assessment: holistic, analytic primary trait, and multiple trait scales. For differing purposes of research, the types of rating scales may vary (Weigle, 2002; North, 2003; Fulcher, 2003). For example, Weigle (2002) categorizes multiple trait scales as analytic scales because the two scales differ more in procedure for developing and using the scales but remain the same in the description of the rating scales themselves. For the purpose of the present study, the distinctions between multiple trait scales and analytic scales are recognized. The four different types of scales are related to each other in terms of four criteria: how many scales are used or scores are reported; whether it can be generalized to a variety of tasks or a specific task or type of tasks; whether it is performance-based or ability-based; and the use in ESL/EFL contexts (see Table 2.1). Although distinctions are found between different scale types, there is not always a clear-cut boundary between them (Weigle, 2002; North, 2003).

Table 2.1 Four types of rating scales (based on Weigle, 2002, p. 109; North, 2003, p. 64)

Scales	Primary trait	Multiple trait	Holistic	Analytic
Score or scale	single score	multiple scores	single score	multiple scores
Task generalization	Specific to a particular task or a type of tasks		Generalized to a variety of task types	
Generic or context-specific	Context-specific		Generic	
ESL/EFL setting	Not commonly used		Commonly Used	

As shown in Table 2.1, only a single score is given to a writing script assessed with a primary trait scale or a holistic scale, while multiple scores are given to a writing script assessed with a multiple trait scale or an analytic scale. As for task generalization, the

primary trait and multiple trait scales are intended to be specific to a task or a type of tasks, while the holistic scale and the analytic scale are intended to be generalized to a variety of task types. North (2003, p. 63–4) and Fulcher (2003, p. 90) ascribe the distinction between the context/task-specific (primary trait and multiple trait) scales and generic scales (holistic and analytic scales) to two opposing underlying views of assessment. On the one hand, it is believed that different tasks elicit performance of different types and quality, therefore a separate rating scale for each task or type of task is needed. On the other hand, it is believed that the rating scale focuses on qualities in the performance which reflect the competence underlying it; therefore, the construct to be assessed may be less distorted by test elicitation methods and therefore the rating scale is generalizable to different tasks. Primary trait and multiple trait scales are commonly used in L1 settings while holistic and analytic scales are commonly used in ESL/EFL settings.

In the remaining sections, each type of scale is introduced in detail and their merits and limitations are discussed in relation to the context of this study.

### **2.2.1 Holistic scales**

Holistic scoring requires raters to respond to writing as a whole and assign a single, holistic score to the writing. A holistic scale is usually comprised of descriptions of several features of writing scripts at different levels. A well-known holistic rating scale is the Test of English as a Foreign Language (TOEFL) writing scoring guide. Part of the holistic scoring rubric for the TOEFL computer-based test (CBT) of writing is presented in Figure 2.1. As can be seen from the figure, the scale contains several features and descriptions of writing scripts typical for level 6. These features include rhetorical organization, syntactic feature, language use, word choice, appropriateness, and effectiveness of task fulfillment.

6 An essay at this level effectively addresses the writing task, is well organized and well developed, uses clearly appropriate details to support a thesis or illustrate ideas, displays consistent facility in the use of language, demonstrates syntactic variety and appropriate word choice, though it may have occasional errors
--

Figure 2.1 Holistic scoring rubric for TOEFL CBT writing prompts – Band 6  
(Lee, Gentile & Kantor, 2008)

Holistic scoring is regarded as authentic, for the reason that the process of rating holistically resembles a real reading process, which emphasizes reaction on the basis of the global impression of a reader to a writing script instead of attention to each criterion or component (White, 1993). Holistic scoring is commonly regarded as efficient, and thus more appropriate for large-scale assessment where a quick holistic score is needed. Holistic scales are occasionally used in conjunction with benchmark essays when holistic scales do not provide enough detailed information. However, holistic scales also have disadvantages. Weigle (2002, p. 114) gives a very useful summary of the disadvantages of holistic scales. They are: First, a holistic score does not give information on test performance for different aspects of writing, therefore little can be known about the quality of each aspect of writing. Second, a holistic score is hard to interpret. Raters do not always use the same criteria to arrive at the same score, thus lending the score to multiple interpretations. Third, holistic scoring is unable to reflect the uneven profiles of language learners. Research shows that learners do not develop in all aspects of language proficiency in the same way at roughly the same rate. Roid (1994, cited in North, 2003, p. 71) found that 60% of students showed an uneven profile in research comparing holistic scores to analytic scores on a six-trait scale for first language writing. Fulcher (1993) demonstrates that there is a regression in the accuracy aspect of ESL speaking proficiency as learners at higher elementary/lower intermediate levels struggle to express more complex meanings. Fourth, holistic scores have also been shown to correlate with relatively superficial characteristics such as length of essay/writing and handwriting.

### 2.2.2 Analytic scales

Analytic scoring requires raters to attend to each category of writing performance (e.g., vocabulary, coherence, grammar) and assign a score to each one separately. Accordingly, in an analytic scale, a uniform scale or a separate scale is attached to an individual category or trait or aspect of writing performance. A single score, aggregated from separate scores, is reported or, when needed, separate scores are also reported. These separate scores are weighted equally or unequally in the aggregation. A widely used analytic scale is the ESL Composition Profile by Jacobs et al. (1981). In the Jacobs et al. scale, scripts are rated on five traits of writing: content, organization, vocabulary, language use and mechanics. An implicit four-level scale is used for assessing each trait. Each trait is differentially weighted: content (30 points), organization (20 points), vocabulary (20 points), language use (25 points) and mechanics (5 points). Part of the Jacob et al. scale is presented in Figure 2.2. For the trait of content, scripts which are judged to be ‘knowledgeable, substantive, containing thorough development of thesis and relevant to assigned topic’ are scored in the range of 27 to 30. ‘EXCELLENT TO VERY GOOD’ is included as part of the scoring criteria. The four-level scale is not explicitly stated but can be indicated by a score range, e.g., 27–30, or an overall quality descriptor, e.g., ‘EXCELLENT TO VERY GOOD’.

ESL COMPOSITION PROFILE		
STUDENT	DATE	TOPIC
SCORE	LEVEL	CRITERIA
COMMENTS		
Content	30-27	EXCELLENT TO VERY GOOD: Knowledgeable • substantive • thorough development of thesis • relevant to assigned topic
	26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • most relevant to topic, but lacks detail

	21-17	FAIR TO POOR: •limited range; •frequent errors of word/idiom form, choice, usage; • meaning confused or obscured
	16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate

Figure 2.2 ESL Composition Profile: content (Jacobs et al., 1981)

Three major advantages of the analytic scales are recognized by researchers (e.g., Weigle, 2002, p. 120; North, 2003, p.71–2). They are (1) Analytic scales highlight the uneven profiles of learners (mentioned in the holistic scale review – see Section 2.2.1). (2) Analytic scales are useful in rater training and rating. With explicit and balanced criteria, inexperienced raters more readily get accustomed to new rating scales, and all raters use the same criteria. (3) Analytic scales provide more detailed information on test takers’ performance in different aspects of writing, thus providing more diagnostic information to teachers and learners.

However, rating each category separately takes longer than holistic rating. If separate scores are combined to form a composite score, the information from each scale is lost. Experienced raters who are familiar with a particular analytic scale may assign a combined single score first and then adjust their analytic scores accordingly (Weigle, 2002, p. 120).

### 2.2.3 Primary trait scales

Primary trait scoring is mostly associated with Lloyd-Jones’ (1977) work for the National Assessment of Educational Progress (NAEP). The idea underlying primary trait scoring is that students respond to a writing task that elicits a narrowly defined type of discourse and writing samples are rated by reference to defining features of that particular discourse. A typical primary trait scoring guide includes (a) a writing task, (b) a statement of the primary rhetorical trait of that type of discourse, (c) an interpretation of the task

hypothesizing the expected response to the task, (d) an interpretation of the relationship between the task and the primary trait, (e) a scoring guide, (f) sample scripts, and (g) explanations of why scores are assigned to sample scripts. A primary trait scoring guide can include several salient categories (or traits) on which each script is to be judged. Part of the scoring guide for primary trait scoring, developed by Lloyd-Jones (1977), is reproduced in Figure 2.3.

NAEP Scoring Guide: Children on Boat	
<b>Background</b>	<i>Primary trait.</i> Imaginative Expression of Feeling through Inventive Elaboration of a Point of View
<b>Final Scoring Guide</b>	
ENTIRE EXERCISE	
0 No response, sentence fragment	
1 Scorable	
2 Illegible or illiterate	
3 Does not refer to the picture at all	
9 I don't know	
USE OF DIALOGUE	
0 Does not use dialogue in the story.	
1 Direct quote from one person in the story. The one person may talk more than once. When in doubt whether two statements are made by the same person or different people, code 1. A direct quote of a thought also counts. Can be in hypothetical tense.	
2 Direct quote from two or more persons in the story.	
POINT OF VIEW	
0 Point of view cannot be determined or does not control point of view.	
1 Point of view is consistently one of the five children. Include "If I were one of the children..." and recalling participation as one of the children.	
2 Point of view is consistently one of an observer. When an observer joins the children in the play, the point of view is still "2" because the observer makes a sixth person playing. Include papers with minimal evidence even when difficult to tell which point of view is being taken.	

TENSE	
0	Cannot determine time, or does not control tense. (One wrong tense places the paper in this category, except drowned in the present.)
1	Present tense-past tense may also be present if not part of the “mainline” of the story.
2	Past tense-If a past tense description is acceptable brought up to present, code as “past”. Sometimes the present is used to create a frame for past events. Code this as past, since the actual description is in the past.
3	Hypothetical time-Papers written entirely in the “If I were on the boat” or “If I were there, I would.” These papers often include future references such as “when I get on the boat I will” If the part is hypothetical and rest past or present and tense is controlled, code present or past. If the introduction, up to two sentences, is only part in past or present then code hypothetical.

Figure 2.3 Primary trait scale (Lloyd-Jones, 1977)

The writing task associated with the scoring guide in Figure 2.3 requires test takers to look at a printed photograph of five children playing on an overturned rowboat and write about that event to a good friend imagining they are one of the children, in a way that expresses strong feelings. In the scoring guide, the primary trait to be assessed is stated; four defining traits are decided: entire exercise, use of dialogue, point of view, and tense of the imaginative expression. Under each trait, descriptions of expected writing responses and codes are listed.

Primary trait scales provide detailed and specific scoring criteria. Therefore, they have the potential to provide rich information about students’ abilities if enough samples are collected from students (Weigle, 2002). However, primary trait scales are not commonly used in second or foreign language assessment. One of the reasons is probably that it is very time-consuming and expensive to develop. It takes about 60 to 80 hours of expertise per task to be developed. Another reason could be that there is a lack of research to support the assertion that raters can focus on one trait during the whole rating process (Hamp-Lyons, 1991). Furthermore, it depends heavily on scale developers’ expertise to make decisions about the primary trait of a certain discourse and features of performance related

to that trait. Such expertise is not as common in second language assessment as in L1 assessment.

#### **2.2.4 Multiple trait scales**

Multiple trait scales were developed by Hamp-Lyons (1991) for the assessment of writing ability of entrants of the University of Michigan. They were introduced as a contrast to the primary trait scale and the analytic scale. They include several scales and these scales are used to rate different traits that are selected as salient (i.e. rather than including every element of writing ability that may be manifested in the context) in describing test performance in a certain context (e.g., second language assessment). Unlike some analytic scales, scores on each category in multiple trait scales are reported and never aggregated into a single score. Unlike the primary trait scale, the scale is not specific to a particular writing task but to a range of writing tasks which are defined by the same design criteria and the same context (e.g., the same testing population). In designing a rating scale for the Michigan English Language Assessment Battery (Hamp-Lyons, 1991), the researcher collected two groups of university faculty responses (from writing teachers and teachers of other disciplines) to students' test texts. She analyzed these responses and firstly decided on six categories which would cover all the major features in the two groups of responses: content, argument, text structure features (cohesion), evidence of planning (coherence), language control and planning/organization. She then combined 'content' and 'argument' into a trait cluster called 'ideas and argument', combined 'planning/organization', 'coherence' and 'cohesion' into 'rhetorical features', and kept 'language control' as a separate category, aiming to incorporate all criteria and responses into more general terms. North (2003) argues that the empirical and contextualized development methodology of the multiple trait scale started with Smith and Kendall (1963, cited in North, 2003). The predecessors of the multiple trait scale innovatively invited scale users to select categories which were salient and could be consistently interpreted by the scale users, and the most reliable descriptors for these categories were



then scaled. Part of the Michigan writing assessment scoring guide is presented in Figure 2.4. There are three trait scales: ‘Ideas and arguments’, ‘Rhetorical features’, and ‘Language control’. Writing scripts are assessed at six levels. The figure presents attributes of scripts at level 6 in terms of three major traits.

	Ideas and arguments	Rhetorical features	Language control
6	The essay deals with the issues centrally and fully. The position is clear, and strongly and substantially argued. The complexity of the issues is treated seriously and the viewpoints of other people are taken into account very well.	The essay has rhetorical control at the highest level, showing unity and subtle management. Ideas are balanced with support and the whole essay shows strong control of organization appropriate to the content. Textual elements are well connected through logical or linguistic transitions and there is no repetition or redundancy.	The essay has excellent language control with the elegance of diction and style. Grammatical structures and vocabulary are well chosen to express the ideas and to carry out the intentions.

Figure 2.4 Michigan Writing Assessment Scoring Guide – Level 6 (Hamp-Lyons, 1991)

The advantages of the multiple trait scale lie in the way it is different from other scales (Hamp-Lyons, 1991, p. 249–55). First is the salience of test performance on a distinct type of writing prompts, rather than every trivial element of test performance. The second is that raters are more readily able to negotiate scores over their rating decisions through the shared language of criteria or standards that originated from the development process. The third is more diagnostic information as scores are not combined in a single score and scale is based on the investigation of real work sample scripts in context. The fourth is the validity – the scale development involves careful observation of a context through prompt specification and the shaping of the scale to fit with those observations, thus having content validity (i.e., whether a test is relevant to and contains a representative sample of a given area of content or ability) and construct validity (i.e., whether it can be

inferred from observations in the test (i.e., test performance) what a test purports to measure). That is, the scale development process reflect the content and construct validity of a test. The fifth is the positive effect on teaching of the detailed information that multiple trait scales provide.

One of the disadvantages of the multiple trait scale is that the categorization between different aspects may not be balanced, as the scale is not theory-based. For example, mechanics are not included in the scale as they are not regarded as salient within the context. Another disadvantage is that it can only be used with writing tasks that meet the initial design criteria for which the scale is developed.

### **2.2.5 Implications for the new rating scale**

A review of four major types of rating scales of language performance seems to show that analytic scales and multiple trait scales are more appropriate to the assessment of language performance in the EFL context than primary trait scales and holistic scales. The reasons are (1) They both emphasize the uneven profile of EFL learners. (2) They provide separate scoring criteria on each aspect of the writing abilities, thus potentially being more diagnostic and useful in classroom-based assessment for Chinese EFL teachers and learners. (3) They are more convenient in training raters. With separate trait scales, raters with little rating experience can understand the scoring criteria more easily (e.g., writing teachers and students). Despite these common advantages, the two scales differ in score reporting and the way traits are selected. If these differences are taken into account, analytic scales are more appropriate to the context of this study. First, a combined single score and separate scores are both needed in this study. A combined single score is needed on occasions where students' performance as a whole is considered, for example, mid-term tests, final exams. With multiple trait scales, separate scores on different categories are never combined into a single score. Second, the selection of categories or traits is more comprehensive or generic in analytic scales, while only salient categories are

included in multiple trait scales. In this study, it cannot be decided at the initial stage of development what kind of characteristics of the Chinese college level EFL learners' writing abilities is salient. Although analytic scales are time-consuming in use, the consumption of time can be balanced by the detailed diagnostic information they can provide.

### **2.3 Important features of rating scales**

To develop a rating scale, developers usually need to consider a number of important features: functions, categories/traits, distinguishing being levels, descriptor formulations, number of bands or scoring levels to be used, weighting of different categories/traits and score reporting (Weigle, 2002; North, 2003). Each of these features is considered step by step.

#### **2.3.1 Functions**

Alderson (1991) classifies three functions of rating scales: user-oriented, assessor-oriented, and constructor-oriented. User-oriented rating scales are concerned with the reporting function of a scale when test users, like employers or admissions officers, try to understand what a candidate can do. Assessor-oriented rating scales represent the guiding function in a rating process to ensure the reliability and validity of the subjective judgments involved. Constructor-oriented rating scales emphasize the guiding function in constructing tests at appropriate levels. Alderson (ibid) points out that problems will arise when scales are devised for one purpose but used for another. For example, a user-oriented rating scale may include a score report which describes detailed implications of each band score. An assessor-oriented rating scale must provide reliable rating standards through detailed descriptors. A constructor-oriented rating scale may commonly include a description of a class of tasks which represent certain levels of proficiency for language learners. In the present study, the rating scale to be developed is an assessor-oriented scale, because it is aimed at being used by writing teachers and raters to rate the argumentative

writing performance of Chinese EFL college students.

### **2.3.2 Categories/traits**

North (2003, p. 22–3) summarizes three types of categories: pre-communicative and generic, performance-based, and communicative ability-based and generic. The pre-communicative and generic categories are those adapted from 1960s' elements: grammar, vocabulary and phonology. The main disadvantages of the pre-communicative categories are that grammar may be interpreted in terms of counting mistakes made and the focus is purely on the psycholinguistic process while no concept of communicative meaning is mentioned (North, 2003, p. 23). Performance-based categories are focused on defining aspects of quality of performance, for example, range (range of syntactic structure as well as range of vocabulary). The development of performance-based categories tends to involve detailed investigation by scale developers and discussion of the performance involved, therefore performance-based categories tend to be interpreted more consistently by different raters, thus giving a high reliability (North, 2003, p.26). However, performance-based categories also have disadvantages: some categories are arbitrarily selected, and are not related to theoretical models; some categories are related to test method facts, thus having generalizability issues. Communicative ability-based and generic scales base all categories directly on a model of communicative ability. Test results are therefore expected to show less variation across different kinds of tasks and so generalize to other contexts. However, North (2003, p. 32) points out two typical issues: certain aspects are thought of as parameters of a theoretical model, but it does not necessarily mean they can be isolated as observable components and hence rated or tested separately. The model which has been elaborated is a model of underlying competence and not a model of performance. This means that it has difficulty in coping with what happens when competence is put to use. For example, there is no place for fluency in communicative ability-based scales.

### 2.3.3 Distinguishing between levels

North (2003, p. 42) summarizes three ways to distinguish between levels in a scale:

- 1) Graphic and numerical scales: giving labels to each end of the relevant dimension with a line or row of numbers to represent the dimension
- 2) Labelled scales: adding labels to various points or bands along each dimension
- 3) Defined scales: turning the scale round 90 degrees to give room for longer descriptors.

The graphic and numerical scales were the first attempt to associate different kinds of behavior with different parts of a continuum represented by a scale (Paterson, 1922, cited in North, 2003). The technique of labelling each end of the scale with slots or intervals in the middle is known as the “semantic differential” by Osgood and Tannenbaum (1957, cited in North, 2003) and adopted by some scales (e.g., the rating scale used for the Foreign Service Institute (FSI) system). However, North (2003) criticizes such scales for not being able to specify the behaviors at each level, therefore, causing difficulty in understanding of the criteria for each slot or interval by raters.

Labelled scales articulate the differences between bands through semantic cues (e.g., *poor*, *below average*, *average*, *above average*, *excellent*). Labelled scales can be ambiguous when raters need to mark a point and points and cues are not in a one-on-one relation (North, 2003).

Defined scales add length to labels for each band through detailed descriptors. For a defined scale to be valid, descriptors should be tangible and definite (Thorndike, 1912, cited in North, 2003), and should not be dependent on abstract adjectives (e.g., generally, sometimes) (Champney, 1941, cited in North, 2003).

It seems that the new scale to be developed in this study would be more useful if it is in

the form of defined scales rather than graphic and numerical scales, or labelled scales because the classroom-based assessment needs more detailed descriptors for the production of feedback to teachers and students.

#### **2.3.4 Descriptor formulation**

North (2003, p. 48–59) summarizes three basic ways of descriptor formulation. The first type is abstract formulation. Scales of this type define each band through the degree of presence or absence of the feature concerned, using a continuum of qualifiers and quantifiers (e.g., *almost always*, *generally*, *somewhat*, *generally not*) to distinguish between different levels. These scales are direct development from graphic and labelled rating scales discussed above. The main disadvantage is that descriptors using abstract qualifiers are criticized as imprecise and ambiguous (Brindley 1991; Davidson, 1991; Alderson, 1991), thus giving raters difficulty in understanding the meaning of descriptors for each band. North (2003) argues that the disadvantage of abstract descriptors can be improved through piloting with raters or use of benchmark performance.

The second type is concrete formulation. Scales of such a type define each band with a description of salient features of actual work samples. North (2003, p. 51) argues that concrete descriptors are more likely to be interpreted consistently by raters or users because the ‘focus is on what is salient’ and ‘there is no explicit attempt to create a semantic continuum on which “each descriptor shares phraseology with those above and those below, and the endpoint descriptors focus on presence and absence in a complete sense” (Davidson, 1992).’ However, the major problem with the concrete approach is that little is reported on the decision on assigning particular tasks or features to particular bands (North, 2003). This problem is detailed in Section 2.4.1 on the criticism of terminology of intuitively developed scales.

The third type is objective formulation. Scales of this type define bands through counting

the presence or absence of features. An example scale is presented in Figure 2.5. In this figure, lexical complexity is measured by the number of words in the Academic Word List (AWL) identified in the writing scripts to be assessed.

Band 8	Large number of words from academic wordlist (more than 20)
Band 7	Between 12 and 20 AWL words
Band 6	5-12 words from AWL
Band 5	Less than 5 words from AWL

Figure 2.5 The Diagnostic English Language Needs Assessment (DELNA) rating scale – Lexical complexity (Knoch, 2009)

Scales with objectively formulated descriptors have the advantage of enabling raters to make more objective judgments through countable measures, thus reducing unreliable judgments (Knoch, 2009). The disadvantage of this type of scale is that it is time-consuming, because rating each feature through counting requires the text to be read multiple times (North, 2003), though it can be read automatically if it is in electronic form.

As discussed above, it seems that both concrete and objective formulations are preferred in a rating scale, rather than an abstract formulation, as the former two have a considerable effectiveness on their own compared to the latter (e.g., inter-rater reliability).

### **2.3.5 Number of bands**

Weigle (2002, p. 123) summarizes three factors influencing the number of bands or levels included in a rating scale. The most important one is the range of performance that can be expected of the population of test takers. Normally, one wants to have enough bands to show progress. The second factor is the purpose of the test. If the test is used to make pass/fail decision, fewer bands are needed; if the test is used to place students in different courses, more bands are needed. The third factor is the capability of raters to make

distinctions between bands. North (2003, p. 18) cautions that the number of bands should not exceed the number that raters are capable of making reasonably consistent distinctions between. Pollitt (1991, cited in North, 2003, p. 18) argues that the number of bands can be derived from the inter-rater reliability (i.e., the consistency in distinguishing between bands); for example, an inter-rater reliability of .08–.09 justifies the use of between 4 and 6 bands. Thus, he suggests that 5 bands assumes a well-developed scale and well-trained raters.

In addition to Weigle's (2002) three factors, North (2003, p. 19) points out another issue, that not all bands are appropriate for all categories. That is, scale developers may have difficulty in defining descriptors for each level of a certain category (e.g., pronunciation), or raters may have difficulty in distinguishing that number of levels even when each level is defined. For example, raters tended to adopt a play-it-safe method and mainly award central level scores when they had problems understanding scale descriptors.

With all of the considerations taken into account, it seems that the number of bands may not be satisfactorily decided on. It remains an empirical question whether the number of bands can show the progress of learners sufficiently, or can be reasonably consistently distinguished between by raters. Researchers suggest trialing the scale over a variety of scripts and raters and seeking their feedback on the appropriate number of bands to be included (North, 2003; Weigle, 2002).

### **2.3.6 Weighting of categories**

There are two types of weighting in the rating scales: one is equally weighted, and the other is unequally weighted. Jacob et al. (1981) put differential weights on various components, with content receiving the most weight and mechanics the least. Hamp-Lyons (1991) suggests an equal weighting of all components. She believes if one component is weighted more heavily than others, a holistic scale is more appropriate.



Weigle (2002, p. 124) proposes two aspects to be considered in the weighting: whether a theory of writing on which a scale is built prescribes certain aspects that are more or less important/relevant/involved than others, and whether statistical factors, such as the amount of variation within each aspect and correlations between aspects, need to be considered.

The issue of weighting of categories is taken into account in the development of the rating scale of argumentative writing of Chinese EFL college students.

### **2.3.7 Score reporting**

Weigle (2002, p. 124) argues that score reporting is dependent on the purpose of testing. She further argues that a composite score which combines separate scale scores is preferred when decision on placement, exit, or exemption needs to be made, while scale scores are reported preferable when a more accurate profile of test takers' abilities, and more useful diagnostic information are needed.

The issue of score reporting is taken into account in the development of the rating scale of argumentative writing of Chinese EFL college students.

## **2.4 Design method of rating scales**

The scale development methods are mainly concerned with how categories or traits, and descriptors for each level are developed. There are four major types of methods: intuitive method, theoretically-based method, empirical method and a hybrid method (Turner and Upshur, 2002; Fulcher, 2003; North, 2003; Montee and Malone, 2014). Each method has its merits and limitations. In this section, each method is introduced, together with its merits and limitations. Their relevance to the context of this study is also discussed and a method that is suitable for the rating scale of argumentative writing ability is selected.

### **2.4.1 Intuitive method**

Fulcher (2003) classifies scales which are developed mainly through expert, committee, and experiential judgments as intuitively developed. In his classification, an expert is an experienced teacher or a language tester. They draft a rating scale based on an already existing scale, a teaching syllabus, or a needs analysis. Feedback on the usefulness of the scale may be collected from scale users. A committee is a small group of experts. Their consensus is used to inform the wording of descriptors and the levels of scales. Once a scale is developed, it may be refined by scale users who accumulate the knowledge of sample performances, in relation to levels of a scale, through rating.

Typical intuitive-developed rating scales are the Foreign Service Institute (FSI) family of rating scales, which are widely used in the development of curriculum and in assessment contexts. The intuitively developed criteria or a priori criteria in the scales in the FSI family is beneficial as it provides a common terminology for testers and curriculum designers (Schultz, 1986), and it can be used in a wide testing situation (Bachman, 1990). Montee and Malone (2014, p. 7) argue that a rating scale developed based on an existing scale widely used and understood by stakeholders can facilitate interpretation and the use of scores when performance on a specific test needs to be compared to an external standard. However, intuitive-developed rating scales are not always well-understood or consistently interpreted by raters and are increasingly criticized. Fulcher (2003, p. 92–7) provides four major criticisms of the design method of the scales in the FSI family. One frequently mentioned criticism is on the use of the native speaker as the top level of a scale. The concept of native speaker as a referencing point in a scale is criticized because native speakers vary considerably in their ability (Bachman & Savignon, 1986; Lantolf and Frawley, 1985; Davies, 1990, cited in Fulcher, 2003). Another criticism is on the vagueness of band descriptors. Band descriptors of these scales are vague and of little practical use for the actual rating process (Fulcher, 1989; Matthews, 1990; Hieke, 1985, cited in Fulcher, 2003). Fulcher (2003) further argues that raters may produce reliable

ratings using vague descriptors through rater training and socialization, but this should not mask the problem of vagueness of band descriptors. Still another criticism is on the progression from zero to native speaker. Pienemann, Johnston, and Brindley (1988) argue that the progression does not reflect language development because the progression is not theoretically coherent and empirically verifiable. Fulcher (2003) further critiques that the progression may reflect unvalidated theories of second language acquisition that correspond to scale designers' intuition and experience. Yet another criticism is on the lack of empirical evidence for scale descriptors. Following Jones (1981) and Alderson (1991), Fulcher (2003) argues that there are few investigations on the definition of terms used in the scale descriptors, and on correspondence between language samples and the scale descriptors, thus influencing the usefulness of the rating system.

#### **2.4.2 Theoretically-based method**

Theoretically-based scales are originally referred to as those derived from theories of language acquisition and are intended to reflect language-learning progression. Pienemann et al. (1988) made the first attempt to develop rating scales for assessing second language attainment. Although their scoring criteria are derived from specific natural interlanguage speech samples (e.g., simple words, formulae at stage 1), these scoring criteria are restricted to a number of pre-selected syntactic and morphological structures. Mislevy (1993, p. 343, cited in North, 2003, p. 11) argues that it is impossible to build a rating scale directly on a learning process unless it reflects 'a simplified description of selected aspects of the infinite varieties of skills and knowledge that characterize real students'. Another theoretically-based rating scale is developed by Bachman and Palmer (1982) on the basis of research on communicative competence by Hymes (1972) and Canale and Swain (1980). The categories or traits of the scale are directly derived from three major components of communicative competence, with each level defined in perceived developmental features (e.g., limited vocabulary (a few words and formulaic phrases), small vocabulary, vocabulary of moderate size, large vocabulary,

extensive vocabulary). A number of researchers argue for the necessity of a theoretical basis for rating scales. Fulcher (1995) argues that the scales based on theoretical models are generic and context-independent, therefore their rating results are more generalizable and transferable across task types and contexts. McNamara (1996, p. 49) argues that ‘an a-theoretical approach to rating scale design in fact provides an inadequate basis for practice’. North (2003, p. 22) also comments that ‘one cannot avoid theory’.

Despite these advantages, theory-based rating scales are also criticized. Turner and Upshur (2002, p. 3) summarize four commonly mentioned criticisms of theory-based rating scales: (a) Scoring criteria are not ordered in line with progression of language development, (b) scoring criteria are often not relevant to test performance (c) scoring criteria are improperly grouped at different proficiency levels, and (d) relative wording is used in descriptors, which may lead to a false profile of a test taker.

### **2.4.3 Empirically-developed method**

Empirically-developed scales attempt to address some of the criticisms that a priori scales, as well as theoretically-based scales, have (Turner & Upshur, 2002). In empirically derived approaches, scoring criteria are developed or selected by test developers by working with sample performances from the test – for example, the data-based/data-driven approach proposed by Fulcher (1993) for developing a fluency scale of oral proficiency, the approach for developing the empirically derived, binary-choice, boundary definition scales (EBBs) (Upshur & Turner, 1995), or scaling descriptors by North (1995, 1996/2000) for the development of a common framework scale for language proficiency. In the data-based/data-driven approach, the scale developer groups the samples into a predetermined number of levels, identifies verbal phenomena (e.g., pauses) related to a trait of oral proficiency (e.g., fluency), classifies explanatory categories (e.g., pauses indicating the end of a turn) to explain the reason for the phenomena, tallies categorized phenomena, conducts statistical analysis on the discrimination of categories

between the different levels, and drafts descriptors based on the statistical results (e.g., non-linear relationship) of the discrimination of explanatory categories. Fulcher's fluency scale has been shown to be stable when used in different speaking tasks and by different raters (North, 2000). Another advantage of this scale is that the content of the scale is based on the description of discourse features identified through analysis of actual speech performance, thus the use of qualifiers to distinguish levels in the scale is lessened (Fulcher, 1993).

EBBs are comprised of a hierarchical set of binary questions (yes/no) on the features of the performance being rated (Upshur and Turner, 1995). Through answering these questions, raters are able to distinguish between levels of language performance. The development of the EBBs requires a group of teachers or a research team to divide a group of learners' performances, which are selected as representative of the full range of ability to be assessed, into better and poorer subgroups. Then they are asked to discuss and identify the attributes with which they make dichotomous decisions. On the basis of these attributes, yes/no questions are formed to distinguish the different levels. This process repeats until the performance is separated into a number of subgroups and a number of questions are formulated. Upshur and Turner (1995) argue that the EBBs are different from the traditional scale in that instead of having the midpoints defined in descriptors, the EBBs describe the boundaries between categories, thus simplifying the estimation process and enhancing measurement accuracy. They also argue that the EBBs are simple to use as instead of having to attend to several features at a time for a descriptor during rating, raters only need to answer a critical question to distinguish between levels of performance each time. They further argue that unlike scales which embody false assumptions about the development of ability, and features which may not be present in the performance that is being rated, the EBBs are empirically derived from the expert analysis of sample performance. However, Fulcher (2003) argues that the EBBs rely heavily on decisions of expert raters in the development process. Upshur and Turner

found increased inter-rater reliability but no validation studies were carried out.

Alternatively, scale developers might ask raters which criteria are most important in their decision-making processes and then use these to construct the scale, for example, scaling descriptors used by North (1995, 1996/2000) for the development of the common European framework of scale for language proficiency. The main development procedure includes: (1) Developing a descriptor pool. A range of rating scales is broken down into 2,000 sentence-length descriptors. Existing descriptors are classified into different types of communicative activities and different aspects of communicative language proficiency. New descriptors are written to fill perceived gaps in the descriptive scheme. (2) Developing questionnaires. A number of pairs of teachers are given an envelope of band descriptors and asked to sort them into four or five given, related, categories and then divide descriptors belonging to the same categories into six piles. Questionnaires are developed which contain 50 descriptors and a common five-level scale and are targeted at each level and with balanced content. Questionnaires are linked by common items (North & Schneider, 1998, p. 251). (3) Teachers are asked to use descriptor questionnaires to rate language learners of different levels. The multi-faceted Rasch model (Linacre, 1989) is used to place each descriptor onto a common logit scale using the rating data. Cut-off points are established. This method is particularly useful for developing a common scale used to assess learners of a variety of second languages across a wide geographical area, with different educational systems and curricula. However, North and Schneider (1998) acknowledge that their method is a-theoretical because it is not based on empirically validated descriptions of language proficiency or on a model of the language learning process. It is a linear proficiency scale with equal intervals based on a theory of measurement.

A major limitation for empirically developed scales is that since they are derived from test performances, the descriptors may not be generally applicable outside of the specific

testing context. In addition, essay samples and examiners need to be carefully selected as the essays examined in the development procedure directly shape the comments that raters make and the criteria that they choose (Turner, 2000). However, Turner and Upshur point out that,

the lack of generality of these scales (empirically-developed scales) is not in dispute, but more general, theory-based rating scales have not been shown to be equally valid for the various task types that empirically derived scales are designed for. For performance testing, therefore, such scales are advocated, in part because of their content relevance.

(2002, p. 53)

Hudson (2005) also notes that although theoretically-based scales reflect current knowledge of language acquisition, they may not reflect all the real-world tasks an examinee needs to perform.

#### **2.4.4 A hybrid approach**

In addition to these three design methods, a hybrid approach that mixes two approaches is used. A hybrid approach can address limitations the two approaches demonstrate independently (Montee & Malone, 2014). Knoch (2009) adopts a hybrid approach which combines the theoretically-based approach and the empirically-developed approach in developing a rating scale for the Diagnostic English Language Needs Assessment (DELNA) (a diagnostic test for tertiary-level students in the University of Auckland). Knoch (2009) proposes a taxonomy in which features of Bachman and Palmer's (1996) model of communicative competence, Grabe and Kaplan's (1996) model of text construction and their writing taxonomy, the models of rater decision-making by Milanovic et al. (1996), Sakyi (2000) and Cumming et al. (2001; 2002), and Lado's (1961) Four Skills Model are grouped into different traits or categories in a rating scale. The scale developer then uses the taxonomy as a basis to decide which aspects are testable and which are not. The empirical approach is realized through applying discourse analytic

measures used in the study of second language acquisition to the DELNA test essays and building descriptors on the data analysis results.

Following Alderson (2005), who describes extensively features of diagnostic assessment, Knoch (2009) argues that a rating scale for diagnostic assessment should be able to provide useful diagnostic information on strengths and weaknesses of a learner's writing and should meet five criteria (2009, p.67-68). First, different aspects of the writing should be assessed separately and the scale organized in a way that discourages raters from displaying a halo effect. Second, scale descriptors should provide enough information for raters to rate reliably. Third, the criteria should reflect current understanding of writing development, because it represents the de facto test construct, and the descriptors should be derived from actual performance of students. Fourth, descriptors should be objectively formulated, specific and avoid vague terminology. Fifth, a score report should be issued to offer detailed feedback to students.

Knoch (2009) finds that raters using the theoretically-based empirically-developed new scale tend to provide more reliable ratings and raters respond more positively to the new scale than the original intuitively developed scale. Although Knoch (2009) also finds that the development approach is very time-consuming and is not considered practical in terms of use and development by raters, it should not be argued that such scales are not usable because the choice of a scale is more a matter of striking a balance between factors that contribute to validity of a rating scale (e.g., construct validity vs practicality) (Weigle, 2002). Bachman and Palmer (1996) suggest that the purpose of a test is important in deciding which qualities of the test are more suitable. Likewise, their suggestion seems to be true for the choice of scale development method. In the current study, the classroom assessment context is similar to the diagnostic assessment context in Knoch (2009) as the overall purpose of both assessments is to provide detailed information on strengths and weaknesses to teachers, raters and writers. According to Turner (2012, p.65), classroom



assessment results should facilitate both teaching and learning. Therefore, it is assumed that the hybrid approach used by Knoch (2009) in developing the rating scale of diagnostic assessment is suitable for the development of a rating scale for classroom assessment of Chinese college students' argumentative writing ability in the current study.

#### **2.4.5 Implications for the new rating scale**

Based on the preceding review of four kinds of scale development approaches and the purpose of the new rating scale to be developed, I argue that a hybrid approach that combines the theoretically-based approach and the data-driven approach is more appropriate for the current study than other approaches used alone or combined. First, intuition-based and theoretically-based approaches are unable to create clear and precise scoring criteria for diagnostic purposes in classroom-based assessments because of the vague terminology and lack of relevance to actual performance in descriptors. Second, though the EBB approach and the scaling descriptor approach take into account empirical evidence from student performance, they are not appropriate in the current study. The EBB approach is heavily dependent on experts' intuition. That is, regular writing teachers may have difficulty in interpreting those critical questions in EBBs consistently. The scaling descriptor approach relies heavily on existing rating scales. That is, the weaknesses of existing rating scales would be built in the new scale if the approach was adopted. In the current study, existing scales are criticized for vague descriptors and lack of criteria for quality of argumentation, therefore, the scaling approach is not appropriate. Third, the data-driven approach is not sufficient as the features or attributes of performance was pre-determined, which is not the case for the current study. In the current study, it remains a question what features of argumentative writing would be included in the new scale. Fourth, Knoch's (2009) hybrid approach that combines the theoretically-based approach and the data-driven approach addresses the limitations mentioned above by ensuring that scoring criteria are both meaningful and relevant to test performance while still being generalizable to similar contexts. This approach also tends to produce

more detailed and concrete descriptors which are necessary for classroom assessment. Knoch (2009) uses the more general term the empirically-developed approach to cover the discourse analysis and statistical analysis identified with the data-driven approach. Although the theoretically-based and data-driven approach is assumed to be suitable for the current study, it does not mean Knoch's approach can be used directly in this study. The empirically-based and data-driven approach used in the current study is specified in Chapter 3.

## **2.5 The current TEM4 rating scale and its issues**

### **2.5.1 The considerations of the selection of the current TEM4 rating scale**

The TEM4 test is administered to evaluate English language teaching and learning at the end of the foundation stage of Chinese college students enrolled in English programs in accordance with the Teaching syllabus (Zou, 1997). The TEM4 rating scale of the writing subtest is selected for the assessment of argumentative writing in the present study because Chinese tertiary-level EFL learners at the foundation stage are generally systematically taught English writing of different genres (i.e., in writing classes) and their writing ability is assessed accordingly using the TEM4 rating scale.

### **2.5.2 The current TEM4 rating scale**

The current rating scale of the TEM4 writing test was re-developed in 2010 on the basis of the original holistic scale developed in 1997. The current rating scale is an analytic scale. It comprises three major components (e.g., ideas and arguments), and various subcomponents (e.g., rhetorical organization) and descriptors related to each subcomponent (e.g., theme sentence in each paragraph). Details of the rating scale are shown in Figure 2.6 (an English version, translated by me from the original Chinese in Li, 2010, p. 111-112). Writing performance on three different components is rated against

three different-pointed scales. These scales are also labeled with qualitative adjectives, such as ‘poor’ and ‘fair’, in between different points to indicate the shades between the different points.

<p>Notes: The full score for the writing test is 15.0. Test takers’ writing performance is scored according to Ideas and arguments, Language use and Mechanics..</p> <p>Ideas and arguments: the full score is 7. Various scores from 1 to 7 can be assigned based on test takers’ writing performance: 1---2---3---4---5---6---7</p> <p>Poor Fair Good Excellent</p> <p>Language use: the full score is 6. Various scores from 1 to 6 can be assigned based on test takers’ writing performance: 1---2---3---4---5---6</p> <p>Poor Fair Good Excellent</p> <p>Mechanics: the full score is 2. Various scores from 0.5 to 2 can be assigned based on test takers’ writing performance: 0.5-----1-----1.5-----2</p> <p>Poor Fair Good Excellent</p> <p>Blank paper, a number of words and sentences irrelevant to the writing task, and a copy of writing instructions are scored 0. Raters do not need to calculate the total score.</p>	
Ideas and arguments	Ideas and content
	Relevance, substantial, clear standpoint, claim supported, insightful views
	Effectiveness
	Clear, fluent, convincing
Ideas and arguments	Rhetorical organization
	Theme sentence in each paragraph; coherence, and cohesion between sentences; coherence and cohesion between paragraphs; clear standpoint in first part; naturally-drawn conclusion in the last part; natural and reasonable arrangement of paragraphs
	Correctness
	Grammar, sentence structure, collocation, idiomatic expressions, wording
	Richness

Language use	Rich vocabulary, varied sentence structures
	Appropriateness
	Tone, authentic language
	Fluency
Mechanics	No less than word limit
	Correct spelling
	Correct punctuation
	Correct capitalization
	Artistic handwriting
	Neat layout
Total score:	

Figure 2.6 The rating scale of the TEM4 writing section (Li, 2010)

Raters are required to rate testing texts trait by trait and then total sub-scores on each trait. Zero scores are assigned for responses like blank paper, a number of words and sentences irrelevant to the writing task, and a copy of the writing instructions.

### 2.5.3 The development of the current TEM4 rating scale of writing

Development of the current TEM4 rating scale is described in Li (2010). The development of the scale is phased. The first phase is to establish a construct of Chinese EFL learners' writing ability. The construct is established through the comparison of TEM4 raters' and experts' views of Chinese EFL learners' writing ability based on students' test responses from the TEM4 writing tests. Two methods are used for this phase: questionnaire and guided rating. The second phase is to establish the weighting of different dimensions. The weighting is based on raters' expertise and experience.

The two-phase development can be divided into a number of steps (Li, 2010, p. 65–70):

1. The researcher collects research on text analysis of ESL/EFL writing scripts, and a

number of well-known ESL/EFL writing scales, and selects 45 descriptors (e.g., spelling is correct) that describe different aspects of ESL/EFL written text.

2. The researcher develops a questionnaire based on the 45 descriptors selected in the first step. The descriptors are presented in positively formulated subject-plus-predicate phrases and grouped into nine dimensions (i.e., layout, accuracy, appropriacy, richness, organization, idea, sensitivity to culture, fluency and effectiveness of communication). Each descriptor is given a five-level Likert scale to measure what raters thought of the descriptor in terms of its degree of importance in representing the EFL writing ability of Chinese EFL learners majoring in English. A part of the questionnaire is shown in Figure 2.7.

Accuracy	
(5) correct vocabulary spelling	1-----2-----3-----4-----5
(6) correct punctuation usage	1-----2-----3-----4-----5
(7) correct capitalization	1-----2-----3-----4-----5
(8) complete sentence structure	1-----2-----3-----4-----5
(9) correct grammatical rules	1-----2-----3-----4-----5
(10) sentential cohesion	1-----2-----3-----4-----5
(11) correct sentence structure	1-----2-----3-----4-----5
(12) word accurately	1-----2-----3-----4-----5
(13) word properly	1-----2-----3-----4-----5
(14) correct set phrase use	1-----2-----3-----4-----5

Figure 2.7 The questionnaire on the writing abilities of Chinese EFL learners majoring in English - Accuracy (Li, 2010)

The questionnaire is reviewed by five Chinese experts in applied linguistics and language assessment. It is then sent to 100 raters of Test for English Majors – Band 8 (TEM8) in April 2008 for piloting of the questionnaire. TEM8 is a test administered

to evaluate English language teaching and learning at the end of the four years' study of Chinese college students enrolled in English programs. At the time of Li's (2010) study, the TEM8 writing subtest required test takers to respond to a prompt (e.g., the introduction of a controversial topic) and an outline with a 400 or so word expository or argumentative essay. In the most recent adjustments of the items of TEM8, administered in 2016 (Wang, 2016), TEM8 designers abandoned the outline and replace it with reading material (e.g., two excerpts with two opposing points of view on one topic). The TEM8 rating scale is generally similar to that of TEM4 (see Figure 2.6) in terms of format and content (e.g., categories and descriptors) except for two sets of numerical scales attached to ideas and content, and language use (two general category). The full score for the TEM8 writing subtest is 20 rather than the 15 in TEM4. Ideas and content is assessed with a 10-point scale (i.e., 1---2---3---4---5---6---7---8---9--10) and a verbal scale (i.e., *very poor* - *poor* - *fair* - *good* - *excellent*). Language use is assessed with an 8-point scale (i.e., 1---2---3---4---5---6---7---8) and a verbal scale (i.e., *very poor* - *poor* - *fair* - *good* - *excellent*). Other parts of the rating scale remain the same as that of TEM4.

3. The level of Likert scale, the construct validity of the questionnaire, and the reliability of the questionnaire are investigated. Modifications are made to inaccurate wording, repetition and improper categorization of the descriptors.
4. The modified questionnaire is given to 200 TEM4 raters who participated in the rating of written responses of TEM4 in May 2008. Opinions on the degree of importance of the descriptors for representing the writing ability of Chinese EFL English majors are investigated. Again, modifications are made to inaccurate wording, repetition and improper categorization of the descriptors. Thirty items of descriptors are selected from the total 45 and nine dimensions are confirmed.
5. Twenty-two volunteer raters are recruited to rate 23 written scripts from the TEM4 writing subtest administered separately in 2005. The raters are required to finish three tasks: rating writing scripts holistically, giving reasons representing at least four

aspects of writing ability to support the holistic rating and listing them in the sequence of importance, and giving their ideas on the weighting of the different aspects listed in the second task in the form of the percentage score. The scale used for holistic rating includes five levels and five point ranges: excellence (13–15 points), fair (10–12 points), average (7–9 points), poor (4–6 points), and very poor (0–3 points). The rating reasons are collected and grouped into 15 categories according to key words in each rating reason. Fifteen dimensions are further merged into eight dimensions. The average weighting and the frequency of occurrence of the different dimensions of rating reasons in the form of a percentage score are calculated and compared.

6. The descriptors and categories collected from the questionnaire phase and those collected from the rating process are compared and collapsed. An initial scale is formulated which comprises four broad categories (i.e., ideas and arguments, language ability, organization, layout), thirty-one descriptors (e.g., vocabulary spelling is correct), a six-point numerical scale (ranging from 0–5) indicating the degree of match between the writing scripts to be assessed, and the descriptors grouped in four categories. A labelled scale ranging from ‘completely not matching’ and ‘basically matching’, to ‘completely matching’ is attached to the numerical scale to indicate the continuum of match. A weighting system is formulated (i.e., 120% for ideas and arguments, 100% for language ability, 60% for organization, and 20% for layout).
7. Ten experts are invited to review the initial rating scale and give advice on the appropriacy of categorization, clearness of the descriptors, properness of weighting, and operationalization. Modifications are made: a zero point is assigned to blank paper, copying testing instruction or writing prompts; the numerical scale is changed to range from 1 to 5 points; mechanics and related descriptors that used to be included in ‘ideas and arguments’ are singled out and merged with layout, and organization is merged into ‘ideas and arguments’ Three categories (ideas and arguments, language use, mechanics) are formulated; the weighting system is assigned to three categories

(140% for ideas and arguments, 120% for language use, and 40% for mechanics).

#### **2.5.4 The strengths and issues of the current TEM4 rating scale of writing**

The current TEM4 rating scale is more valid and reliable than the original holistic TEM4 rating scale in several ways: 1) It fulfills the purpose of assessment of Chinese EFL learners in TEM4 by providing analytic scales for different aspects of writing ability. As mentioned by Weigle (2002) (see Section 2.2.2), analytic scales acknowledge the uneven profiles typical of L2 learners (learners of English as a second language) and EFL learners. Thus the scale is more appropriate for the purpose of the assessment of Chinese EFL learners. In the validation study of the existing TEM4 rating scale, Li (2010; 2014) finds that 1) TEM4 raters provide more positive comments on the existing scale than on the original rating scale, in terms of positive washback to instruction of writing and accurate rating descriptors. 2) The TEM4 rating scale reflects raters'/readers' perceptions of the construct (i.e., writing ability) of the TEM4 writing subtest. Rating is in essence a reading process. Cumming et al (2001) emphasize the raters' role as a reader in the rating process and develop the TOEFL rating scale through investigating raters' decision-making behaviors (e.g., what aspects are attended to in certain rating behaviors). 3) It provides reliable criteria for TEM4 raters. Li (2010; 2014) found a higher consistency (i.e., Kendall's Coefficient of Concordance, .746) in rating results between different TEM4 raters in the existing rating scale than the original rating scale (i.e., Kendall's Coefficient of Concordance, .673) and better performance in terms of rater separation, candidate separation, and variation in ratings using Rasch analysis of rating results produced by the existing rating scale than those produced by the original scale.

Little research has been conducted on the use of the TEM4 rating scale outside the TEM4 context, for example, in the classroom assessment context. Despite little research on its use in the classroom context, its use can be expected in the classroom assessment context



as it is the only rating scale of writing that is developed specific to the target population of TEM4 and the current study, and the TEM4 test is in nature a criterion-based test and aims to provide positive feedback to the writing classroom. Therefore, for the purpose of this study, I will discuss a number of concerns about the validity of its use in classroom assessment.

Firstly, the scoring criteria only provide graduation or levels of quality rather than detailed descriptors. The use of qualifiers such as ‘poor’ and ‘average’ do not provide more information than a score to students and writing teachers. The diagnostic feedback based on such descriptors poses interpretation issues to student and writing teachers. Secondly, heterogeneous categories rated on a single scale fail to provide consistent interpretation of scores to score users. For example, ‘ideas and content’ and ‘rhetorical organization’ are combined to form ‘ideas and arguments’. When a rater assigns a point of five using the ‘ideas and arguments’ scale, a number of interpretations may occur, as it may refer to a good ideas and content but poor rhetorical organization, or good rhetorical organization but poor ideas and content. This problem is reported by raters when using the current scale (Li, 2010). Thirdly, although the written samples used for development of the rating scale are representative of different levels, they are responses to one writing task. Therefore, the generalizability of the rating scale to the type of tasks that the test claims to assess (i.e., expository writing, argumentative writing) is doubted. In all, although the current TEM4 rating scale is more valid and reliable than the original TEM4 rating scale, the validity of its use remains a question in the classroom assessment context in which detailed and clear diagnostic information is expected.

In order to overcome these limitations, this study sets out to analyze a large number of writing samples from three writing tasks to solve the generalizability issue, develop separate scales for each category to provide more diagnostic information to students and writing teachers, and build the results of data analysis into descriptors to make descriptors

more concrete and avoid vagueness.

## **Chapter 3     A theoretical model of argumentative writing ability and its operational framework**

In Chapter 2, I have argued that the rating scale of Chinese EFL learners' argumentative writing ability should be (1) based on a theory of language or writing development and (2) based on an empirical investigation of written scripts. This chapter has two aims. The first aim is to provide such a theoretical model for the development of a rating scale of argumentative writing ability. To reach this aim, first, existing theories or models are reviewed. Then, the reasons why these models are unable to be used alone, but should be adapted and put together, are discussed. Finally, a theoretical framework of argumentative writing ability is built.

The second aim is to establish an operational framework of discourse measures for the empirical investigation of argumentative writing scripts. To reach this aim, the chapter reviews discourse measures that have been used to investigate language development in second language acquisition and writing studies. Discourse measures that have been successful in distinguishing the development of different components of writing ability are selected. Finally, an operational framework of discourse measures is formulated.

### **3.1     Theoretical basis for scale design**

Although it has been argued that a theoretical model is necessary for the scale design, it remains a question as to what kind of theoretical model to base a rating scale of argumentative writing on. Ideally, the progression in a language proficiency scale should be based on stages of attainment in the learning process. However, North (2003, p. 10) argues that the attempt to describe the stages of learning surpasses the state of our knowledge of the learning process. He further argues that it is sensible to have a valid conceptual framework and try and incorporate relevant insights from theory (p. 22). Therefore, he suggests that models of language use are a logical starting point.

Three models of language use that are relevant to the current study are selected. They are: i) the model of communicative language ability (the CLA model) (Bachman, 1990; Bachman and Palmer, 1996), ii) the taxonomy of academic writing skills, knowledge bases and process (Grabe & Kaplan, 1996) and iii) the model of writing competence (Connor & Mbaye, 2002) (see Table 3.1). The CLA model is a model that describes how different elements of language use (e.g., knowledge, strategies, and characteristics of test takers) interact with each other in different testing situations. It is relevant to this study because of its explicitness in describing what is involved in language test performance, thus providing potentially useful information regarding assessing argumentative writing performance in the current study. As Fulcher and Davidson (2007, p. 39) comment, if a model is ‘fine grained’, it can be used to develop criteria for the evaluation of language performance at different levels of proficiency. McNamara (1996, p. 66) considers the CLA model a refinement and elaboration of Canale and Swain’s (1980) model of communicative competence. Skehan (1998) argues that the CLA model is grounded in linguistic theory and more empirically based than previous models of communicative competence developed by Hymes (1967), Canale and Swain (1980) and Canale (1983). Grabe and Kaplan’s (1996) taxonomy is a collection of information involved in writing from an ethnographic perspective and is categorized into different writing skills, knowledge bases and processes. It is, according to Grabe and Kaplan (1996), a useful way to identify any gaps that can be further investigated in writing research. Since the CLA model is developed as a general language ability model, Grabe and Kaplan’s (1996) taxonomy is expected to complement the CLA model by providing more writing-related information. Connor and Mbaye’s (2002) model of writing competence is built on the model of communicative competence developed by Canale and Swain (1980) and Canale (1983) and empirical studies on persuasive and argumentative writing (Connor, 1990; Connor and Lauer, 1985, 1988). Compared with the CLA model and Grabe and Kaplan’s taxonomy, this model is more tailored to argumentative writing and therefore more

relevant to the current study.

Although these models or taxonomy were expected to provide a theoretical basis for the development of the rating scale in the current study, I found after a review that there are a number of limitations with them which prevent a rating scale being built on them directly. Three of four limitations are acknowledged by Knoch (2009) in her attempts to develop the DELNA scale on these models. These limitations also exist for the current study. First, these models or taxonomy are in fact a model or taxonomy of underlying ability/knowledge/competence. One of the typical problems with the use of a model of competence, described by North (2003, p. 32), is that it has difficulty in coping with what happens when performance is assessed. For example, fluency, the most obvious feature of performance has no place in a model of competence. Second, these models or taxonomy are hard to operationalize. North (2003) describes it as another typical problem with the model of competence. He argues that, while certain aspects are conceived of as parameters of a theoretical model, this does not necessarily mean they can be isolated as observable components and hence rated or tested separately. Third, these models or taxonomy fail to account for the fact that some of the competences might be more important in some situations than in others. Fourth, these models or taxonomy fail to account for argumentation ability adequately. This last limitation is further discussed below as it is most relevant to the current study and one might argue that argumentation ability should not be assessed in a language test.

An increasing body of text analytical studies shows that the quality of arguments is a good predictor of writing quality (e.g., Connor, 1990; Chase, 2011). Studies on raters' decision-making behaviors also show that raters do attend to features of discourse types. For example, in the descriptive framework of decision-making behaviors, Cumming, Kantor, and Powers (2001; 2002) show that raters 'discern rhetorical structure', 'assess reasoning, logic or topic development', 'assess text organization... discourse functions or genre',

and ‘rate ideas or rhetoric’. However, there are few explicit mentions of argumentation ability in current theoretical models or taxonomy. In the CLA model, argumentation ability is generalized as rhetorical organizational ability. For example, ‘rhetorical organization’ under the textual competence component is described as ‘the overall conceptual structure of a text (e.g., common methods of development like narration, description, comparison, classification and process analysis)’ (Bachman, 1990, p. 88). In the model of text construction, Grabe and Kaplan (1996, p. 61) claim that the model should address important hypotheses and findings, of which two are related to type of discourse or writing: a theory of text type variation is possible and is needed for comprehension, production, and assessment research; learning to write requires manipulation of many complex structural and rhetorical dimensions, with greater complexity occurring in expository/argumentative writing. However, despite the acknowledgment of the existence of the distinctions in discourse types, Grabe and Kaplan do not explain how the model accounts for such distinctions. In their taxonomy of academic writing skills, knowledge bases and process, the minimal mention is restricted to ‘knowledge of genre structure and genre constraints’, and ‘knowledge of organizing schemes (top-level discourse structure)’ and neither of them is further explained.

The model which explicitly mentions the competence of argumentative writing is Connor and Mbaye’s (2002) model of writing competence. They build on the communicative competence model proposed by Canale and Swain (1980) and Canale (1983) and empirical studies on persuasive and argumentative writing (Connor, 1990; Connor & Lauer, 1985, 1988), and propose the inclusion of features that writing research has found important, for example, ‘appeals’, ‘pertinence of claims’ and ‘warrants’. However, the authors do not specify why they are grouped as strategic competence, but the concept of strategic competence is clearly distinct from the definition provided by Bachman (1990), ‘a general ability that makes use of other abilities to finish a task’ (p. 106), including ‘goal-setting, assessment, and planning’ (p. 100).

Some might argue that if this is the case, one should give up the idea of developing a rating scale based on a theoretical framework. Yet a theoretically-based scale has the advantage of being transferrable across contexts and assuring generalizability of results. If a theoretical model is necessary, it is preferable to investigate what such a model or theory should look like. According to McNamara (1996, cited in Knoch, 2009, p. 73), such a model should satisfy three requirements: (1) It should be rich enough to conceptualize any issue which might potentially be relevant to cope with performance. (2) There should be a careful research agenda to investigate the significance of the different measurement variables proposed in the model. (3) These variables to be investigated should be appropriate and practical to assess in a given test situation. Following McNamara's (1996) first requirement, I now propose a theoretical model of argumentative writing ability (the AWA model). The AWA model retains all available components from the current models or taxonomy that would account for argumentative writing performance. I also propose an operational framework of analytic measures in relation to components of the theoretical model to meet McNamara's second and third requirements. By establishing the operational model, the operationalization problem, that aspects that are conceived of as parameters of a model of competence or components in theory-based scale design cannot be separated and observed, is addressed.

### **3.2 Building the theoretical model of argumentative writing ability (the AWA model)**

The theoretical model of argumentative writing ability (the AWA model) is a multi-component hierarchical model. It is built through comparing all the components and subcomponents of three theoretical models (see Table 3.1). The comparison is possible because all three models have roots in the models of communicative competence, especially Canale and Swain's (1980) and Canale's (1983) models, thus providing a common foundation for comparison. The AWA model comprises language competence,

argumentation competence, strategic competence and topical knowledge at a higher level of hierarchy, and subcomponents at a lower level of hierarchy. Following Bachman's (1990) and Bachman and Palmer's (1996) CLA models, different components and subcomponents of the AWA model interact with each other in language use. Similarly, strategic competence and topical knowledge, though necessary in language use, are not investigated in the current study as they are either hard to assess or not the focus of argumentative writing assessment (Bachman, 1990; Bachman & Palmer, 1996; McNamara, 1996). Strategic competence refers to 'mental capacity for implementing the components of language competence' (Bachman, 1990, p. 85). Oller (1983, cited in Bachman, 1990) hypothesizes that a general factor of language proficiency (e.g., strategic competence) is the principal function of intelligence, while Carroll (personal communication with Bachman, 1990) holds that intelligence is not totally independent but distinct from language abilities. Bachman (1990, p. 106) argues that it is inaccurate to identify strategic competence with intelligence and it should be left to validation studies of constructs to decide whether it should be measured. Topical knowledge (also knowledge of the world in Bachman, 1990) is 'loosely defined as knowledge structure in long-term memory' (Bachman & Palmer, 1996, p. 65). Bachman and Palmer (1996) argue that topical knowledge provides an information base that enables language learners to use language with reference to the world, and is thus involved in all language use. McNamara (1996) argues it is odd to exclude topical knowledge from the assessment but he also acknowledges it is complex to assess it. Bachman and Palmer (1996) caution that the assessment of topical knowledge that one candidate has is unfair to those who do not have it. Considering the aim of the argumentative writing assessment in the current study is to assess general language competence and argumentation competence, and the nature of strategic competence and topical knowledge, I decided not to assess strategic competence and topical knowledge in the current study but maintain their existence in the theoretical model of argumentative writing ability as they are essential for any argumentative writing to be composed. For the purpose of this research, the AWA model comprises only



language competence and argumentation competence.

The theoretical basis for the new rating scale provided by the AWA model is mainly focused on language competence and argumentation competence and their subcomponents. Language competence is essential in an argumentative writing test as the writing activity is impossible if no language is used. The language competence component is developed by comparing its subcomponents as specified in existing models. Equally important in argumentative writing test, I argue, is argumentation competence/knowledge. The aspects that are conceived of as belonging to argumentation knowledge are developed based on textual analytical studies of argumentative writing.

### **3.2.1 General components**

Three questions arise during the development of the AWA model: (1) Are there overlaps and differences between different models? What are they? (2) Which of the components and subcomponents should be included in the AWA model? The three models discussed in Section 3.1 are adapted and mapped on a common table (see Table 3.1). The components of strategic competence and the knowledge of the world (or topical knowledge in Bachman and Palmer's (1996) term) in these models are removed from the comparison but are presented for the comparison between original models. In Table 3.1, the first column includes six components of language competence: linguistic/grammatical competence, textual/discourse competence, functional competence, sociolinguistic competence, strategic competence and knowledge of the world (topical knowledge). The second, third and fourth columns are specific subcomponents from different models. According to Bachman and Palmer (1996), these components are essential for a language ability model, but it needs to be discussed whether each of these components can be assessed in an argumentative writing test.

It becomes clear from the table that some models are more extensive than others. Grabe

and Kaplan's (1996) taxonomy and Connor and Mbaye's (2002) model of writing competence do not incorporate functional knowledge, while the CLA model does. Bachman and Palmer (1996) define functional knowledge (illocutionary competence in Bachman's (1990) model) as the knowledge of conventions for performing acceptable language functions (e.g., a function performed by saying something, or a speech act). While acknowledging speech acts as language functions, Bachman (1990) does not seem to be satisfied with identifying language functions solely with speech acts. Bachman (1990) introduced a broader framework drawing on Halliday (1973, 1976), including four functions: ideational (expressing meaning), manipulative (affecting the world), heuristic (extending knowledge), and imaginative (extending our environment for esthetic purposes, e.g., creating metaphors) into this category. Grabe and Kaplan (1996) do not take language functions as a separate category, rather, they treat language functions as part of sociolinguistic competence, of which the primary feature is appropriacy. That is, they view functional knowledge of language as whether it is appropriate in different contexts, rather than knowing how to express or interpret the functions that language performs. The former focuses on appropriacy, while the latter focuses on knowledge itself. Grabe and Kaplan's (1996) functional uses of written language include apologizing, denying, etc. It is not clear how Connor and Mbaye (2002) view functional knowledge as they do not mention it in their model. It seems that argumentative writing is not a good task to assess functional knowledge as the functional use of written language in argumentative writing is solely-to convince and persuade (e.g., speech acts, Grabe and Kaplan's (1996) functional uses of language), while Bachman and Palmer's (1996) more general point of view is hard to operationalize. It can therefore be argued that a writer's functional knowledge cannot be assessed on the basis of argumentative writing as it is not easy to make an inference about a writer's functional knowledge from a single type of writing. Therefore, functional knowledge is not included in the AWA model.

Table 3.1 Comparison between the CLA model, part of Grabe and Kaplan's (1996) taxonomy, and Connor and Mbaye's (2002) model of writing competence

Model Component	The CLA model	Part of Grabe and Kaplan's (1996) taxonomy	Connor and Mbaye's (2002) model of writing competence
Linguistic/grammatical competence	1. Vocabulary 2. Morphology 3. Syntax 4. Phonology /graphology	1. the written code a. Orthography, b. Spelling, c. Punctuation, d. Formatting conventions (margins, paragraphing, spacing, etc.) 2. phonology and morphology 3. Vocabulary 4. Syntactic knowledge	1. Grammar, 2. Vocabulary, 3. Spelling, 4. Punctuation
Textual/ discourse competence	5. Cohesion 6. Rhetorical organization	5. intra-sentential and inter-sentential marking devices (cohesion, syntactic parallelism) 6. informational structuring (topic/comment, given/new, theme/ rhyme, adjacency pairs) 7. semantic relations across clauses 8. Knowledge to recognize main topics 9. genre structure and genre constraints 10. organizing schemes (topic-level discourse structure)	5. Discourse organization, 6. Cohesion, 7. Coherence
Functional knowledge	7. Ideational functions 8. Manipulative functions 9. Heuristic functions 10. Imaginative functions	11. Functional uses of written language (e.g. a. Apologize, b. Deny)	N/A
Sociolinguistic competence	11. Sensitivity to Dialect or Variety 12. Sensitivity to Register 13. Sensitivity to natural or idiomatic expression 14. The ability to interpret cultural references and figures of speech	12. Register 13. Audience consideration (e.g., a. number in audience, b. degree of familiarity with audience) 14. Awareness of sociolinguistic differences across languages and cultures 15. Self-awareness of roles of register	8. Written Genre Appropriacy, 9. Register, 10. Tone
Strategic competence	14. Goal-setting 15. Assessment 16. Planning	14. Writing process skills (online processing skills; not linear) Goal planning routines, generating content, propositional integration, etc. 15. Writing process strategies (executive control or metacognitive strategies) Monitoring text production. Generating additional content, considering task problems, etc.	11. Audience/Reader awareness, 12. Appeals, 13. Pertinence of Claims, 14. Warrants
Knowledge of the world	N/A	16. Declarative (semantic, topical) 17. Episodic (events, personal experiences, interactional) 18. Procedural (processes, routines, conventions)	N/A

It is also clear from Table 3.1 that all models have linguistic/grammatical competence, textual competence and sociolinguistic competence, but they differ in the descriptions of their subcomponents. Grabe and Kaplan's (1996) taxonomy is more detailed than the other two models in linguistic and textual competence. For linguistic/grammatical competence and textual competence, all models include syntax or grammar, vocabulary, cohesion, coherence, and rhetorical organization. These are the essential knowledge types that a writer needs to compose a writing product and they are commonly assessed (e.g., Bachman & Palmer, 1996). For other components, phonology cannot be assessed through writing assessment, margins and spacing are normally assessed in computer-based assessment, while spelling, punctuation, paragraphing, and morphology are usually assessed in writing assessment. Cumming et al. (2001; 2002) show that raters see spelling, punctuation and morphology as important in their rating process.

The three models are roughly similar and mainly contain knowledge of register and differences across languages and cultures for sociolinguistic competence. Register (Halliday, McIntosh, & Stevens, 1964, cited in Bachman, 1990) describes three aspects of language use: field, mode and style. The field refers to subject matter of the language use. Mode refers to spoken and written mode. The style of discourse includes: frozen, formal, consultative, casual, and intimate. Sociolinguistic competence is defined as the sensitivity to conventions of language use that are determined by the features of specific language contexts (Bachman, 1990, p. 94). It includes the sensitivity to dialect or variety, the sensitivity to native or idiomatic expressions, and the sensitivity to cultural references and figures of speech (Bachman, 1990). It is not clear to what extent these subcomponents influence the quality of argumentative writing. For example, students are expected to use American English and British English rather than other varieties (e.g., Jamaican English) in formal writing. However, since only American English and British English are exposed to Chinese EFL students and the teaching and assessment of the distinction between American English and British English is generally lacking in the Chinese EFL context,

therefore, knowledge of dialects or varieties can be least expected to contribute to the quality of argumentative writing for Chinese EFL students. For naturalness of language use, it remains a question whether it is practical for raters who are non-native speakers of English to make a reliable judgment and it even poses a question to native speaker raters as native speakers' language use varies from person to person.

For idiomatic expressions, cultural references and figures of speech, it remains an empirical question as to whether raters react positively to them in an essay. Bachman (1990, p. 97) argues that many cultural references and figures of speech are incorporated into the lexicon of any language, and can thus be considered part of lexical, or vocabulary, competence. He further argues that regardless of this, knowledge of these two components is required whenever these meanings are referred to in language use. Therefore, assessing the sensitivity to idiomatic expression, cultural reference and figures of speech in written texts is possible and probably could be operationalized as the number of idiomatic expressions, cultural references and figures of speech used in an argumentative essay. Assessing knowledge of register and genre seems not possible with a single type of genre or register for Chinese EFL college students' argumentative writing, unless certain textual features have been identified with that type of genre. In this case, the inclusion of more of these textual features could be interpreted as showing a greater competency in this type of genre, although this does not mean that frequency is the only important aspect of textual features of genres. However, the only possible knowledge of register that seems to be concerned is the distinction between formal and informal writing. Argumentative writing can be regarded as formal compared with the language used in conversations in informal settings. Therefore, textual features that are typical of spoken language can be expected to occur less in formal writing. Based on this discussion, sociolinguistic knowledge to be assessed in argumentative writing will be restricted to knowledge of the difference between formal and informal writing and knowledge of idiomatic expression, cultural reference and figures of speech.

From the above, I argue that linguistic competence, textual competence, and sociolinguistic competence are essential for general writing ability. Linguistic competence includes vocabulary, syntax, morphology, spelling, and punctuation. Textual competence includes cohesion, coherence and rhetorical organization. Sociolinguistic competence includes idiomatic expression, cultural reference, figures of speech, and informal and formal writing style. The ‘appeals’, ‘pertinence of claims’ and ‘warrants’ proposed in Connor and Mbaye’s (2002) model serve as a starting point for argumentation competence.

### **3.2.2 Argumentation knowledge**

Investigating what comprises argumentation knowledge seems impossible without exploring existing models of argumentation. A literature review of existing models of argumentation and empirical studies on argumentation shows that argumentation is complex. The ability to argue can refer to the ability to produce an argument – basically, a claim and a reason supporting it – and it can also refer to the ability to evaluate an argument; argumentation can refer to a process of arguing and it can also refer to the product resulting from it; argumentation can refer to justifying a standpoint and it can also refer to refuting a standpoint. For the purpose of this study, the knowledge discussed in this study is restricted to the perspectives which consider argumentation ability as production ability and argumentation as a product resulting from this ability. The knowledge of argumentation, therefore, is locally defined as the understanding of how reasons are put forward to defend a standpoint in a discourse.

With this definition in mind, I conducted a literature review of textual studies on argumentative writing and argumentation models investigated in argumentative writing. The literature review suggests three aspects of knowledge of argumentation: the structure, substance (e.g., content), and appeals of argument or argumentation. In the following

sections, the theoretical basis of these three aspects of argumentation is discussed, followed by research that has investigated these three aspects.

### 3.2.2.1 *Structure of argumentation*

The most influential argumentation theory in written discourse studies is Toulmin's model or scheme. Toulmin (1958/2003) provides a model of good argument. The model includes six elements. A claim is an assertion put forward for general acceptance and data are given to justify the claim. When the connection between the claim and the data is challenged by the other party of an argument, a warrant is put forward to authorize the connection. A warrant can be part of rules, principles, hypothesis, and inference-license. When the warrant is further challenged by the other party, a backing is put forward to establish the trustworthiness of the warrant. Qualifiers are usually added to the claim to indicate the strengths and limitations of the conclusion drawn from supporting data. Rebuttals are part of the conclusion, which describes the circumstances that might undermine the force of the supporting arguments. Figure 3.1 graphically presents how these elements are interrelated within a single argument (Toulmin, 1958/2003, p. 97).

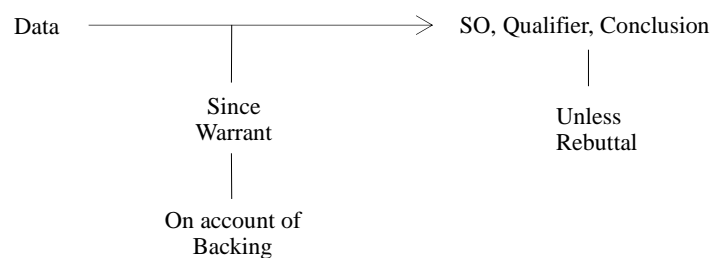


Figure 3.1 Toulmin's argument elements (Toulmin, 1958/2003, p. 92)

Despite the clear relations between different elements within a single argument, Toulmin's model is limited when applied to actual argumentation. In actual argumentation, an argument could serve as a claim or data for another argument ('chains of arguments' in Toulmin's term); a claim could be supported by more than one data (multiple lines of

arguments); or there could be the existence of two-sided arguments: arguments supporting a position the writer is in favor of and arguments supporting a position the writer is against (counterargument). Many argumentation researchers have noticed these situations (Toulmin, 1958/2003; van Eemeren, Grootendorst and Henkemans; 2002; Wolfe, Britt and Butler, 2009). Toulmin (1958/2003) states that an argument is liable to become the starting point for a further argument, and this second argument tends to become the starting point for a third argument, and so on. He further argues that there are situations in which multiple parties in an argument put forward different grounds, and the different grounds serve as a separate starting point for another argument. van Eemeren et al. (2002) describe the different relations in complex arguments and distinguish between three types: multiple argumentation consists of a standpoint and more than one reason which are independent of one another in support of the standpoint; coordinative argumentation consists of a standpoint and more than one reason which are dependent on one another in support of the standpoint; subordinative argumentation consists of an initial standpoint and supporting data which are also supported by another layer of reasons. They also argue that these structures of argumentation can occur in combination. Perkins (1985, cited in Wolfe, et al., 2009) states that researchers tend to generate more arguments (reasons) in favor of a position they support than reasons on the other side ('myside bias'). Students tend to have myside bias (Coffin, Hewings & North, 2012; Nussbaum, Kardash, & Graham, 2005; Wolfe & Britt, 2008).

Toulmin's model has another limitation. The identification of elements in the model is difficult. Sampson and Clark (2008), for example, provide examples of statements made by students that were interpreted as any of claims, warrants, qualifiers or rebuttals depending on the perspective of the reader. Simon (2008) also notes that identifying statements as data, warrants, or backings can be ambiguous, especially when a claim is implicitly stated. In order to overcome these limitations, some researchers have adapted Toulmin's model. For example, Crammond (1997; 1998) proposes a model of argument



structure based on Toulmin's model by adding rebuttals (i.e., opposing arguments), an alternative solution (to a controversial issue), counter-rebuttals (i.e., rebuttals to the opposing arguments) to acknowledge the existence of opposing arguments. He also proposes a semantic coding system to facilitate identifying argumentative elements, like warrants, backings that have been argued as difficult to identify. But this system is too complex to apply by human raters and it is not discussed in this study.

A model which takes into account all features that exist in actual argumentation is proposed by Ferretti, Lewis, and Andrews-Weckerly (2009, cited in Chase, 2011). Based on van Eemeren et al.'s (2002) structure of argumentation mentioned above, Ferretti et al. (2009) propose a model of argumentative writing. The model includes the basic argument elements in Toulmin's model (i.e., claims and data), the logical relations in van Eemeren et al.'s (2002) structure of argumentation (i.e., coordination, subordination), and the opposing arguments (e.g., alternative standpoint, counterarguments). In addition, it also acknowledges other elements typical of argumentative writing: the introduction of a controversial topic as the introduction element, the reinforcement of one's standpoints as the conclusion element, the cohesive connection of a chain of arguments as a functional marker, the achievement of rhetorical effects by repetitions, the element which does not contribute to the strength of argumentation as irrelevant information on the topic. Chase (2011) further distinguishes reasons that offer direct support for the standpoint as level-1 reasons, and reasons subordinate to level-1 reasons that are offered as support for reasons above them as level-2 reasons and below.

In view of the discussion of existing models of arguments or argumentation, it is proposed that students' structural knowledge of argumentation comprises structure of a single argument (i.e., claim and reason), structure of complex arguments (e.g., chains of arguments, multiple lines of arguments), and a balanced structure of argumentation (i.e., including both myside arguments and yourside arguments). They are included in the AWA

model.

### 3.2.2.2 *Substance of arguments*

Another essential component of an argument or an argumentation is the substance (or content) of the argument. This is another aspect of argumentation competence. The evaluation of quality of argumentation only by structure is partial without considering the content of arguments. That is, a sophisticated or balanced structure which includes both sides of arguments may contribute to a strong argumentation, but without knowing the content presented in these elements little would be known about how strong the argumentation actually is.

Soundness is a concept that is often used to evaluate the quality of an argument as a whole (Hughes & Lavery, 2008, p. 21; van Eemeren et al., 2002, p. 93; Govier, 2013, p. 88). According to Hughes and Lavery (2008), an argument is sound when the premises are true (a premise can be judged to be true or false, and offers support to a conclusion), the inference made from the premises to the conclusion is strong (or valid in van Eemeren, et al.'s term) and the conclusion is true. The inference is referred to as logical strength or validity and is often presented in one of two forms:

a) 1. If..., then... 2. ...Therefore: 3....

b) 1. If..., then... 2. Not...Therefore: 3. Not....

van Eemeren et al., (2002, p. 95) state that invalid reasoning can be made valid by adding an 'if...then...' conditional. Hughes and Lavery (2008, p. 21) argue that logical strength is the property of the connection between the conclusion and the premise and never of the statement in the premise or conclusion. In both these cases the research suggests that the soundness of an argument is an issue of not only the form, but also the content of the argument, and the content of an argument is independent of the logical strength of the argument. Since the content of an argument is more relevant to the present study, the evaluation of the soundness of an argument is restricted to the content, while the form or

logical strength is out of the scope of this study.

Three criteria have often been proposed for evaluating the soundness of argument in actual argumentation: acceptability, relevance and sufficiency/adequacy of premises (Hughes & Lavery, 2008) (or cogency of arguments in Govier (2013)). The acceptability of premises refers to the condition where there is good evidence to support the reader in believing the premises, even if they are not known to be true, and there is no good evidence indicating that the premises are false. The relevance of its premises refers to the condition where premises state evidence to support the conclusion. The adequacy of the premises refers to the condition where all premises considered together give sufficient support to justify belief in the conclusion.

Acceptability and relevance are operationalized into analytic scales in the empirical studies that investigate the quality of argument (Schwarz, Neuman, Gil, & Ilya, 2003; Stapleton and Wu, 2015). In Schwarz et al. (2003), acceptability and relevance are evaluated separately on a scale of 0–2 and then added together to form an assessment of ‘soundness’ on a scale of 0–4. In an exploratory study of how the quality of reasoning can be assessed in argumentative essays with good structure, Stapleton and Wu (2015) operationalize acceptability and relevance on one scale with ‘being relevant’ as a necessary condition for further consideration of ‘acceptability’, although the development of the scale of acceptability and relevance is not justified in their study. However, it seems rational that reasons need to be relevant before they are acceptable. It seems so because, for those which are irrelevant, even if they are based on factual information, they contribute too little to the persuasiveness of an argumentative essay (e.g., the degree to which a reader feels convinced by an argument), and thus are not worthy of investigation of their acceptability. Govier (2013) distinguishes three types of relevance conditions: positive relevance (premises offer reasons in favor of a conclusion), negative relevance (premises offer reasons against a conclusion) and irrelevance (premises do not offer

reasons in favor of or against a conclusion). Negative relevance and irrelevance of premises are flaws of an argument, thus rendering an argument not cogent (i.e., not sound).

While being relevant seems a necessary condition, the acceptability of premises seems to be a matter of degree as readers of the arguments need to appeal to the real world to understand their truth condition while in the real world the truth condition of premises seems not as accurate as that in the natural sciences. A number of researchers have commented on the acceptability of premises (Govier, 2013; van Eemeren et al., 2002; Means & Voss, 1996). van Eemeren et al. (2002, p. 93) state that statements which are facts are more acceptable than those based on values or judgments, while the latter statements need more argumentation to demonstrate their acceptability. Govier (2013) enumerates four general conditions when premises are regarded as acceptable: premises supported by a cogent sub-argument, premises known a priori, common knowledge, and plausible testimony. Means and Voss (1996) operationalize sufficiency as ‘the more the better’, although this may not be true, but it is probably true when the premises are acceptable and relevant.

In addition to the three criteria proposed by the above researchers, van Eemeren et al. (2002, p. 93) further propose a different method of analyzing soundness. An argument needs to meet three requirements to be cogent. Two of them have been mentioned above: valid reasoning, and acceptable premises and conclusion. The third one is that the argument scheme employed must be appropriate and correctly used. An argument scheme is a general scheme in which the premises and the conclusion being defended are linked together. van Eemeren et al. provide an example of how an argument scheme works. In the argument ‘Jack is an experienced teacher, because he spends hardly any time on lesson preparation.’, if the standpoint ‘Jack is an experienced teacher’ can be defended by the reason ‘he spends hardly any time on lesson preparation’, there needs another ‘unexpressed premise’ ‘little time spent on lesson preparation is characteristic of

experienced teachers.’ The unexpressed premise in this case is a specific case of a more general scheme, one thing (being an experienced teacher) is symptomatic of another (spending hardly any time on lesson preparation). The strength of the scheme is determined by a set of critical questions. These questions are context-dependent, that is, what should be criticized and how it should be done are dependent on the scheme. For example, in this case, the critical questions are ‘Aren’t there also other non-experienced teachers that spend hardly any time on lesson preparation?’ or ‘Aren’t there also experienced teachers that spend lots of time on lesson preparation?’. By answering these questions, the strength of the scheme is determined. This evaluation of the appropriate use of an argument scheme seems more objective than the scalar measure based on raters’ subjective judgments, as the critical questions lead to a dichotomous yes/no judgment while the scalar measure indicates the degree of an abstraction, such as weak or strong. However, it is not clear how these critical questions can be converted into descriptors of a rating scale. Therefore, though promising, the use of argument schemes as one criterion for evaluation of the soundness of an argument is not investigated in the study.

As the soundness of argumentation is more concerned with reasons or grounds and conclusion or standpoint, the quality of the claim or the standpoint that is part of an argument is often neglected. The claim or the standpoint is different from the conclusion as the claim or the standpoint is stated before grounds are given, while the conclusion is stated after the premises are given. Cerbin (1988, cited in Marttunen, 1994) emphasizes two essential aspects concerning the evaluation of claims, which are whether a claim includes a contention relating to some theme and whether a claim is written clearly.

The studies and theories reviewed above indicate that the content of argumentation is central to the evaluation of soundness of an argument, and the soundness of an argument can be evaluated through acceptability, relevance and sufficiency/adequacy of premises, and quality of a claim or a standpoint.

### 3.2.2.3 *Appeals of argumentation*

Still another important component of argumentation knowledge is different appeals that the writer makes to convince the audience. It is included in the model as it is related to the persuasiveness of the argumentative writing – an aspect of the quality of argumentative writing. Very few researchers have attempted to define appeals in a direct way, but for those who have tried, they seem to have been almost unanimously identifying appeals with Aristotle's three sources of persuasion: ethos, pathos, and logos (see in Connor & Lauer, 1985). According to Aristotle (1984, p. 2155, cited in Anthony & Gladkov, 2007), ethos 'depends on the personal character of the speaker', pathos 'depends on putting the audience into a certain frame of mind', and logos 'depends on the proof provided by the words of the speech itself'.

Connor and Lauer (1985) operationalize the persuasiveness into fourteen rational appeals, four credibility appeals and five affective appeals. These appeals were well explained in Anthony and Gladkov (2007, p. 125–32) and they are reproduced here. According to Anthony and Gladkov (2007, p. 125), rational arguments are made to appeal to 'the sensible and rational aspect of the reader's mind'. Rational appeals include the use of descriptive example (R1), narrative example (R2), classification (including definition) (R3), comparison (including analogy) (R4), contrast (R5), degree (R6), authority (R7), cause/effect-means/end-consequences (R8), model (R9), stage in process (R10), ideal or principle (R11), and information (facts, statistics) (R12). As shown in Table 3.2, the appeals of descriptive example (R1) and narrative example (R2) are similar in helping readers to infer a general conclusion from a typical specific compelling example, and make readers react to the appeal. The appeal of Classification (R3) places a person or a thing into a certain class and defines it. The appeals of Comparison (R4) and Contrast (R5) build a logical argument on the relations of similarity and difference. The appeal of Degree (R6) is not easy to understand; according to Aristotle (1932, p. 161, cited in Anthony and Gladkov, 2007, p. 126), Degree (R6) can be described by the following

example: if the less frequent thing occurs, then the more frequent thing would occur. The argument based on Authority (R7) relies on people's belief in the prestige of authoritative people, and their acts and speeches are imitated and approved. The appeal of Cause/Effect – Means/End – Consequences (R8) helps the writer to recommend on the reader's part by forecasting effects, consequences or ends because it commonly happens that a given thing has both good and bad consequences. For the appeal of Model (R9), Perelman (1982, cited in Anthony & Gladkov, 2007, p.127) consider it as providing a description of the way a proposed end can be achieved. An argument based on Stage in Process (R10) is used when a gap exists between the concept accepted by the audience and the proposal the writer is defending. Perelman (1982, p. 18) describes the appeal of Ideal or Principle, or Values (R11) as 'A convincing discourse is one whose premises are universalizable, that is acceptable in principle to all members of the universal audience'. When an argument is based on the appeal of Ideal or Principle, or Values, the argument is more easily acceptable by readers. For the appeal of Information (R12), the writer uses facts and statistics to establish the acceptability of his or her claim.

Table 3.2 The definitions of rational appeals (adapted from Anthony & Gladkov, 2007, p.124–5)

Rational appeals	
R1	Descriptive Example Using a compelling descriptive example from one's own or someone else's experience
R2	Narrative Example Using a compelling narrative example. Must contain a beginning, middle, and end of a story
R3	Classification Placing in a class or unit, and describing what that means
R4	Comparison Using comparison to support one's focus
R5	Contrast Using contrast to support one's focus
R6	Degree Arguing that two things are separated by a difference of degree rather than kind, or making an appeal for an incremental change
R7	Authority Using the authority of a person other than the writer

---

R8	Cause/effect-Means/End-Consequences how one event is the cause of another
R9	Model Proposing a model for action that relies on existing programs
R10	Stage in process Reviewing previous steps and looking forward to what steps need to be taken
R11	Ideal or Principle, or Values General knowledge
R12	Information Using supporting facts and statistics, description of reality

---

Table 3.3 presents the definitions of credibility appeals. According to Aristotle (1932, cited in Anthony & Gladkov, 2007), through credibility appeals, a trustworthy image of the speaker is established in the readers' mind. Through the trustworthy image, the discourse is more convincing. Credibility appeals include writer's first-hand experience (C13), writer's respect for audiences' interests and point of view (C14), showing writer-audience shared interests and points of view (C15), and writer's good character and/or judgment (C16). First-hand experience is used as a technique for providing information directly from the writer's experiences, thus establishing the writer's credibility, or giving the impression that the writer is knowledgeable and versed on the subject he/she is talking about. Writers' respect for audiences' interests and point of view helps to create the impression of a good-willed writer in the audience's mind. Writers' showing of shared writer-audience interest helps to build up solidarity with the audience by making the audience a part of it. The writer's good character and/or judgment helps to create a good image of the writer, in turn, making the writer more convincing.

Table 3.3 The definitions of credibility appeals  
(adapted from Anthony & Gladkov, 2007, p. 124–5)

---

Credibility appeals	
C13	First-hand experience Providing information to show first-hand experience or some authority on the subject
C14	showing writer's respect for audience's interests and point of view
C15	showing writer-audience shared interests and points of view
C16	showing writer's good character and/or judgment

---



Table 3.4 presents the definitions of affective appeals. According to Aristotle (1932, cited in Anthony & Gladkov, 2007), affective appeals are made to arouse the audience's emotions because people make very different decisions when under influence of pain or joy, liking or hatred; that is, the audience's decision about a proposed thing can be influenced by their positive or negative emotions. Affective appeals include appealing to the audience's emotional, attitudinal, or moral views or values using vivid pictures and charged language. Vivid pictures create the effect of placing a reader in the situation depicted by the writer. Charged language is usually used to arouse readers' emotions.

Table 3.4 The definitions of affective appeals (adapted from Anthony & Gladkov, 2007, p. 124–5)

Affective appeals
A17 Appealing to the audience's views (emotional, attitudinal, moral)
A18 Vivid picture
Creating a thought, a mind's eye vision.
A19 Charged language
Using strong language used to arouse emotions.

Connor and Lauer's (1985) appeals enable the writer to make a more convincing argument and understanding these appeals seems to be part of the knowledge a writer needs to have to make an effective argument in argumentative writing. Therefore, knowledge of appeals is tentatively included in argumentation knowledge and investigated in this study.

### 3.2.3 Developing the AWA model

Based on the discussion and literature review of previous research, I propose a model of different types of knowledge that are essential for a writer to have in order to write argumentative writing effectively. The model is presented in Table 3.5. The model comprises three essential knowledge components of language competence: grammatical, textual and sociolinguistic. As has been discussed in Section 3.2.1, grammatical

knowledge includes knowledge of written code (i.e., spelling, punctuation, paragraphing, and capitalization), knowledge of morphology (word-part knowledge), vocabulary and syntactic knowledge; textual knowledge includes cohesion, coherence and rhetorical structure; sociolinguistic knowledge includes idiomatic expression, cultural reference, figures of speech, and informal and formal style of writing.

Originating from rhetorical knowledge and genre knowledge in language and writing models, argumentation knowledge is developed as a separate and equally important knowledge component, like grammatical, textual, and sociolinguistic knowledge, by including insights from a different field: argumentation theory. The underlying philosophy is that the ability to argue should be an essential part of argumentative writing which should be assessed, and the criteria of assessment should be represented on a rating scale. Three subcomponents have been identified: structure, substance and appeals.

It is worth noting that Table 3.5 only presents part of the AWA model, which is comprised of language competence, argumentation competence, strategic competence and knowledge of the world. Strategic competence and knowledge of the world are similar to those described in Bachman and Palmer's model of communicative language ability. Since strategic competence and knowledge of the world is not a focus of testing and assessment, they are not presented in the table.

Table 3.5 The AWA model without strategic competence and knowledge of the world

Argumentative writing ability	Language competence	<ul style="list-style-type: none"> <li>Grammatical knowledge</li> </ul>	<ol style="list-style-type: none"> <li>Knowledge of the written code               <ol style="list-style-type: none"> <li>Spelling; b. Punctuation; c. Paragraphing; d. Capitalization</li> </ol> </li> <li>Knowledge of morphology (word-part knowledge)</li> <li>Vocabulary</li> <li>Syntactic knowledge</li> </ol>
-------------------------------	---------------------	---	---

		<ul style="list-style-type: none"> <li>• Textual knowledge</li> </ul>	<ol style="list-style-type: none"> <li>1. Cohesion</li> <li>2. Coherence</li> <li>3. Rhetorical organization</li> </ol>
		<ul style="list-style-type: none"> <li>• Sociolinguistic knowledge</li> </ul>	<ol style="list-style-type: none"> <li>1. Formal and informal writing style</li> <li>2. Idiomatic expression, cultural reference, and figures of speech</li> </ol>
		<ul style="list-style-type: none"> <li>• Argumentation knowledge</li> </ul>	<ol style="list-style-type: none"> <li>1. The structure of argumentation</li> <li>2. The substance of argumentation</li> <li>3. The appeals of argumentation</li> </ol>

### 3.3 An operational framework of argumentative writing ability

As has been argued, an operational framework will bridge the gap between a theoretical model and the operationalization of its components. In this section, an operational framework of argumentative writing ability is built on discourse analytic measures that are used to investigate the development/acquisition of the components of language competence and argumentation competence in second language acquisition and writing studies. In the following sections, empirical studies on the use of different measures of the components of the AWA model are reviewed. At the end of the chapter, a summary is given of suitable measures for the empirical investigation in this study. It is worth noting that the studies from which the measures of language competence are collected are not restricted to those in which argumentative writing is involved. The reason is that doing so may exclude potential measures at an early stage of this study. Furthermore, as studies reviewed are mostly focused on college-level ESL/EFL writers, and argumentative writing is a basic writing skill for college-level students, it is expected that argumentative writing is investigated in most of the studies. Therefore, measures which are found to be suitable based on these studies are assumed to be suitable for the investigation of argumentative writing in this study.

### 3.3.1 Grammatical knowledge

Grammatical knowledge (or linguistic competence) is composed of knowledge of written code, morphology, vocabulary, and syntactic knowledge. In second language acquisition studies, ESL/EFL learners' grammatical competence is evaluated through the analytical measures of complexity, accuracy, and fluency (Ellis & Barkhuizen, 2005). Complexity is defined as the extent of elaborateness and variety of language produced in performing a task (Ellis & Barkhuizen, 2005). Accuracy is characterized as error-free language performance (Foster & Skehan, 1996). Fluency is defined as spontaneous language production (Schmidt, 1992).

#### 3.3.1.1 *Syntactic complexity*

Wolfe-Quintero, Inagaki, and Kim (1998) investigate two aspects of complexity that are related to writing: syntactic complexity and lexical complexity. Syntactic complexity is related to how sophisticated and varied the syntactic structures are. Within second language acquisition, some research sets out to identify the measures of syntactic complexity that can characterize language development and language proficiency (e.g., Wolfe-Quintero et al., 1998; Ortega, 2003; Lu, 2011). Wolfe-Quintero et al. (1998) review 33 syntactic complexity measures in 39 ESL/EFL writing development studies. They find that mean length of T-unit (a syntactic structure developed by Hunt (1965) which contains one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it), mean length of clause (a structure defined by Hunt (1970) as containing a subject and a finite verb, which includes independent and dependent clauses), clauses per T-unit, and dependent clauses per clause are the most satisfactory as they are linearly and consistently associated with proficiency levels (e.g., program levels, school levels, holistic rating levels). They also caution that these measures discriminate poorly between adjacent levels of proficiency and that, statistically, relationships are only inconsistently found in studies where ESL/EFL proficiency levels are conceptualized using holistic rating levels. In order to identify the magnitude of the observed between-proficiency level

difference that can be tentatively expected to be statistically significant, Ortega (2003) compares the results of six measures of syntactic complexity that are most frequently used in 25 college-level ESL/EFL writing studies. He finds that between-levels significant difference, given samples of a medium size (of or about ten participants or more per cell), is likely to be 4.5 or more words per sentence for mean length of sentence (MLS), 2 or more words per T-unit for mean length of T-unit (MLT), slightly over 1 word per clause for mean length of clause (MLC), and at least a 0.20 positive or negative difference in number of clauses per T-unit for mean number of clauses per T-unit (C/T). In a meta-analysis study on the performance of syntactic complexity measures, Lu (2011) applies 14 measures (six from Ortega, 2003, five at least weakly-correlated with proficiency levels, and three recommended for further research from Wolfe-Quintero et al., 1998) to ESL argumentative writing samples from the Corpus of English Majors (Wen, Wang, & Liang, 2005). Lu (2011) found that MLC, MLT, MLS, coordinate phrase (including coordinate adjective, adverb, noun and verb phrases) per clause (CP/C), coordinate phrase per T-unit (CP/T), complex nominals (including nouns plus adjective, possessive, prepositional phrase, adjective clause, participle, or appositive, nominal clauses, gerunds and infinitives in subject) per clause (CN/C), and complex nominals per T-unit (CN/T) showed significant differences in discriminating four proficiency levels (i.e., indicated by school levels) and progressed linearly with increase of school levels. Lu (2011) also found that measures involving the same structure (e.g., MLS and MLT) were highly correlated, while measures involving different structures had moderate, weak or very low correlations (e.g., MLS and T/S (T-unit per sentence)). Mendelsohn (1983, cited in Lu, 2011) states that measures with low correlations are regarded as capturing different aspects of development.

Due to the mixed results, it is decided that two screening criteria should be in place to select the measures with the most potential rather than investigate all of them: measures to be selected should be different in nature (i.e., indicated by very low, weak and moderate

correlations), and the between-proficiency level difference for these measures should be practical to apply for raters (i.e., indicated by magnitude of the difference). Although Wolfe et al. (1998, p. 9) argue that the way that proficiency levels are conceptualized (e.g., program levels, school levels, holistic rating levels) may have an effect on whether a measure can be found successful in discriminating between proficiency levels, it was decided that this factor should not be taken into consideration as few studies conceptualize proficiency levels as is done in this. Based on these findings and the criteria listed above, mean length of clause (MLC), mean length of sentence (MLS), clauses per T-unit (C/T), complex nominals per clause (CN/C) and coordinate phrase per clause (CP/C) were selected and investigated in the pilot study.

### *3.3.1.2 Lexical complexity*

Lexical complexity is a construct of the richness of a writer's lexicon as manifested in writing and speaking (Wolfe-Quintero et al., 1998). The richness of a writer's productive vocabulary entails the use of a wide variety of basic and sophisticated words. A basic assumption in the exploration of lexical complexity in the acquisition of vocabulary knowledge is that the wider the variety of vocabulary shown in the writing, the more competent a writer is in using vocabulary they know.

Wolfe-Quintero et al., (1998) identify three dimensions of lexical richness: variation, sophistication, and density. Wolfe-Quintero et al. provide a comprehensive review of measures of lexical complexity. These measures are still widely used today, of which three general measures (as opposed to those involving particular word class, e.g., verbs) are most commonly used: word types per words (WT/W) (total number of different word types divided by the total number of words) for variation; sophisticated word types per word types (SWT/WT) (total number of sophisticated word types divided by total number of word types) for sophistication; and lexical words per words (LW/W) (total number of lexical words divided by total number of words) for density. Lexical words generally refer

to open-class words (i.e., nouns, verbs, adjectives, adverbs), as opposed to closed-class grammatical words (i.e., determiners, numerals/quantifiers, pronouns, prepositions/particles, conjunctions).

Word types per word (WT/W; also termed Type-Token ratio) was developed by Templin (1957), but it has been criticized for penalizing writers who write longer texts. WT/W decreases as text length increases. A number of methods have been developed to take sample length/size into consideration and improve WT/W. Complicated mathematical transformations have been proposed to improve WT/W (e.g., Root WT/W), however, these measures are not applicable for human raters. Other methods proposed include using samples of equal length and samples produced under a time limit (e.g., Wolfe-Quintero et al., 1998). Thordardottir and Ellis Weismer (2001, cited in Lu, 2012) suggest truncation of sample length to a set length. Malvern et al. (2004, cited in Lu, 2012) suggest a random selection of a standard number of words or selection of a standard number of consecutive words from the sample with a random starting point. However, they also suggest that the truncation of sample length should be used with caution because different sampling methods produce different results. As for samples produced under a time limit, the length problem seems to remain as samples produced under a time limit are still likely to be of different lengths. The problem of length sensitivity is still not solved (personal communication with Glenn Fulcher, 2015).

The sophisticated or advanced words have been defined as those occurring less frequently in a word frequency list (e.g., Lauer, 1994) or words that learners are exposed to at a more advanced school level in an education system (e.g., Linnarud, 1986). Laufer (1994) analyzes four different lexical sophistication measures (sophisticated word types/word types) on pre- and post-compositions written by two advanced university classes. In two of the analyses, she defines sophisticated words as those not on a 2000 word frequency list, and words on a university-level word list (the University World List (UWL), Xue &

Nation, 1984), finding a significant difference for two university classes. In the other two analyses, she defines sophisticated words as those not among the first 1000 most-frequent words, and words not on any of these frequency lists (i.e., the 1000 word frequency list, the 2000 word frequency list, and the UWL), finding no significant difference. Lu (2012) reports no significant difference for the same measure between four levels of oral narrative proficiency of Chinese TEM4 test takers. Lu (2012) defines sophisticated words as those not among the 2000 most-frequent words in the BNC (British National Corpus) word list (Leech, Rayson, & Wilson, 2001). The mixed results may be due to different definitions of sophisticated words. For this study, it would be preferable for there to be a word frequency list that is developed from a large-scale corpus of English texts that Chinese EFL learners are exposed to. However, developing a word frequency list is out of the scope of this study. Therefore, Laufer's (1994) frequency lists were chosen for investigating lexical sophistication in this study as these frequency lists are utilized by a computer program and by using the program the lexical sophistication can be calculated conveniently.

Studies reviewed by Wolfe et al. (1998) suggest that lexical density might not be a good measure of language development as no significant difference was found in this measure between the writing of native speakers and non-native speakers (e.g., Linnarud, 1986) and between writing with different holistic ratings in these studies (e.g., Engber, 1995). A recent study by Lu (2012) looks at this measure in relation to the target population of the present study and also reports no significant difference between four narrative oral proficiency levels of Chinese TEM4 takers. Although no significant results are found for lexical density, since the studies reviewed above either do not involve a similar target population to this study or are not focused on writing proficiency, it is too early to exclude the measure of lexical density from this study. Therefore, lexical density is selected for investigation in this study.



Therefore, measures of lexical complexity to be investigated in the pilot study include sophisticated word types per word type (SWT/WT) and lexical words per word (LW/W). Sophisticated words are defined here as those not on the 2000 most frequent word list (sophisticated word type I), that is those on the UWL and off-list words and those on the UWL (sophisticated word type II).

### 3.3.1.3 *Accuracy*

Accuracy is concerned with errors in language use. The analysis of the accuracy of language use usually resorts to counting the errors in a text. Lennon (1991, p. 182, cited in Polio & Shea, 2014) defines an error as ‘a linguistic form or combination of forms which, in the same context and under similar conditions of production would, in all likelihood, not be produced by the speakers’ native speaker counterpart’. However, identifying errors is problematic. Disagreement arises when decisions need to be made on writers’ intended meaning of incorrect sentences, native speaker counterparts’ choice of words or structures in a certain context, and the appropriacy of language use (e.g., Polio & Shea, 2014). Reliability of error coding and explicit coding schemes are expected to help interpret research results more meaningfully (reliability of accuracy measures) (Polio, 1997).

Wolfe-Quintero et al. (1998) provide a comprehensive view of accuracy measures in second language acquisition studies. Among these measures, error-free T-unit per T-unit (EFT/T) and errors per T-unit (E/T) have been shown to be the best measures that can capture language development and language proficiency. Wolfe-Quintero et al. (1998) suggest error-free clause per clause EFC/C and errors per clause (E/C) for further study. Other measures which involve specific error types (e.g., article, preposition, lexical errors) and error severity are not investigated in this study as it is too trivial to focus on a specific error type in a language scale and error severity is heavily dependent on subjective judgments. Polio and Shea (2014) show that there is a strong correlation between EFT/T

and EFC/C and weighted error T-unit ratio (a measure which takes into account error severity) when they are used on college-level ESL/EFL argumentative writing. According to Mendelsohn (1983) (see Section 3.3.1.1), the three measures which are shown to be strongly correlated can be regarded as capturing the similar aspect of development.

Based on the findings, EFT, E/T, and E/C are selected for the investigation in the current study.

#### 3.3.1.4 *Fluency*

Fluency, according to Koponen and Rigenbach (2000), can be defined in different ways. It may refer to the smoothness of speech in terms of temporal features and it might refer to the automatization of psychological processes. To assess this multi-faceted nature of fluency, researchers have developed a number of measures. In the context of speech research, Skehan (2003) classifies these measures and relates them to four sub-dimensions: breakdown fluency, repair fluency, speech rate and automatization. Breakdown fluency has been measured by silence; repair fluency by reformulation, replacement, false starts, and repetition; speech rate by the number of words/syllables per minute; automation by the length of run. In the context of writing research, these four sub-dimensions of fluency are also applicable to the assessment of writing: writers' pausing (breakdown fluency) (Miller, 2000); the frequency of revisions (repair fluency) (Chenoweth & Hayes, 2001; Knoch, 2009); composing rate (the number of words in a time limit) (Sakyi, 2000; Wolfe-Quintero et al., 1998); and length of bursts of think-aloud protocols between pauses (automatization) (Chenoweth & Hayes, 2001). Kaufer, Hayes, and Flower (1986) define bursts of think-aloud protocols as language parts which are proposed and evaluated in think-aloud protocols and also presented in a text. They hypothesize that the longer the bursts of language parts or proposed text are, the less time writing takes, thus the more automatic the writing skill is. Wolfe-Quintero et al. (1998) also consider a number of length measures (e.g., length of production units, error-free

production units, or complex structures), which are traditionally considered as complexity measures and accuracy measures, as measures of fluency.

Chenowith and Hayes (2001) find that students display a significant increase in writing fluency within a difference of two semesters of language learning. The increase of writing fluency was represented by an increase in burst length (automatization), a decrease in the frequency of revisions (repair fluency), and an increase in the number of words accepted and written down (writing rate). Wolfe-Quintero et al. (1998) find that the number of words in a composition with a time limit show a significant relationship between different proficiency levels. They also find that T-unit length, error-free T-unit length, and clause length consistently increase in a linear relationship with proficiency level across studies, regardless of whether proficiency was defined as program level, school level, or holistic ratings. Knoch (2009) find that the number of revisions and the number of words show a significant difference between different proficiency levels.

In this study, the number of words (composing rate) and the number of revisions (repair fluency) are investigated in the pilot study as fluency measures. Breakdown fluency (e.g., writing pauses) and automatization (e.g., length of bursts) are not investigated as they are more concerned with the writing process than with written products.

#### *3.3.1.5 Mechanics*

Mechanics has been quantified in terms of spelling, punctuation, capitalization and indentation (Polio, 2001), the order of punctuation marks (Mugharbil, 1999), and the number of paragraphs (Kennedy & Thorp, 2002). In a review of text studies on various features of second language writing, Polio (2001) points out that most studies that have quantified mechanics to date (e.g., Pennington & So, 1993; Tsang, 1996) have made use of the Jacob et al.'s ESL Composition Profile: content (Jacobs et al., 1981). In their scale, mechanics are measured by the frequency of spelling, punctuation, capitalization and

indentation errors.

Polio (2001) further points out that it is not clear whether mechanics is a construct, as little evidence can be found to support the view that different aspects of mechanics (e.g., punctuation, spelling) are related. This may explain why, in some widely used scales, there is no mechanics. For example, in the IELTS analytic scale for academic writing (2005), spelling is encapsulated in the subscale of lexical resource, and punctuation in the subscale of grammatical range. Polio (2001) also points out that there seems to be the lack of a theoretical basis for, and little interest in, measuring mechanics in ESL/EFL writing research. For example, in studies exploring accuracy, spelling is often disregarded.

Other studies which involve measures of knowledge of written code focus on the developmental aspect of mechanics rather than their accurate use (Mugharbil, 1999; Kennedy & Thorp, 2002, cited in Knoch, 2009). Mugharbil (1999) investigates the order of punctuation marks acquired by second language learners and finds that the period is the first question mark acquired and the semi-colon the last. But it seems that the developmental aspect is less relevant to the study than its correct use, as advanced learners of language are expected to master punctuation and their use of certain punctuation is more a matter of personal style than a proficiency issue. Kennedy and Thorp (2002) compare the number of paragraphs produced by IELTS takers at levels 4, 6 and 8, but no conclusive result is found.

Although there is little evidence to support the view that mechanics should be measured as a separate construct in rating scales, it is still used in many scales, especially in the current TEM4 and TEM8 scales. It can thus be assumed that the correct use of the written code is part of the test performance that is expected from a high-level language learner. The same seems true for the number of paragraphs. Despite Kennedy and Thorp's (2002) results, the number of paragraphs was investigated in the current study as it could indicate

the ability to organize different argumentative elements in different paragraphs. That is, writers with different proficiency levels can produce a varying number of paragraphs on different elements of argumentative writing (e.g., introduction, conclusion, reasons). The hypothesis is that more proficient writers will develop introduction, first-level reasons (see Appendix 5 for definition) and conclusion into separate paragraphs, while less proficient writers will combine one or two or some of these elements into a differing number of paragraphs, for example, introduction, main body, and conclusion, or introduction, reason 1, reason 2, and conclusion.

However, this measure is limited as it fails to penalize very short paragraphs, unnecessary paragraph breaks, and ordering of information (Knoch, 2009, p. 176). This measure also needs to be used with caution as students may be encouraged to mechanically write more paragraphs to score high while neglecting unnecessary paragraph breaks.

Mechanics, therefore, is measured in the pilot study through the correct use of punctuation, spelling, capitalization and the number of paragraphs.

### **3.3.2 Textual knowledge**

As shown in the AWA model in Table 3.5, textual knowledge includes coherence, cohesion and rhetorical organization, also termed top-level argumentative writing structures in this study. Top-level argumentative writing structure in this study refers to arrangements of argumentation in relation to paragraphs. That is, the distribution of argument elements across paragraphs. I invented this measure for the analysis of the rhetorical structure of argumentative writing in the data analysis for this study. The evaluation of textual competence focuses on the evaluation of cohesion, coherence, and top-level argumentative writing structure.

### 3.3.2.1 Cohesion

Cohesion is mainly identified with Halliday and Hasan's (1976) theory of cohesion. In their work, cohesion involves ways in which interpretation of one element is dependent on another in a text. That is, when one element in a sentence presupposes another element in another sentence, cohesion is set up between these two sentences. Through cohesion the information in a text is organized and a unified whole is formed. A single occurrence of cohesion is termed a cohesive tie (Halliday & Hasan, 1976, p. 3). The presupposing element is termed a cohesive element by Halliday and Hasan. Halliday and Hasan identify five types of cohesive device: *reference*, *substitution*, *ellipsis*, *conjunction*, and *lexical cohesion*.

*Reference* refers to the semantic interpretation of one item by making reference to another in the surrounding text or in a situation. The reference made to an item identified in a situation is exophora and the reference made to an item identified in surrounding text is endophora. Exophora does not contribute to cohesion. The endophoric reference includes two types: anaphoric (a reference to an item in preceding text) and cataphoric (a reference to an item in the following text). There are three types of reference items: personal pronoun (e.g., 'Harry Potter is my favorite book. *It* teaches me the importance of friendship.'), demonstrative pronoun (e.g., 'I visited my grandparents last year. *That* was pleasant.'), and comparative reference (e.g., 'Mr Thomas went to London. His father went even *further*').

*Substitution* refers to the replacement of a word or a phrase by another item with the same structural function or at the same lexico-grammatical level (Halliday & Hasan, 1976). Substitution is normally anaphoric and rarely cataphoric. There are three types of substitution: nominal (e.g., 'I bought a dictionary yesterday. Did you buy *one* too?'), verbal (e.g., 'Can you help me with my math please?-Yes, I can *do* it.'), and clausal (e.g., 'It seemed that everyone was happy. I would say it seemed *not*.'). Substitution is more

frequent in spoken texts.

*Ellipsis* refers to the omission of an item normally required by the grammar which can be understood from the preceding linguistic context (occasionally the following one) and which therefore need not be repeated. Ellipsis is normally anaphoric and rarely cataphoric. There are three types of ellipsis: nominal, verbal and clausal. Ellipsis can be regarded as ‘substitution by zero’ (Halliday & Hasan, 1976, p. 142). Like substitution, ellipsis is more common in spoken texts (e.g., ‘What are they doing now?-Preparing exams.’)

*Conjunction* refers to semantic relations expressed through conjunctive adjuncts connecting two or more segments in texts. Halliday and Hasan (1976, p. 242–3) summarize common conjunctive adjuncts in a list. Examples of conjunction are additive (e.g., *and, additionally*), causative (e.g., *therefore, hence*), temporal (e.g., *then, next*), adversative (e.g., *but, however, rather*).

*Lexical cohesion* refers to the cohesive effect achieved through selecting certain vocabulary items in relation to lexical items occurring earlier in the context. The relations between two cohesive lexical items include repetition (e.g., *bears, bear; leave, left*), synonym/near-synonym (e.g., *fall, drop*), hyponym (e.g., *flower, rose*), meronym (e.g., *class, boys*), antonym (e.g., *boys, girls*), general words (e.g., *thing, person*), collocation (e.g., *doctor, ill*) (Halliday & Hasan, 1976; Hasan, 1984; Halliday, Matthiessen, & Matthiessen, 1994/2014).

Halliday and Hasan’s (1976) categories of cohesion have been applied in a number of studies. Mixed results are found on the relationship between the quality of writing or program levels with cohesive devices. Witte and Faigley (1981), comparing the cohesion of high- and low-level essays composed by L1 freshmen, found that there was a higher density of cohesive ties and cohesive ties spanned shorter distances in high-level essays

than in low-level essays. They also found that lexical cohesion was the most frequently used cohesive tie – about two thirds in both high and low-level essays – but the specific type of lexical cohesive ties used varied. Low-level essays relied more on repetition while lexical collocations used in high-level essays were triple those used in low-level essays. In contrast, Neuner (1987) found that none of the ties were used more in high-level essays than in low-level essays. However, he found a difference in lexical cohesion between good essays and poor essays. In good essays, cohesive chains were sustained over greater distances, and more different words as well as less frequent words were used in the lexical chains. Jafapur (1991) investigated the cohesion of ESL writing. He found that the number of cohesive ties and the number of different types of cohesion successfully discriminated between different proficiency levels. Kennedy and Thorp (2002) found that IELTS writers at levels 4 and 6 used markers like ‘however’, ‘firstly’, ‘secondly’ and subordinate conjunctions more than writers at level 8. Banerjee and Franceschina (2006) investigated the use of demonstrative reference over five different IELTS levels. They found that the use of ‘this’ and ‘these’ increased with proficiency levels while the use of ‘that’ and ‘those’ stayed relatively constant or decreased.

Some recent studies investigated the cohesion of Chinese EFL students’ writing. Liu and Braine (2005) explored the use of reference, conjunction and lexical cohesion in argumentative essays composed by Chinese EFL college students. They found that the essay scores significantly co-varied with the total number of cohesive devices and were highly correlated with lexical devices in particular. They also identified a number of major types of problems with reference devices and lexical cohesion in the writing. They found that Chinese EFL writers tended to shift the use of singular and plural pronouns, omit or misuse the definite article, underuse comparatives and overuse the phrase ‘more and more’. They also found that writers use a restricted choice of lexical items, and misuse collocation. However, they did not compare the difference in cohesion between different proficiency levels. Yang and Sun (2012) investigated the (correct) use of reference,



conjunction, substitution, ellipsis, and lexical cohesion in year-2 and year-4 Chinese college English majors' writing. They found significant differences and a large effect size in the (incorrect) use of reference and lexical cohesion between the year-2 and the year-4 college students. They also found no significant relationship between the total number of cohesive devices and proficiency levels in Chinese EFL writing. However, Zhang (2000) found no significant relationship between the number of cohesive devices and writing quality in Chinese EFL writing.

Based on these findings, the number of references, lexical chains, conjunctions and incorrect use of cohesive devices are selected for the analysis in the pilot study.

### 3.3.2.2 *Coherence*

Coherence, according to Grabe and Kaplan (1996), can be approached from the perspective of the writer and the reader. From a writer perspective, ideas are organized in a text to convey meaning. de Beaugrande and Dressler (1981, cited in Hoey, 1991, p. 11), consider coherence as 'the configuration of concepts and relations that underlie the surface text are mutually accessible and relevant'. From a reader perspective, Yu (1996, cited in Watson-Todd, Khongput, & Darasawang (2007) views coherence as 'less intangible ways...which reside in how people interpret texts rather than in text itself'. According to these two perspectives, coherence can be viewed as both a textual feature and the reader's expectation of upcoming textual information. In this study, the reader's perspective of the coherence is not investigated as it is not easy to convert readers' interpretations of coherence to scale descriptors.

Several analytical models and abundance of measures have been developed to operationalize coherence, however, only a few are useful in this study. Therefore, only models and measures that can be operationalized for a rating scale are reviewed. Topical structure analysis (TSA) and metadiscourse markers are selected and are introduced in

the following.

### *Topical structure analysis (TSA)*

TSA was developed by Lautamatti (1978) for the analysis of topic development in written texts and further extended by Connor and Schneider (1990). In TSA, a written text is considered as progressions of topics (or themes) (i.e., what a sentence is about) and comments (or rhemes) (i.e., what is said about the topic) between sentences. Researchers using TSA analysis believe that coherence is achieved through semantic relations (i.e., exact repetition, synonymous relation) between the topic of one sentence and the topic or the comment of another. Lautamatti (1978) describes three types of progression: parallel progression, sequential progression, and extended parallel progression. Parallel progression refers to a number of successive sentences of which topics are 'the same' (p. 73) or have 'the same referent' (p. 82). It is illustrated as  $\langle a, b \rangle$ ,  $\langle a, c \rangle$ ,  $\langle c, d \rangle$ , with 'a' representing the same topic of three sentences and 'b', 'c' and 'd' three different comments of these sentences. For example, in two consecutive sentences extracted from Lautamatti (1978, p. 78) '(3) New-born infants are completely helpless. (4) They can do nothing to ensure their own survival.' the topic 'New-born infants' in sentence 3 and 'They' in sentence 4 have the same referent. Sequential progression refers to the relation of a minimal pair of two sentences, in one of which 'the predicate or rhematic part' (1978, p. 73) provides the topic for the next. It is illustrated as  $\langle a, b \rangle$ ,  $\langle b, c \rangle$ ,  $\langle c, d \rangle$ , with 'b' representing the comment of the first sentence and the topic of the second sentence, and 'c' representing the comment of the second sentence and the topic of the third sentence. For example, in two consecutive sentences extracted from Lautamatti (1978, p. 78) '(5) They are different from young animals. (6) Young animals learn very quickly to look after themselves.', 'young animals' in the rhematic part of sentence 5 occurs at the theme position in sentence 6. Extended parallel progression refers to the relation in which a topic of a preceding sentence is resumed after a sequential type of progression. It is illustrated

as  $\langle a, b \rangle$ ,  $\langle b, c \rangle$ ,  $\langle a, d \rangle$ , with 'a' the topic of the first sentence resumed at the topic position of the third sentence, and the comment of the first sentence occurring at the topic of the second sentence. For example, the topic 'a child' in sentence 4 'Without care from some other human being or beings, be it mother, grandmother, sister, nurse, or human group, a child is very unlikely to survive.' is resumed in the topic 'the human infant' in sentence 8: 'It is during this very long period in which the human infant is totally dependent on others that it reveals the second feature which it shares with all other undamaged human infants, a capacity to learn language.' 1978. p. 78). Connor and Schneider (1990) further extended TSA by dividing sequential progression into three sub-categories according to different semantic relations involved in sequential progression: directly-related, indirectly-related, and unrelated. Directly-related sequential progression includes comment-topic relation (i.e., the comment of the previous sentence becomes the topic of the following sentence), word derivations, and part-whole relations. Indirectly-related sequential progression involves topics related by semantic sets (i.e., scientists, their inventions and discoveries, and the invention of radio, telephone and televisions). Unrelated sequential progression refers to sentential topics which are not clearly related to either the previous sentence topic or the discourse topic. Simpson (2000, p. 301) adds extended sequential progression. This refers to the comment or the topic of a previous sentence being taken up as the topic of a non-consecutive sentence after an interruption of a number of topics.

A number of studies use TSA to explore the relationship between proficiency levels and the use of different types of topic progression. The results of these studies are slightly different. High-level writers tend to use more extended parallel progression and sequential progression, while low-level writers tend to use more parallel progression (Schneider & Connor, 1990). Knoch (2007) investigated the types of topical progression in the analysis of 602 ESL/EFL test essays of five levels, and found that higher level writers tended to use more direct sequential progression, indirect progression, and

superstructure (e.g., linking devices). Low-level writers tended to use more coherence breaks, unrelated sequential progression, and parallel progression.

### *Meta-discourse markers*

Meta-discourse markers are a set of linguistic devices which function to guide and direct readers in their process of understanding the content and the writer's attitude to the content of the text (Crismore, Markkanen, & Steffensen, 1993). Metadiscourse is a term used to describe how writers use language means (e.g., words, phrases, or sentences) to help readers interpret, evaluate, and react to the propositional content they write (Vande Kopple, 1985; Crismore, 1989; Hyland, 2005) and it is often referred to as "discourse about discourse". Metadiscourse markers have two general functions: textual and interpersonal (Crismore et al., 1993). Hyland (2005) provides more detailed definitions of these two functions, though with different terms (i.e., renaming textual and interpersonal functions as interactive and interactional functions). Textual metadiscourse markers, according to Vande Kopple (1985, cited in Hyland, 2005, p. 26), are used to indicate 'how we link and relate individual propositions so that they form a cohesive and coherent text and how individual elements of those propositions make sense in conjunction with other elements of the text'. Interpersonal meta-discourse markers, according to Lyons (1977, cited in Hyland, 2005, p. 26), are used to 'help us express our personalities and our reactions to the propositional content of our texts and characterize the interaction we would like to have with our readers about that content'. Metadiscourse markers are in various linguistic forms from a word (e.g., *however*), to a full sentence (e.g., *Here are the reasons.*).

Textual metadiscourse markers include logical markers, sequencers, reminders, topicalisers, code glosses, illocutionary markers and announcements (Crismore et al., 1993). Logical markers mark additive (e.g., *and*, *in addition*), adversative (e.g., *but*,

*however*) and conclusive (e.g., *finally*, *overall*) relationships between discourse parts. Sequencers express the order relationship (e.g., *in the first place*, *secondly*) between different discourse parts. Reminders refer back to previous sections in the text to retake an argument, expand it or summarize previous argumentation (e.g., *as has been mentioned before*). Topicalizers indicate a topic shift to the reader (e.g., *as for*, *with regards to*). Code glosses help readers grasp the intended meaning of propositional content by elaborating or redefining it (e.g., *that is*, *namely*, *for example*). Illocutionary markers explicitly name the act the writer performs (e.g., *I am going to explain this idea*). Announcements refer forward to future sections in the text in order to prepare the reader for upcoming ideas (e.g., *In the following sections*).

Interpersonal metadiscourse markers include hedges, certainty markers, attributors, attitude markers, and commentary. Hedges allow writers to withhold full commitment to the truth-value of their statements through the use of epistemic verbs (e.g., *may*, *might*), probability adverbs (e.g., *probably*, *perhaps*) and epistemic expressions (e.g., *it is likely*). Certainty markers allow writers to express total commitment to the truth-value of the text (e.g., *undoubtedly*, *clearly*). Attributors refer to the source of information (e.g., *As X said*). Attitude markers allow the writer to express affective values towards text and readers through deontic verbs (e.g., *have to*, *must*), attitudinal adverbs (e.g., *unfortunately*), attitudinal adjectives (e.g., *It is absurd.*), and cognitive verbs (e.g., *I think*, *I feel*). Commentaries help establish reader-writer relations through texts. It includes rhetorical questions (e.g., *what is the difference between A and B?*), direct address to readers (e.g., *you must know*), inclusive expressions (e.g., *we all believe*, *let us summarize*), personalization (e.g., *I do not want*), asides (e.g., *Diana (ironically for a Spencer) was not of the Establishment*). In addition to linguistic forms, some punctuation markers, according to Crismore et al. (1993), also serve the function of metadiscourse markers. A colon, a comma, an underlining, parentheses, and brackets fall into code glosses. Exclamation marks, underlining, and capitalization fall into attitude markers.

An assumption for research into students' knowledge of metadiscourse is that a good command of knowledge of metadiscourse leads to a reader-based discourse. A reader-based discourse implies the writer's awareness of the needs of the reader and written productions more accommodating to the reader's better understanding. In turn, a reader-based discourse contributes to a high overall writing quality.

Studies show that a higher frequency, and a greater variety of use, of metadiscourse makers is attained by higher proficiency writers than lower proficiency ones (Intaraprawat & Steffensen, 1995; Cheng & Stephensen, 1996; Basturkmen & Randow, 2014; Tan & Eng, 2014). Intaraprawat and Steffensen (1995) investigated the use of metadiscourse markers in essays written by university ESL students. They found that good EFL essays produced a higher density of metadiscourse markers (i.e., number of metadiscourse markers per T-unit), a higher variety of metadiscourse markers within each category, and fewer errors in meta-discourse markers than poor EFL essays. Cheng and Stephensen (1996) investigated the effect of instruction of metadiscourse on the quality of essays through an experiment and found that essays produced by the experimental group scored significantly higher than essays produced by the control group. An in-depth qualitative analysis of the high scoring essays shows that the improvements can be attributed to the use of metadiscourse markers, which make the texts more accommodating towards readers. Basturkmen and von Randow (2014) examined textual metadiscourse markers in 10 higher-graded and 10 lower-graded postgraduate essays and found slightly more instances of textual metadiscourse in the higher-graded writing samples than in the lower-graded writing samples, but no significant difference was found. Tan and Eng (2014) examined Hyland (2005)'s metadiscourse markers in Malaysian College Students' writing and found that higher proficiency Malaysian college writers exhibited a higher frequency of use of, and a greater variety of, metadiscourse markers than low proficiency writers.

An additional aspect of coherence has been identified by Knoch (2007): coherence breaks. Knoch defines this as when an attempt at coherence fails because of an error, <a, b> <failed attempts at a, or b or linker, c>. She illustrates a coherence break in an example, in ‘The reasons for the change in the graph. *It’s* all depends on their personal attitude’ (p. 121), ‘it’ fails to refer to the topic or comment in the previous sentence because of the number of pronouns. The underlying principle is that the more the coherence breaks, the less coherent the passage is. However, the coherence break that is caused by an error in reference use (e.g., it) overlaps with incorrect use of reference as a cohesive device investigated in cohesion. Since reference has been investigated in cohesion, coherence breaks caused by cohesive devices are investigated in cohesion analysis. Therefore, Knoch’s (2007) coherence breaks are adapted as errors in metadiscourse markers.

Based on the findings above, the measures of coherence include the proportion of topical progression patterns, number of metadiscourse markers, and number of errors in metadiscourse markers.

### 3.3.2.3 *Top-level argumentative writing structures*

Top-level argumentative structure in the current study refers to the organization of argumentative elements in relation to paragraphing. Another term that relates to the organization of argumentative writing is superstructure. This refers to ‘the organization plan of any text and... the linear progression of the text’ (Connor & Lauer, 1988, p. 142) and has been investigated in written texts analysis (e.g., Hoey, 1994). The underlying theory of the superstructure relates to the linear cognitive process of problem-solving when readers approach different written texts. Kummer (1972, cited in Connor & Lauer, 1988) identifies a situation-problem-solution-evaluation linear structure in argumentative writing. In each of these parts, background materials are provided that oriented readers to the problem, an undesirable state, responses to the undesirable state, and the evaluation

of the outcome of the suggested solution. Connor and Lauer (1988) conducted a superstructure analysis of the persuasive writing of high school native speakers of English, and find no significant difference in the presence of four parts of the superstructure at three levels of writing quality, and no evidence in the presence of the evaluation part in the average performance of writing at three levels. Instead of the situation-problem-solution-evaluation linear structure, this study treats superstructure as argumentative moves in relation to paragraphing because I believe argumentative moves are more accurate to capture the textual structure of argumentative writing. A move is a section of a text that performs a specific communicative function (Swales, 1990). Argumentative moves include introduction, myside argument (the writer's main standpoint and its justification), counterarguments, rebuttals, and conclusion.

This study tentatively investigates the arrangements of argumentative moves in relation to paragraphs. It is expected that the argumentative moves can more precisely capture the structure of argumentative writing than linear progression plan. The underlying assumption is that there are variations between students at different writing levels in arranging argumentative elements at the paragraph level. That is, in essays of low quality, moves would be squeezed into paragraphs disproportionately. This indicates poor knowledge of the organization of paragraphs in accordance with the argument structure. On the contrary, in the essays of high quality, most paragraphs would be devoted to the justification of the author's standpoint, which may indicate a more in-depth justification. The presumed measure for top-level structure for argumentative writing is the number of paragraphs per argumentative move.

### **3.3.3 Sociolinguistic knowledge**

The evaluation of the quality of students' writing from the sociolinguistic perspective is mainly focused on awareness of the difference between formal and informal register, natural or idiomatic expressions, cultural reference and figures of speech (see Section



3.2.1).

The key aspect of register knowledge that can be assessed is style (including frozen, formal, consultative, casual and intimate). Little research has been conducted on the assessment of register knowledge. Bachman and Palmer (1996) develop a rating scale for knowledge of register in the assessment of telephone company employees' writing ability (see Table 3.6). The scale distinguishes between four levels of awareness of distinctions between formal and informal registers: zero, limited, moderate, and complete. It can be implied that register knowledge is mainly assessed from the perspective of formality (style).

Table 3.6 Moderate level of a rating scale of register knowledge (Bachman & Palmer, 1996, p. 288)

Moderate	Moderate knowledge of register  Range: moderate distinction between formal and informal registers  Accuracy: good, few errors
----------	---

In second language acquisition, the distinction between spoken English and formal academic style is often studied. As has been mentioned in Section 3.2.1, this study views this type of distinction (register proficiency) as expected of EFL college students. Shaw and Liu (1998) explored the developmental changes in language use in 144 entry EFL postgraduates taking a summer course of English for Academic Purpose in a British university. Their research shows significant changes, from the features of spoken English to those more typical of formal writing. These features include personal pronouns (e.g., *I, me, my, we, us, it, you, your*), contractions (e.g., *he's, it's*), forms of metadiscourse markers (personal pronoun + active verb, e.g., *I conclude/As a conclusion, I can see*; passives, e.g., *it can be concluded that*; non-finite, e.g., *To conclude*), colloquial words (e.g., *a bit, a lot, lots, thing, nice, big, little*), and use of 'because' for clausal causes (i.e., 'knowledge-base' use of 'because' rather than causal use). Schleppegrell (1996, cited in Shaw & Liu, 1998)

argues that the use of because for indicating the ground ('knowledge base') for the preceding claim rather than the cause of something is characteristic of the spoken language.

A review of studies on natural or idiomatic expressions shows that this knowledge is part of superordinate knowledge which is termed in a number of confusing ways: phraseology, fixed expressions, prefabricated chunks, formulaic language (Howarth, 1998), or multi-word lexical items (Read, 2000). All of these terms are concerned with any linguistic forms with more than two words (e.g., phrases, sentences). According to Howarth (1998), word combinations can be graded on a collocational continuum in terms of openness and interpretation. On this continuum, idiomatic expressions comprise figurative idioms and pure idioms. Figurative idioms have metaphorical meanings and literal interpretation (e.g., *under the microscope*). Pure idioms have a unitary meaning which cannot be derived from the meanings of the components and are the most opaque and fixed category (e.g., *under the weather*). Both these two types allow no substitution of any component within the items. Quantitative measures of idiomatic expressions are mainly derived from a corpus-based approach (e.g., Bestgen & Granger, 2014) and idiom tests (McGavigan, 2009, cited in Milton, 2009) and collocation tests (Gyllstad, 2007, cited in Milton, 2009). These measures include corpus-based complex computational indices, idiom and collocation test scores. None of them are suitable for this study as computational indices cannot be interpreted by human raters and test scores cannot be transferred to scale descriptors. A review of studies show that few studies have focused on the knowledge of cultural references and figures of speech in second language acquisition.

Therefore, the measures of sociolinguistic competence are restricted to measures of register, which are mainly identified with those used in Shaw and Liu (1998), and include the number of personal pronouns, contractions, informal metadiscourse markers, colloquial words, 'knowledge-base' use of 'because'.

### **3.3.4 Argumentation knowledge**

As discussed in Section 3.2.2, argumentation knowledge is assumed to consist of structure, substance, and appeals of argumentation. In this section, studies on the relationship between the quality of writing with the structure, soundness, and persuasiveness of argumentation are investigated. Measures which can be used for the analysis of argumentation in the present study are selected.

#### *3.3.4.1 The structure of argumentation*

It was proposed in Section 3.2.2.1 that Chase's (2011) model of argumentative elements was the most comprehensive to date and was suitable to account for all the structural aspects of argumentative writing. Therefore, Chase's (2011) model is selected to operationalize structural aspect of argumentative writing and is introduced below.

Chase's (2011) model includes a number of argumentative elements: (1) Introduction (I) is a foreshadow of what is to follow in the writer's presentation of the argument and may also outline the writer's purposes or goals; (2) Conclusion (C) offers a closing at the end of an essay to what is written; (3) Standpoint (SP/SN) describes the writer's belief or opinion about a controversial topic; (4) Reasons (R) answers the questions "why" the writer holds a certain standpoint; (5) Coordinative reasons (R1a/b) are dependent on one another to defend a standpoint; (6) Subordinative reasons (R1.R1) consist of a series of reasons where one reason represents a standpoint for the following reason; (7) Convergent reasons (R1, R2) consist of more than one reason for the same standpoint; (8) Alternative standpoint (AS) is directly opposed to the writer's stated standpoint; (9) Counterargument (CA) is a criticism or objection could be used to undermine a person's standpoint; (10) Rebuttal (RB) is a statement that refutes, weakens or undermines an alternative standpoint, or counterarguments; (11) Reasons for rebuttal (RB.R) support a rebuttal; (12) Non-functional Unit (NF) include: repetitions (NFR), other information that does not appear to be relevant to the topic (NFI), and illegible or nonsensical information (NFU); (13)

Functional marker (FM) serves a particular purpose for the writer, and is often used as a transition to introduce reasons, arguments, and standpoints; (14) Rhetorically Functional Repetitions (RFR) occur when the writer restates previously expressed reasons, arguments or standpoints (p. 98–100). Following Wolfe et al. (2009), Chase (2011) further categorizes these elements into myside functional elements, yourside functional elements, extra functional elements, and non-functional elements. According to Wolfe et al. (2009), myside arguments represent the author's standpoint, supporting reasons for the author's standpoint, and elaborations of the author's standpoint, whereas yourside arguments represent the alternative perspective, counterarguments of the author's standpoint, and rebuttals of the counterargument. In Chase's (2011) model of argumentative elements, author's standpoint(s), level-1 reasons, and reasons below level-1, are subsumed into myside functional elements (or myside arguments). Counterargument(s), rebuttal(s), alternative standpoint(s), and reason(s) for alternative standpoint(s) are subsumed into yourside functional elements (or yourside arguments). Other elements like introduction, conclusion, title, functional markers, and rhetorically functional repetitions are covered by extra functional elements. Irrelevant information on the topic is categorized as non-functional elements.

Studies show that the quality of argumentative writing is positively correlated with the number of argument elements (e.g., claim, reasons) (e.g., Chase, 2011; Ferretti, MacArthur, & Dowdy, 2000; Rusfandi, 2015), the existence of counterarguments and rebuttals (e.g., Crammond, 1998; Wolfe, et al, 2009; Liu & Stapleton, 2014; Rusfandi, 2015, Qin & Karabacak, 2010), depth of justification (Liu & Stapleton, 2014; Crammond, 1997, 1998) and variety of argument structures (Crammond, 1997).

Chase (2011) explored the extent to which the elements of argumentative discourse, along with other aspects (e.g., coherence, cohesion) contribute to the overall quality of argumentative writing composed by 112 college students in an American college. She

found that functional elements (i.e., myside functional elements, yourside functional elements, extra functional elements) were significantly correlated with the quality of argumentative writing and accounted for 47% of the variance of overall essay scores. Ferretti et al. (2009) investigated whether the elaboration of writing goals based on Toulmin's argument elements in a writing prompt influenced the overall persuasiveness of argumentative writing composed by 4th and 6th grade students in a USA school. They found that students given the elaborated goal produced a greater number of alternative standpoints, level-1 reasons for alternative standpoints, and rebuttals compared to students given the general goal. They also found that elements of argumentative discourse contributed to 45%–70% of the variance in overall persuasiveness. Rusfandi (2015) investigated the potential use of the argument-counterargument structure in English L2 essays written by Indonesian EFL learners and found that claim, refutation, sub-claim, and justification were all significantly correlated with overall essay scores. The presence of the four elements accounted for 61.7% of the variance in the overall essay scores.

Liu and Stapleton (2014) investigated the efficacy of explicit instruction in counterarguments and rebuttals on writing quality of 125 Chinese EFL college students and found that explicit instruction had an effect on the amount of counterarguments and rebuttals in the case of the experimental group, and the overall essay quality of both control and experimental groups was significantly positively correlated with frequency of data, counterarguments, and rebuttals. Wolfe, et al. (2009) explored the effect of argumentation schema on the quality of claims, the number of reasons or supporting statements, and other side arguments (counterarguments and rebuttals) in 60 native speakers of English in an American university. They found that the clarity, the frequency of reasons, and the inclusion of other side arguments increased significantly in the experimental group, who receive argumentation schema-based tutorials. Their study implies that clear claims, number of reasons, and other side arguments are indicative of student argumentative essays of good quality. Crammond (1998) shows that there is a

significant group effect in the number of countered rebuttals among 6th, 8th, 10th graders and expert writers. Qin and Karabacak (2010) investigated argumentative essays of 133 second-year university English majors in a Chinese university using the Toulmin model of argument. They found that counterargument and rebuttals were significant predictors of overall writing quality, although counterargument and rebuttals were much fewer than the arguments for writers' standpoint (i.e., myside arguments).

Other studies suggest that argument complexity (Crammond, 1997) and types of argument (Coffin, 2004) are closely related to the quality of argumentative writing. Crammond (1997) investigated the relationship between argument complexity and persuasive writing skill of 36 6th, 8th and 10th graders and seven expert writers. Argument complexity was examined through two measures: maximum depth of an argument structure (i.e., the number of arguments in the longest embedded argument), and maximum variety of substructures used to elaborate an argument structure (i.e., sub-claims, backing for data, backing for warrant, warrant). He found that the maximum depth of an argument and the maximum variety of substructures in an argument showed a steady increase as the persuasive skill level increased and they significantly differentiated across three grades of student writers and expert writers.

The measures of the structure of argumentation to be explored include the number of opposing elements (e.g., counterarguments, rebuttals), depth of argument structure, and variety of argument elements. The latter two terms are redefined using Chase's (2011) argumentation analysis because Crammond's (1997) argumentation analysis is too complex to adopt as the identification of different argument elements is dependent on a rule-based semantic grammar (i.e., Backus-Naur form (BNF) grammar) which describes the semantic structures underlying the argument elements. The depth of argument is examined by the number of level-2 and above reasons. Variety of argument elements is examined by the number of different types of functional elements. An additional measure

is the number of non-functional elements. It is assumed that the number of non-functional elements is negatively correlated with writing quality.

#### 3.3.4.2 *The soundness of arguments*

The soundness of arguments is often evaluated against three criteria: (1) the acceptability of the reason; (2) the relevance or support the reason provides for the claim; (3) the sufficiency/adequacy the reason provides for the claim (Connor, 1990; McCann, 1989; Stapleton & Wu, 2015).

Studies on the evaluation of soundness of argument in second language writing are mainly classified into two groups: one group involves the use of an existing rating scale (Connor, 1990; McCann, 1989), and another group involves analytic measures (Stapleton & Wu, 2015). For example, McCann (1989) uses a scoring guide based on Toulmin's model of argument (Toulmin, 1958/2003; Toulmin, Rieke & Janik, 1979). The readers assign a rating to six argumentative traits: claims, data, warrant, proposition, recognition of opposition, and response to opposition, using analytic scales for each trait. However, the scoring criteria are subjective. For example, the scoring criteria for a rating of 4 for data quality is written as 'The data that are offered are relevant but not complete.' The writer leaves much for the reader to infer from the data (McCann, 1989, p. 75). It remains unclear as to how the ratings can be used for the development of descriptors in the rating scale, and therefore the existing scales are not discussed in this study.

The second group of studies is more relevant to my research because their use of relevance and acceptability as two criteria seem to be more easily operationalized by teachers when assessing argumentative essays than the first group of studies. Stapleton and Wu (2015) explored the quality of arguments in secondary students' persuasive writing in terms of surface structure and substance of arguments. The substance of arguments was investigated through the evaluation of the quality of reasoning. Stapleton and Wu (2015)

selected the 20 most common reasons after viewing 125 scripts, and asked 46 doctoral student raters to assign a point score to each reason in terms of its relevance and acceptability, following a scale of 0–3: Not relevant = 0; Not acceptable = 1; Weak = 2; Acceptable = 3. The scale of quality of reasoning on relevance and acceptability was shown to be effective in Stapleton and Wu (2015) and thus is used in this study to evaluate the soundness of arguments. The sufficiency of the reasons is quantified as the number of acceptable reasons as is the case in Means and Voss (1996). The reason for the quantification of sufficiency can be found in Section 3.2.2.2.

### 3.3.4.3 *Appeals of argumentation*

In Section 3.2.2.3 it was shown that appeals of argumentation are characteristic of argumentative and persuasive writing. The evaluation of appeals of argumentation in second language writing is mainly identified with Connor and Lauer (1985; 1988). In Connor and Lauer (1985), the effectiveness of appeals is judged in terms of appropriateness of content, sensitivity to the reader, and evidence of control. Appeals are rated as effective or ineffective. No explicit rating scale is provided. In Connor and Lauer (1988), the use of appeals is judged using three scales, each ranging from 0 to 3. One of the scales for the rational appeal scale is presented in Table 3.7, which shows that the quality of appeal is evaluated in terms of appropriacy, quantity and development.

Table 3.7 Rational appeal scale

0	No use of the rational appeal
1	Use of some rational appeals, minimally developed or use of some inappropriate (in terms of major point) rational appeals.
2	Use of a single rational appeal or a series of rational appeals with at least two points of development
3	Exceptionally well developed and appropriate single extended rational appeal or a coherent set of rational appeals.

Connor and Lauer (1985) explored the effectiveness of the use of rhetorical appeals in



relation to the quality of 150 essays written by high school students in America, England and New Zealand. They found that the ineffective uses of the rational and credibility appeals were significantly correlated with low holistic ratings of essays, the ineffective rational appeals were significantly correlated with essays judged to be low in quality, and four specific appeals (i.e., classification, contrast, shared interests and views, and emotion in the audience's situation) were significantly correlated with high rated essays. Connor and Lauer (1988) explored the use of rhetorical appeals in relation to the quality of essays of the same population using a 0–3 rating scale of effectiveness for rhetorical appeals. They found that there was a significant difference among the three groups of compositions (indicated both by countries and mean holistic scores) on the effectiveness of rational appeals, credibility appeals and affective appeals. Although there has been a significant correlation between effective appeals and writing quality (e.g., in Connor & Lauer, 1985), and a significant difference among different groups of compositions (e.g., in Connor & Lauer, 1988), since my study is focused on discourse analytic measures of writing quality, the effectiveness scale is converted into three analytic measures: the number of types of appeals, the number of specific appeals and development of appeals. The appropriateness of the appeals used in students' writing is not investigated as there is no existing measure for this. The assumption of the investigation is that writing quality increases with the number of appeal types and the number of specific appeals. Therefore, the persuasiveness of the essays will be evaluated in terms of the number of types of appeals and specific appeals, and development of appeals.

### **3.4 Conclusion**

Overall, this chapter has shown that a theoretical model of argumentative writing ability and its operational framework can provide a theoretical basis and guide for the rating scale design process. I have shown that the knowledge components identified as important to argumentative writing can be operationalized to varying degrees and with varying success. Table 3.8 shows four knowledge components in the left-hand column,

constructs or subcomponents in the center, and measures which were chosen for operationalization of these constructs in the right-hand column. Each discourse analytic measure is trialed during the pilot study phase, which is described in the following chapter.

Table 3.8 The operational framework of argumentative writing ability

Knowledge (competence) types	Constructs or subcomponents	Measures
Grammatical knowledge	A. CAF  (complexity, accuracy, fluency)	Syntactic complexity Mean length of clause (MLC) Mean length of sentence (MLS) Clauses per T-unit (C/T) Complex nominals per clause (CN/C) Coordinate phrases per clause (CP/C) Lexical complexity Sophisticated word types per word types (SWT/WT) Lexical words per words (LW/W) Accuracy Error-free T-unit per T-unit (EFT/T), Number of errors per T-unit (E/T), Number of errors per clause(E/C) Fluency Number of words (composing rate) Number of revisions (repair fluency)
	B. Mechanics	Number of spelling errors Number of punctuation errors Number of paragraphs
Textual knowledge	C. Cohesion	Number of reference Number of lexical cohesion Number of conjunction Number of incorrect use of cohesive devices
	D. Coherence	Proportion of topical progression patterns, Number of metadiscourse markers Number of errors in metadiscourse markers
	E. Topic-level argumentative structure	Number of paragraphs per argumentative move

Sociolinguistic knowledge	F. Register	Number of personal pronouns, contractions, formal and informal metadiscourse markers, colloquial word, 'knowledge-base' use of because
Argumentation knowledge	G. The structure of argumentation	Proportion of opposing elements Number of level-2 and above reasons Number of different functional elements Proportion of non-functional elements
	H. The soundness of argumentation	A 0–3 scale of relevance and acceptability
	I. Appeals of argumentation	Number of different types of the appeals The development of appeals

## **Chapter 4 Methodology- Analysis of writing scripts**

### **4.1 Research design**

This study is two-phased: (1) the development of a rating scale for classroom assessment of the argumentative writing of Chinese EFL college students majoring in English, and (2) investigation of the usability of the new rating scale. Each phase has a different research design due to the different purposes.

In the first phase, a pilot study and a main study were implemented to develop the new rating scale by identifying discourse measures that can be used to differentiate between different levels of argumentative writing performance of Chinese college EFL learners. The hypothesis for the pilot study and the main study is that most of the discourse analytic measures identified in previous writing research can successfully distinguish between different levels of argumentative writing performance of Chinese EFL learners. Firstly, a pilot study was undertaken to analyze a small number of writing scripts using discourse measures identified in the literature review. The pilot study aimed to trial these discourse measures and select those that were successful in differentiating between different levels of writing performance and could easily be transferred to descriptors for raters to use. Then, a large number of writing scripts were analyzed using discourse measures selected from the pilot study. At end of the main study, those measures which were successful in distinguishing between different levels of argumentative writing performance were identified and used to develop descriptors of a new rating scale.

This phase uses a quantitative research design. Frequency of occurrence of these variables was analyzed using descriptive statistics. In addition, an inferential analysis – an Analysis of Variance (ANOVA) – was conducted to investigate whether differences existed between measures of variables. The statistical results of both descriptive statistics and inferential analysis provided the basis for the development of level descriptors of the

newly developed scale. Writing tasks were created to elicit argumentative writing performance. Participants were grouped into different proficiency levels according to a proficiency score assigned by raters based on their performance.

The purpose of the second phase of the study was to investigate whether the new rating scale was usable by Chinese EFL argumentative writing teachers. For this purpose, three Chinese EFL argumentative writing teachers were selected to rate thirty writing samples using the new rating scale. Their rating results were analyzed to investigate how reliable the ratings were that the raters produced. Then, after the rating, the three raters'/writing teachers' opinions about the new rating scale were elicited using a questionnaire.

In this phase, a quantitative and qualitative research design was used. The inter-rater reliability of three Chinese EFL argumentative writing teachers was analyzed using inter-rater reliability analysis, which is quantitative and mainly involves numerical ratings and statistical procedures. The questionnaire feedback from the three raters was analyzed using content analysis, which involves finding themes and is therefore qualitative.

For reasons of readability, the methods, results and discussion sections of the two phases are described separately, with Chapters 4 and 5 being on the development phase, and Chapters 6 and 7 on the usability study phase.

## **4.2 Research questions**

Two main research questions are investigated, with one guiding each phase. Two subsidiary questions are further developed to explicate the second main research question and guide quantitative and qualitative analyses in the second phase. These questions are as follows:

**Phase 1:**

Which discourse analytical measures are successful in distinguishing between argumentative writing samples from Chinese EFL college students majoring in English at different proficiency levels?

**Phase 2:**

Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing?

- 2a. How reliable are the ratings produced by Chinese EFL argumentative writing teachers using the new rating scale?
- 2b. What are raters/Chinese EFL argumentative writing teachers' perceptions of the new rating scale?

**4.3 Research instruments****4.3.1 Writing tasks**

Three argumentative writing tasks were first created before the pilot study was undertaken through a small test development project, and then these tasks were administered to elicit student writers' writing performance. Existing argumentative writing tasks, TEM4 writing tasks in particular, were not used or adapted for the data collection for two reasons: first, current TEM4 writing tasks are unable to elicit student writers' organizational competence as they provide outlines in writing prompts; as shown in Figure 4.1, students are informed of how to organize paragraphs and content by following an outline. Second, the target population were preparing for the TEM4 test at the time of data collection – student writers may outperform their real writing competence if they happen to write on a prepared prompt, thus biasing the results.

My Idea of a University Arts Festival

You are to write in three parts.

In the first part, state specifically what your idea is.

In the second part, provide one or two reasons to support your idea OR describe your idea.

In the last part, bring what you have written to a natural conclusion or a summary.

Figure 4.1 An excerpt writing prompt from the TEM4 writing task

The aim of the small test development project is to develop writing tasks that represented formative and summative assessments of the target population's argumentative writing ability in the classroom context. It was decided that teachers' reflections on how argumentative writing had been taught and assessed, and their opinions on how students' achievements should be assessed, should be collected and used for the development of writing tasks. Three experienced writing teachers were chosen, as their feedback was expected to be insightful.

The writing task development project was phased:

1. Interviewing three Chinese EFL head writing teachers. Semi-structured interviews were conducted with three teachers. The interview questions included (1) How has your argumentative writing class usually been conducted? (2) How have students' achievements been assessed? (3) What are the proper parameters for each characteristic of an argumentative writing task? An interview sheet was handed out to each interviewee with a taxonomy of writing task parameters, based on typical existing argumentative writing task prompts (see Appendix 1), to help them answer the third interview question. The interviews were tape-recorded.

2. Drafting writing tasks. Following the teachers' feedback, I selected topics from online debating competitions. I created prompts with two opposing ideas, typical of their teaching of argumentative writing. Other parameters like word counts, time limits, inclusion of marking criteria and formats of writing instructions were also decided by considering the teachers' feedback. For example, advice was taken from one of the writing teachers. He suggested the adaptation of the format of instruction of the American College Test (ACT) writing test for the current writing tasks, in which detailed instructions on argumentative writing are given, such as "In your essay, take a position on this question. You may write about either one of the two points of view given, or you may present a different point of view on this question. Use specific reasons and examples to support your position". All teachers confirmed that their students could write 300 words within 50 minutes and believed that an extended length could provide the possibility of more space to address the opposing side of a controversial issue. 10 writing tasks were drafted.

3. Revision of writing tasks. 10 writing tasks were sent to one of the three EFL writing teachers for revision. After revision, these tasks were sent to a postgraduate, who was an English high school writing teacher in the UK, for second-round revision. Based on their feedback, confusing words and parts were replaced with clearer expressions, and less-frequently used words were annotated with Chinese translations.

4. Difficulty of writing tasks was investigated on a small sample from the target population using questionnaires (see Appendix 2). Twenty students who volunteered to take part in the questionnaire were required to rate the difficulty of each of the ten writing tasks on a five-point Likert scale and provide reasons for the difficulty rating they gave. Tasks were then ranked according to the average of their difficulty ratings. Three writing tasks at middle difficulty level were chosen for data collection in both the pilot and the main study, as shown in Figure 4.2, Figure 4.3 and Figure 4.4. Since writing tasks were administered as mid-term tests in class, each teacher randomly assigned each of the three tasks to each class of students, with one writing task to one class of students. For the task they were assigned, student writers were expected to write a 300-word essay, responding



to two opposing positions on a controversial issue, in 50 minutes. In the essay, students were required to take a position and support their position with evidence and examples. Students were also required to provide an appropriate title.

Weibo has become one of the major social networks for millions of Chinese netizens, providing a platform for users to share daily life and comment on heated topics. Recently, a popular website conducted an online survey on whether Weibo websites should initiate a real name identification system. The results have shown that the 57 percent of participants argue that real name authentication (认证) can create a healthy and harmonious network environment, while 45 percent of participants contend that real name authentication disregards users' right of privacy and the rest choose not to respond. Should Weibo require real name authentication? Write an argumentative essay outlining your point of view.

You have 50 minutes to plan, write and revise your essay. Write at least 300 words. In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Try to convince your readers with relevant arguments, evidence and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference. Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.

Figure 4.2 Writing task I – Weibo

Is tertiary education (高等教育) worth going to? This has been a controversial issue for many years. Some believe it remains a good investment. Others argue that increasing unemployment rates for college graduates dissuades high school graduates from applying for college. What do you think?

You have 50 minutes to plan, write and revise your essay. Write at least 300 words.

In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Try to convince your readers with relevant arguments, evidence and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference. Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.

Figure 4.3 Writing task II – Higher education

PM2.5, one of air quality indices, has drawn world's attention to China's air pollution. Enormous emission of air pollutants by industries has mostly caused this environmental disaster. The factories discharging air pollutants are often blamed as a major culprit of this environmental disaster. Some experts believe shutting down these factories can improve the country's air. Others believe that shutting down them would curb the economic growth and leave people unemployed. What do you think? Should factories be closed down to improve air quality?

You have 50 minutes to plan, write and revise your essay. Write at least 300 words.

In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Convince your readers with reasons and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference. Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.

Figure 4.4 Writing task III – Air pollution

### 4.3.2 Rating scale

The TEM4 rating scale was used to rate the written performance on three argumentative writing tasks. The TEM4 rating scale was selected as it is the most relevant to the target population and it can provide reasonably reliable criteria for ranking the writing scripts into different proficiency levels, although its validity remains a question in classroom assessment of argumentative writing (see more details in Section 2.5.4). The TEM4 rating scale (see Figure 2.6) is an analytic rating scale. As noted in Section 2.5.2, it comprises

three subscales: ideas and arguments, language use, and mechanics. The scale of ideas and arguments provides scoring criteria on the relevance of the content, quality of arguments and rhetorical organization. The scale of language use includes scoring criteria on accuracy, richness, appropriateness and fluency. The scale of mechanics includes scoring criteria on the accuracy of spelling, punctuation, capitalization, handwriting, and neat layout. Each scale has attached a score range (e.g., ‘1---2---3---4---5---6---7’), and under the score range four degree adjectives (e.g., ‘Poor Fair Good Excellent’) are aligned evenly along the score range to indicate the degree of writing performance.

## 4.4 Research participants

### 4.4.1 Raters

I contacted five teachers from the university where I was working to take part in the study. They were selected as they had at least three years teaching and rating experience of EFL writing. Therefore, their rating using the TEM4 rating scale was expected to be more consistent and reliable. Three of them agreed to take part. Their teaching and rating experience information is presented in Table 4.1. They had an average of more than five years’ teaching experience of EFL writing, and an average of more than three years of TEM4 writing rating experience.

Table 4.1 Raters’ background

Raters	English teaching experience (years)	TEM4 rating experience (years)	Specialty
1	5	5	Translation
2	7	3	English language teaching
3	7	3	English language teaching

#### 4.4.2 Students

The student participants were year-two and year-three EFL college students. They were selected because they were required to learn argumentative writing as one of their writing skills in an intermediate writing course and were taught and assessed on their achievements. Four universities in Chongqing where I was working were identified. They were chosen on the basis of their university rankings and can be roughly regarded as each representing a different rank in the target population. The rationale for this was to represent the writing proficiency level distribution of the target population whose argumentative writing abilities are assessed nationwide. Six teachers and their students agreed to take part in the study. 623 college students in total were recruited. Demographic information on these students are presented in Table 4.2 to Table 4.5. As shown in Table 4.2, female students made up about 90% of the total population, while male students were less than 10%. Such distribution is normally seen in the discipline of arts and humanities because male students normally tend to choose the discipline of science.

Table 4.2 Gender distribution

Gender	Male	Female	Total
Student	59	564	623
Percentage	9.5%	90.5%	100%

In Table 4.3 it can be seen that over half of the student participants majored in English education, while less than 5% of student participants majored in translation. Student participants majoring in foreign affairs management account for 17.5%. Those majoring in international relationship and public policy account for 9.3% and 8.7%.

Table 4.3 Major distribution

Major	English education	Foreign affairs management	International relationship	Public policy	Translation	Total
Student	379	109	58	54	23	623
Percentage	60.8%	17.5%	9.3%	8.7%	3.7%	100%

In Table 4.4 it can be seen that 81.1% students enrolled in 2013 and 18.9% student participants enrolled in 2012. This was due to the fact that the Teaching Syllabus of English Majors stipulates that writing courses are compulsory for year-2 students, while year-3 students are occasionally assessed on argumentative writing for mid-term performance.

Table 4.4 Year distribution

Year	2012	2013	Total
Student	118	505	623
%age	18.9%	81.1%	100%

Table 4.5 presents the distribution of student writers in different universities. Anonymous letters were used for different universities. They were ranked from high to low, according to the Nation's Discipline Evaluation University Ranking of Foreign Languages and Literatures (2012), A, B, C and D. As shown in Table 4.5, student participants from universities of A and D constitute 35.5% and 32.9%, while student participants from universities of B and C constitute 11.9% and 19.7%.

Table 4.5 University background

University	A	B	C	D	TOTAL
Student	221	74	123	205	623
Percentage	35.5%	11.9%	19.7%	32.9%	100%

#### **4.5 The development of the Chinese EFL college students' argumentative writing corpus (the CEAW corpus)**

The development has four stages: administration of writing tests, collecting writing samples, text digitization, and rating writing performance.

##### **4.5.1 Administration of writing tests**

Student participant information sheets (see Appendix 12), student informed consent forms (see Appendix 13), and test papers (see Appendix 9, Appendix 10, and Appendix 11) were delivered to the three head writing teachers interviewed for developing writing tasks in three universities. Test papers also requested demographic information. Writing tests were then administered separately in writing classes at different times in one week. The same data collection procedure was followed for each class. Before the test, a student participant information sheet and a student informed consent form were handed out to student participants for consent of participation. For those who did not sign the consent form, their writing scripts were not collected in the corpus. After that, the writing test was administered. Only one student (in University B) did not sign the informed consent form and did not respond to the writing task as the student asked for sick leave soon after administration of the test. At the end of test, test responses and informed consent forms were collected. While it was not possible to ensure that students taking the test first did not communicate content to others, it seems unlikely due to their tight class schedule, and the fact that they did not know that other students would be exposed to the same prompts. It is worth noting that these students took the test seriously because they were told that

their performance would be taken into account in mid-term summative assessments. It is reasonable to think that the test performance collected is a ‘real’ reflection of the target population’s writing proficiency, as expected to be elicited from the writing test.

#### **4.5.2 Collecting writing samples and text digitization**

Writing scripts were collected by each writing teacher after the tests were administered. I then fetched them from the universities. Before text digitization, student informed consent forms were checked by myself. All but one of the students signed consent forms and their writing scripts were included in the corpus to be built. Five postgraduate students in the university where I was working were then recruited to type the hand-written scripts into computers and save each document as a plain text file named after the student ID number. The text files were saved in folders named after a class number and an acronym of the corresponding university name. Student typists were required to replicate the original writing scripts by including all errors but excluding letters or words crossed out by the writers. Additionally, they were required to consult the researcher if any letter or word was found to be indiscernible. Each student typist was given a stipend for their time and effort based on the number of written scripts typed.

During the typing session, I resolved a small number of indiscernible parts identified by the student typists. After the typing, I checked all the documents in the text files against the original writing scripts and corrected a small number of typing errors. The CEAW corpus was then ready.

#### **4.5.3 Rater training exercise and rating**

Rater participant information sheets (see Appendix 14), and rater informed consent forms (see Appendix 15) were delivered to three raters. All three raters signed the consent forms. After each informed consent form was signed and collected from the three raters, a rater training exercise was held before actual rating took place. I selected 10 original writing



scripts which were roughly regarded as being different in writing quality and made three copies of each script for the training exercise. A rating sheet (see Appendix 6) was handed out to each rater. The rating sheet also requested demographic information. In the training exercise session, the three raters rated the ten scripts separately, filled in their ratings on the rating sheet and then compared their rating results. They identified the parts which contributed to rating discrepancies, discussed these parts with reference to the relevant rating criteria in the rating scale, and then a consensus on the interpretation of the scale descriptor was reached. However, the rating scale was not revised as a result.

Before the training exercise, rating packages were prepared. Each rating package included a TEM4 rating scale (see Figure 2.6), and two groups of writing scripts. The preparation of the rating package mainly involved numbering writing scripts, copying and printing the scripts, dividing the scripts into piles, covering demographic information on them and numbering piles of scripts. Running IDs were first assigned to the 623 writing scripts and then two copies were printed. These copies were divided into six piles based on their running IDs: Nos. 1–200, Nos. 201–400, and Nos. 401–623. Students' demographic information on the answer sheet was covered to avoid rater bias over student writers from low ranking universities. Each pile of writing scripts was numbered. Then rating packages were ready. After the training exercise, rating packages were handed out to the three raters, and they rated them separately at home. Raters were asked to write their ratings down at the top of each script. They were also asked to have a break when they felt it necessary to avoid fatigue. Then raters handed in the rated writing scripts with the ratings on. Since each writing script was rated by two raters, following the TEM4 rating practice, writing scripts where the score difference was equal to or larger than 3 points were identified. The originals of these writing scripts were copied and sent to a third rater for additional rating. These rated writing scripts with ratings on them were collected. The final score for each writing script was the average of the two ratings. For those writing scripts with a third rating, the final score was an average of the two ratings with the smallest score difference.

Since the final score was an average of two ratings, scores with intervals of 0.25, 0.5, and 0.75 were produced. The reported scores were all rounded (with 0.5 being rounded up). The score distribution is presented in Table 4.6. The table shows that the percentage of students scoring 1 to 5 is each less than 1%. The percentage of students scoring 7, 8, 9, and 10 is 10.3%, 27.0%, 35.6%, and 17.0% respectively. Students scoring 6 and 11 are relatively the same percentage, 3.4% and 4.2% respectively. There are no students scoring 0, 3, 13, 14, or 15.

Table 4.6 Distribution of final scores awarded to scripts in the CEAW corpus

Final score	Frequency	Percentage
1	1	0.2%
2	1	0.2%
3	0	0%
4	3	0.5%
5	5	0.8%
6	21	3.4%
7	64	10.3%
8	168	27.0%
9	222	35.6%
10	106	17.0%
11	26	4.2%
12	6	1.0%
13	0	0%
14	0	0%
15	0	0%
Total	623	100%

The descriptive analysis of the scores for student participants is presented in Table 4.7. As is shown in this table, the minimum score is 1 and the maximum score is 12. The mean score is 8.47 and standard deviation is 1.28.

Table 4.7 Descriptive analysis – Final scores awarded to scripts in the CEAW corpus

	N	Min	Max	Mean	SD
Score	623	1	12	8.47	1.28

Spearman correlation analysis was undertaken to investigate the inter-rater reliability between the three raters. The results are presented in Table 4.8. As shown in this table, the correlation coefficients between the scores of raters 1 and 2, raters 2 and 3, and raters 1 and 3 are .75, .66, and .65 respectively. According to Green (2013, p. 86), a correlation coefficient of .7 and upwards between raters in rating of writing or speaking is generally hoped for. In this study, it was decided that a correlation coefficient of .65 and upwards between raters is satisfactory as it is close to the expected value suggested by Green.

Table 4.8 Inter-rater reliability for the three raters

Rater	Correlation coefficients
Raters 1 and 2	$r = .75$ , $n = 253$ , $p = .000$ , two-tailed
Raters 1 and 3	$r = .66$ , $n = 206$ , $p = .000$ , two-tailed
Raters 2 and 3	$r = .65$ , $n = 285$ , $p = .000$ , two-tailed

#### 4.5.4 Data entry

I created an Excel document and keyed in the test takers' ID number, major, university, grade, running ID and class, and the three sub-scores. Overall scores, and final scores (an average of two overall scores) were calculated and saved in the Excel file. I also keyed in

the three raters' name, English teaching experience, TEM4 rating experience, and specialty.

#### **4.6 Pilot study**

The aim of the pilot study is to trial discourse analytic measures that had been identified in the literature on a small sample of writing scripts created in the current study and select suitable measures for the main analysis. The suitable measures should meet two criteria: (1) They should distinguish between writing at different writing levels; (2) They should be easy to use by human raters.

##### **4.6.1 Data selection**

At the time of the pilot study, the corpus of argumentative writing was created, containing 623 students' writing scripts on three writing tasks. All of these scripts were rated on the 15-point TEM4 rating scale. The score of the corpus ranges from 1 to 12. For the purpose of the pilot study, a small sample of writing scripts was needed. It was decided that scripts scoring low, middle, and high should be selected. The rationale for this is that the pilot study was conducted on only a small number of scripts and it was assumed that the analysis of three levels would yield clearer results. 623 writing scripts were divided into three piles based on their proficiency levels: low level (scoring 1–5), middle level (scoring 6–10), and high level (scoring 11–15). Fifteen scripts, five at each of the three proficiency levels, were randomly selected. The three groups of scripts are referred to here as 'low' for scripts scoring 1–5, 'middle' for scripts scoring 6–10, and 'high' for scripts scoring 11–15.

##### **4.6.2 Data analysis and results**

The analysis of the pilot study involved the manual coding of different measures in fifteen writing scripts by the researcher and descriptive analysis of distribution of different measures at three different levels using IBM SPSS (Statistical Package for the Social Sciences) Statistics for Windows, version 22.0 (IBM Corp., 2013). Given the scale of the

pilot study, no double coding was conducted. However, coding agreement analysis was conducted on the main study. The following sections describe the definition for each measure and the methods used in manual coding, and give exemplar coded samples from the analysis (i.e., texts coded are in bold) and results of the pilot analysis. Reasons why certain measures were selected for the main analysis are also given. It should be noted that, at the stage of the research design, it was decided that only those measures which can distinguish between different proficiency levels (i.e., a linear relationship with the level of writing) would be regarded as successful measures. However, in the analysis stage, it was found that this principle was inconclusive as different types of topical progression are closely interrelated with each other and should be investigated together in describing the coherence.

#### *4.6.2.1 Accuracy: error analysis*

The measures of accuracy to be investigated are error-free T-unit ratio (EFT/T), the number of errors per T-unit (E/T), and the number of errors per clause (E/C). In order to calculate these measures, errors, T-units, and clauses were identified. According to Lennon (1991), an error is defined as ‘a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers’ native speaker counterparts’ (p. 182). The Corpus for English Majors (the CEM) error taxonomy was adapted and used in the error coding. Three new error types that were not mentioned in the CEM error taxonomy but occurred in the writing scripts were added to the adapted taxonomy. They are wrong paragraphing (i.e., text which contains a different theme or topic is grouped into a paragraph where the text does not belong), incomprehensible sentences or phrases created by the literal translation of Chinese proverbs and set phrases, and unreadable sentences. A number of existing error subtypes were merged into more general types for simplicity and convenience of coding. The new merged error subtypes are errors in set phrases that used to belong to errors of different word classes (i.e., nouns, verbs, prepositions, and

conjunctions), noun errors related to either countability or number, errors in transitive or intransitive verbs, and errors in definite or indefinite articles. These new error subtypes were given new code names (e.g., wd2 for errors of set phrases). The adapted error taxonomy is presented in Appendix 3.

A sample excerpt from error analysis is reproduced in Figure 4.5. Errors are marked in bold. Codes are in brackets. As shown in the figure, ‘/’ is indicative of T-unit boundaries and the numbers 1–5 are indicative of number of T-units. [tn4] is the code for errors in tense-related verb forms and [vp2] is the code for misuse of finite and non-finite verbs.

1In recent years, air pollution has **attract**[tn4] increasing attention./2 Our environment is polluted by tremendous industrial emissions./3 More and more factories have been built in China for economic development./4 Some people hold that we should shut down these factories for **improve** [vp2]country's air./5 However, I can't agree with it./

Figure 4.5 Sample text for error analysis

Table 4.9 Descriptive statistics – Accuracy

Measures	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
Error free T-units per T-unit	.11	.10	.33	.18	.42	.08
errors per T-unit	1.48	.33	1.09	.40	.92	.30
errors per clause	1.05	.32	.76	.31	.65	.19

Results for the analysis of accuracy can be seen in Table 4.9, which presents the means and standard deviations for each measure at three proficiency levels (low, middle, high). It is clear from the analysis that all three measures were successful in distinguishing

between the different levels. It shows that, as the level of the essays increased, students made less errors per T-unit or clause, while error-free T-units and the percentage of error-free T-units increased. The error-free T-unit ratio was selected for the main study because this measure might be easier than the other three for raters to apply and is not sensitive to the length of the script.

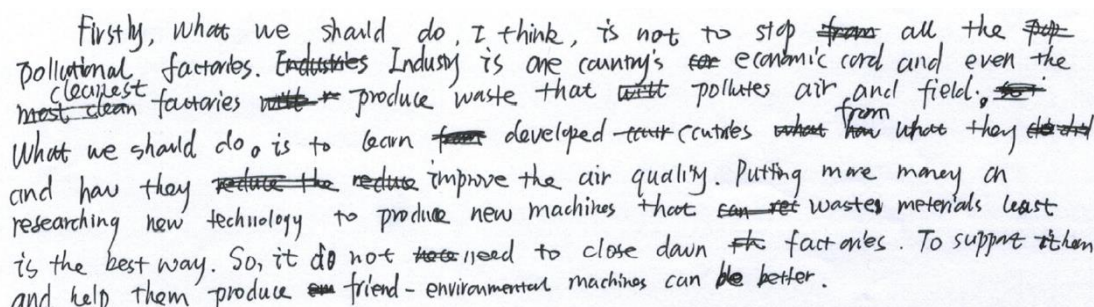
#### 4.6.2.2 *Fluency*

Fluency of writing was assessed through number of words and number of revisions. The number of words is a measure of composition rate. This use was possible because the essays in the writing assessment were written under a time limit. However, it was possible that some students did not utilize the whole time available; therefore, the interpretation of this measure needs to be cautious. In actual coding, wrong words were also counted as words written.

The number of revisions was a measure of repair fluency. Following Knoch (2009), revisions were defined as insertions and deletions of letters, or words, or longer stretches of writing. When an insertion and a deletion occurred at the same place, the number of revisions would count as two. It should be noted that this measure needs to be interpreted with caution as the coding of revisions did not distinguish between revisions that occurred during the writing process and those that occurred at a later time (e.g., self-corrections after first draft of writing). Researchers regard the former as a breakdown in fluency as it indicates writers have trouble in writing smoothly (e.g., Knoch, 2009) and other researchers regard the latter as a “fundamental” ability of writers and is irrelevant to proficiency levels (Personal communication with Nick Smith, 2019).

The writing scripts used for coding were photographic copies of the original writing scripts. The copied writing scripts retained the revisions, which made the assessment of repair fluency possible. However, repair fluency is not applicable in computer-assisted writing assessments. Figure 4.6 shows a sample excerpt from the study. As shown in the

figure, there are 20 revisions, for example, ‘from’, ‘pop’, ‘Industries’, ‘cor’, ‘most clean’, ‘will’. All of these revisions are in the form of deletions.



Firstly, what we should do, I think, is not to stop ~~from~~ all the ~~pop~~ polluting factories. ~~Industries~~ Industry is one country's ~~the~~ economic cord and even the ~~most clean~~ <sup>cleanest</sup> factories ~~will~~ produce waste that ~~will~~ pollutes air and field. ~~from~~ What we should do is to learn ~~from~~ developed ~~countries~~ <sup>from</sup> ~~what~~ <sup>how</sup> ~~they~~ <sup>they</sup> ~~do~~ and how they ~~reduce~~ <sup>improve</sup> the air quality. Putting more money on researching new technology to produce new machines that ~~can~~ <sup>use</sup> waste materials least is the best way. So, it ~~do~~ not ~~need~~ to close down ~~the~~ factories. To support ~~them~~ and help them produce ~~an~~ friend-environmental machines can ~~be~~ better.

Figure 4.6 Sample text with revisions

The results for the analysis of fluency can be found in Table 4.10. The table shows that the number of word tokens was a successful measure in distinguishing between three levels, while the number of revisions was not. Therefore, only Number of word tokens is pursued further in the main analysis.

Table 4.10 Descriptive analysis – Fluency

Measures	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
No. word tokens	208.80	86.38	241.80	77.16	286.80	37.71
No. revisions	5.00	2.83	14.20	17.41	4.20	2.05

#### 4.6.2.3 Grammatical complexity

The measures of grammatical complexity identified in the literature review were applied to the data. These measures include (1) mean length of clause (MLC): number of words divided by the number of clauses; (2) mean length of sentence (MLS): number of words divided by the number of sentences; (3) clauses per T-unit (C/T): number of clauses divided by the number of T-units; (4) complex nominals per clause (CN/C): number of



complex nominals divided by the number of clauses; (5) coordinate phrases per clause (CP/C): number of coordinate phrases divided by the number of clauses.

Sentence is defined as a group of words punctuated usually with a period, exclamation mark or question mark, and in some cases ellipsis marks or closing quotation marks (Lu, 2011). The definitions of T-unit and clause are given in Section 3.3.1.1. Coordinate phrases are structures that include adjective, adverb, noun, and verb phrases in coordination (Lu, 2011). Following Cooper (1976, cited in Lu, 2011), nominals (termed complex nominals in Lu, 2011) include (1) nouns plus adjective, possessive, prepositional phrase, adjective clause, participle, or appositive; (2) nominal clauses; and (3) gerunds and nominal infinitives (e.g., *To visit London had been her dream for years*).

It is a hot point<sup>1</sup> that whether<sup>2</sup> tertiary education worth going to or not.

Figure 4.7 Sample text with complex nominals

In Figure 4.7, a sample of complex nominals in a sentence is given. There are two complex nominals indicated by numbers 1 and 2: ‘a hot point’ (a complex nominal which comprises nouns plus adjective) and ‘point that whether tertiary education worth going to or not’ (a complex nominal which comprises nouns plus nominal clauses).

Table 4.11 Descriptive analysis – grammatical complexity

Measures	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
Coordinate phrases per clause	.21	.07	.15	.10	.18	.09
Complex nominals per clause	1.28	.42	1.30	.50	1.36	.33
Clause length	9.96	1.24	9.23	1.26	10.48	1.47
Clauses per T-unit	1.44	.14	1.44	.16	1.43	.23

Sentence length	17.67	2.08	15.75	2.01	15.54	3.88
-----------------	-------	------	-------	------	-------	------

Table 4.11 shows that only Complex nominals per clause was successful in distinguishing between different levels, while others were not, because only Complex nominals per clause increased as the level of writing increased. This might be because these measures were not suited to the data, as shown in Coordinate phrases per clause, Clause length and Sentence length. This might also be because different levels of writing did not distinguish between each other in terms of grammatical complexity as shown in Clauses per T-unit. I decided not to analyze grammatical complexity by Complex nominals per clause in the main analysis because the counting of complex nominals is time consuming; therefore, the use of this measure in the new rating scale is not practical.

#### 4.6.2.4 *Lexical complexity*

Three measures of lexical complexity, lexical sophistication I (sophisticated word type I per word type, SWT/WT I), lexical sophistication II (sophisticated word type II per word type, SWT/WT II), and lexical density (lexical words per word, LW/W) were identified in the literature. I applied Web Vocabprofile (Cobb, 2002), an online computational word frequency analyzer, to calculate sophisticated words and lexical words. The online analyzer can provide the number of word types and tokens of a text which can be found on the first 1000-word frequency list, the second 2000-word frequency list, the Academic Word List (an updated version of the University Word List developed by Coxhead (2000) and adopted by widely-used online word frequency analyzers such as Cobb's (2002) Web Vocabprofile and Laufer and Nation's (1995) Lexical Frequency), and not on any of these lists. Sophisticated word type I in the current study is defined as those not on the 2000 most frequent word list; that is, those on the AWL and those off-list words, and sophisticated word type II, those on the AWL. Lexical density was also provided by Web Vocabprofile. The online analyzer did not provide the definition of lexical words calculated. I decided to use the definition provided in Lu (2012) in the pilot study; lexical

words are defined as “nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs, *be*, and *have*), and adverbs with an adjective base, including those that can function as both an adjective and adverb (e.g., *fast*) and those formed by attaching the –ly suffix to an adjectival root (e.g., *particularly*)” (Lu, 2012, p. 192, italics added).

Table 4.12 Descriptive statistics – Lexical complexity

	Low		Middle		High	
Measures	Mean	SD	Mean	SD	Mean	SD
Lexical sophistication I	.19	.04	.19	.09	.17	.03
Lexical sophistication II	.09	.03	.08	.03	.08	.02
Lexical density	.56	.02	.54	.04	.54	.01

The results of the analysis of lexical complexity can be seen from Table 4.12. It is clear from the table that none of the measures was successful in distinguishing between the three levels of writing. This might be because the three groups of writing scripts were at roughly the same level in terms of lexical complexity. Therefore, none of these measures are further pursued in the main analysis.

#### 4.6.2.5 Mechanics

There are two concerns over the use of the frequency measure of mechanics in terms of punctuation errors, spelling errors, and capitalization errors. First, it is inconclusive in literature as to whether the length of the text is taken into account in the frequency measures of mechanics. Second, it is inconclusive in the use of a measure of the length of a text (e.g., T-units, word tokens). Considering the inconsistency of the use of these measures, I used three sets of measures: the number measure (e.g., the number of spelling errors), the first ratio measure (e.g., the number of spelling errors per word token), and the second ratio measure (e.g., the number of spelling errors per T-unit). It was found: there was no linear relationship between either spelling errors or punctuation errors, and

proficiency levels, in terms of all three measures; there was no linear relationship between capitalization errors and proficiency levels in terms of the number measure; there was a linear relationship between capitalization errors and proficiency levels in terms of both ratio measures. Since ratio measures were very small, only number measures are reported here in Table 4.13.

Table 4.13 Descriptive statistics – Mechanics

	Low		Middle		High	
Measures	Mean	SD	Mean	SD	Mean	SD
No. spelling error	1.80	1.79	4.40	1.51	1.00	0.00
No. capitalization error	.40	.55	.20	.45	.00	.00
No. punctuation error	.40	.55	.60	.89	.20	.45
No. paragraph	3.20	1.10	3.40	.89	3.60	.89

Table 4.13 shows that the number of capitalization errors decreased and the number of paragraphs increased as the writing level increased, while the number of punctuation errors and the number of spelling errors did not decrease as the writing level increased. Since capitalization errors were not common in the sample (less than one), only the number of paragraphs is further analyzed in the main study.

#### 4.6.2.6 Cohesion

As discussed in the literature review, the measures of cohesion worthy of investigation in the pilot study include the number of references, the number of lexical chains, the number of conjunctions, and the number of incorrect uses of cohesive devices. In Halliday and Hasan (1976), cohesion is operationalized across sentences. However, in this study, cohesion was operationalized between T-units, as T-units are the preferable production unit for textual analysis of EFL writing (Hunt, 1965). Therefore, any occurrence of reference that operated within a T-unit was not included in the calculations. Lexical

cohesion was operationalized as operating both within and across T-units, and conjunctions between T-units.

Reference was operationalized as the number of anaphoric and cataphoric personal pronouns (i.e., *I, me, mine, my/he/she/they/their/them/theirs, it, one, one's*), demonstrative pronouns (i.e., *this, that, those, these, here, there, now, then, the*), and comparative reference (i.e., *same, identical, equal, similar, additional, other, different, else, better, more, identically, similarly, likewise, so, such, differently, otherwise, so+adv, more+adv, less+adv, equally*). An example of reference identified in the sample is shown in Figure 4.8.

For one hand<sup>4</sup>, shutting down some factories would hinder economic growth and leave many people unemployed. /**This** step just can reduce the emission of air pollutants temporarily.

Figure 4.8 Sample text with reference

Conjunctions, or linking devices were operationalized as those listed in Halliday and Hasan (1976).

**For one hand**, shutting down some factories would hinder economic growth and leave many people unemployed. /This step just can reduce the emission of air pollutants temporarily./ **For another**, it also has some other terrible consequences. /These factories are important parts of industrial production. /Without them1 the manufacturing industry cannot continue to function. /**And** the relevant industries may also be influenced by this action. /**Besides**, these factories are closed, where do those unemployed can go? /How can they survive in the society without jobs?/ So in a word4, our country needs to take some measures to optimize industrial structure /and develop people's awareness to protect environment. /Try to maintain balance between human beings and those existing ecosystems. /**What's more**, urge factories to attach great importance their producing process. /Try their best to make the manufacturing process more environmental friendly./ Unswervingly pursue the scientific outlook on development, and build an environmental-friendly society./

Figure 4.9 Sample text with conjunctions

Conjunctions are additive (e.g., *and*, *in addition*, *furthermore*), adversative (e.g., *yet*, *though*, *in fact*), causal (e.g., *therefore*, *consequently*, *so*), temporal (e.g., *then*, *after that*, *finally*) (p. 242–3) in the form of words and phrases. Conjunctive devices identified in the sample are exemplified in Figure 4.9. There were very few colloquial uses of *and* (these *ands* often appear at the head of a sentence) and these were not included in the counting. Phrasal *and* was not included either as the coding only occurred between T-units.

Adapted from Knoch's (2009) lexical chains and Halliday and Hasan's (1976) cohesive chains, lexical chains in the current study were operationalized as three or more lexical items in a semantic relation with each other. The semantic relation was identified as repetition, synonymy, near-synonymy, superordinate, general nouns, or collocation. In

Figure 4.10, an example of a complete text is reproduced. Superscripts are used to indicate the number of lexical chains and bold is used to indicate lexical chains. Lexical chains identified in the essay in Figure 4.10 are (1) air quality, air pollution, pollution; (2) China, our country, country; (3) factories, the factories, clothes factories, iron factories, each factory; (4) problems caused by shutting down factories, some examples, not get the huge amount of tax, lack of money, lose their living guarantee, lose their jobs, a heavy burden for our country, steal and plunder, destroy our social civilization, our daily life will be inconvenient; (5) low-income people, old citizen and the disabled, a lot of people; (6) should be closed down, shut them down, shutting down, shut down; (7) take up a win-win policy, develop science technology, using air purifier, equip the purifier, improve their raw material, set up some concrete policies, taking up some reward and punishment, this kind of policy, taking up win-win policy; (8) improve air pollution, decrease air pollution, reduce air pollution, improve air pollution.

**The air quality<sup>1</sup>** is increasingly worse in **China<sup>2</sup>** nowadays. Experts say that most **air pollution<sup>1</sup>** is caused by **factories<sup>3</sup>**, and they believe that shutting down these **factories<sup>3</sup>** can **improve<sup>8</sup>** the **pollution<sup>1</sup>**. But I don't think so. **Factories<sup>3</sup>** shouldn't be **closed down<sup>6</sup>**.

It will cause worse **problems<sup>4</sup>** if we **shut** them **down<sup>6</sup>**. I have some **examples<sup>4</sup>**. Our **country<sup>2</sup>** will not get the **huge amount of tax<sup>4</sup>** if all the **factories<sup>3</sup>** are **closed down<sup>6</sup>**. Once the **country<sup>2</sup>** **lack<sup>4</sup>** of money, **low-income people<sup>5</sup>**, **old citizen<sup>5</sup>** and **the disabled<sup>5</sup>** will **lose<sup>4</sup>** their living guarantee. And with the **shutter down<sup>6</sup>** of the **factories<sup>3</sup>**, a lot of **people<sup>5</sup>** will **lose<sup>4</sup>** their jobs. The increasingly **unemployed people<sup>5</sup>** would be a heavy **burden<sup>4</sup>** for our **country<sup>2</sup>**. Sometimes, poor results **steal<sup>4</sup>** and **plunder<sup>4</sup>**. These will **destroy<sup>4</sup>** our social civilization. Also, our daily life will be **inconvenient<sup>4</sup>**. If the **iron factory<sup>3</sup>** is **shut down<sup>6</sup>**, we will unable to build house. And should we naked if **the clothes factories<sup>3</sup>** were **shut down<sup>6</sup>**? So we should take up **a win-win policy<sup>7</sup>** instead of **shutting<sup>6</sup>** all **factories<sup>3</sup>** down.

Taking about **improving<sup>8</sup>** the **air pollution<sup>1</sup>**, **science technology<sup>7</sup>** becomes vital important. I think the only way to **decrease<sup>8</sup>** the **air pollution<sup>1</sup>** without **shutting down<sup>6</sup>** **factories<sup>3</sup>** is to develop **science technology<sup>7</sup>**. Using **air purifier<sup>7</sup>** may be a good way. I think **each factory<sup>3</sup>** should equip **the purifier<sup>7</sup>** in the vents. And at the same time, they could improve their **raw material<sup>7</sup>** to **reduce<sup>8</sup>** the **pollution<sup>1</sup>** from the very beginning. Also, our **country<sup>2</sup>** should set up some concrete **policies<sup>7</sup>** to control the situation. Taking up some **reward<sup>7</sup>** and **punishment<sup>7</sup>** to encourage **factories<sup>3</sup>** to join the air-protecting team. This kind of **policy<sup>7</sup>** can reduce the **air pollution<sup>1</sup>** while growing the economic.

So I think close **factories<sup>3</sup>** **down<sup>6</sup>** is nonsense. Taking up **win-win policy<sup>7</sup>** is the efficient way to **improve<sup>8</sup>** the **air pollution<sup>1</sup>**.

Figure 4.10 Sample text with lexical chains



Incorrect use of cohesive devices includes the misuse of pronouns and conjunctions which are used as cohesive devices, and misspelling of these cohesive devices. Figure 4.11 presents wrong forms of conjunctions (in bold) and references (missing). It was found in the coding that incorrect use of cohesive devices was not common, therefore, the number of incorrect uses of cohesive devices was not investigated in the pilot study.

Misuse of conjunctions:
<b>For one hand</b> , shutting down some factories would hinder economic growth and leave many people unemployed. This step just can reduce the emission of air pollutants temporarily. <b>For another</b> , it also has some other terrible consequences.
Misuse of reference:
We know that the rate of unemployment having risen year to year, and have no good jobs.

Figure 4.11 Sample text with errors of cohesive devices

Results for the analysis of cohesion can be found in Table 4.14. It is clear from this table that both conjunctions and lexical chains were more frequently used by higher-level writers, while reference was not. Since the coding of lexical chains is very time-consuming and rater reliability would be hard to achieve, lexical chains was not regarded as a suitable measure for the new rating scale. Therefore, only the number of conjunctions is analyzed in the main study.

Table 4.14 Descriptive statistics – Cohesion

	Low		Middle		High	
Measures	Mean	SD	Mean	SD	Mean	SD
No. reference	8.40	8.23	5.60	2.51	9.80	2.59
No. conjunctions	5.00	3.16	7.60	2.07	9.00	2.35
No. lexical chains	3.40	1.14	4.00	1.00	6.60	2.88

#### 4.6.2.7 Coherence

The proportions of topical progression patterns (i.e., the percentage of T-units involved in certain type of topical progression), number of metadiscourse markers, and number of errors in metadiscourse markers were identified as measures for coherence.

##### *The TSA (Topical Sequential Analysis)*

After a closer review of topical progressions discussed in the literature, I decided to make refinements to clarify the confusing parts in Schneider and Connor's (1990) definitions of topical progressions. The first refinement is to solve the confusing categorization of semantically identical and differing relations. In the coding guidelines of the TSA, they categorize the same nouns with different post-modifications (e.g., *the ideas of scientists*, *the ideas of artists*, p. 427) as semantically identical relations, while they categorize the same nouns with different pre-modifications as semantically differing relations (e.g., *a nation*, *a very small, multi-racial nation*). Since this categorization is not explained by Schneider and Connor (1990) and little existing knowledge supports their categorization, I view two noun phrases with the same head noun but different pre-modifications and post-modifications as semantically-related differently.

In the refined Parallel progression (PP), the topic of the previous sentence is semantically identical to the topic of the following sentence. Following mostly Schneider and Connor's (1990) definition, the semantic identification is based on exact repetition (e.g., *rose*, *rose*), pronominalization (e.g., *rose*, *it*), synonymous relation (e.g., *infants*, *children*), number relation (e.g., *a child*, *children*), and polarity relation (e.g., *children*, *no children*). The relations of parallel progression are illustrated in Figure 4.12. Two successive sentences A and X are inter-related through the semantic meaning of their topics. For example, "6

For one hand, *shutting down some factories* would hinder economic growth and leave many people unemployed./7 *This step* just can reduce the emission of air pollutants temporarily./” (see Appendix 4, italics are used to indicate T-unit topics). In these two sentences, the topic of Sentence 6 is “shutting down some factories”, and it is picked up by the topic of Sentence 7, “this step” through the demonstrative pronoun “this”. Therefore, these two sentences are in parallel progression. Italics are used to indicate the topic of each sentence. A double ended arrow is used to indicate this relation.

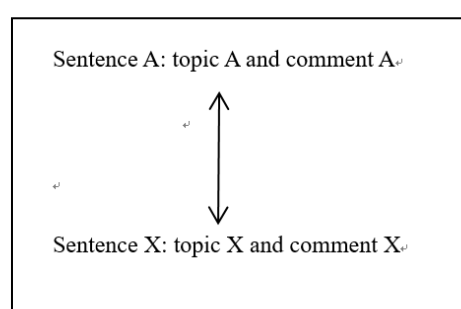


Figure 4.12 Parallel progression

The second refinement is to discard the unjustified dichotomy of directly and indirectly related sequential progression and to re-categorize different types of sequential progression for the convenience of coding. In the coding guideline, Schneider and Connor categorize any two contiguous sentence topics as directly related when they are in such relations as the same head nouns with different pre-qualifiers, word derivations, part-whole relation, repetition of part but not all of a preceding topic (e.g., *two characteristics*, *the first characteristics*), and comment-topic relation in which the comment of a sentence becomes the topic of the immediately following sentence. In contrast, they categorize topics related by semantic sets (e.g., *scientists*, *their inventions and discoveries*, and *the invention of radio, telephone and televisions*) as indirectly related. It seems that their categorization implies that the semantic sets relation is more distant than the topic-topic and the comment-topic relation, while there is no evidence to support this. In addition, it is unclear whether topics are semantically different (e.g., *a nation*, and *a very small, multi-*

*racial nation*) or identical (e.g., *the ideas of scientists, the ideas of artists*) in comment-topic relations. While the dichotomy is unjustified, and the comment-topic relation is too general to code, for the convenience of coding I re-categorized three subtypes of sequential progressions based on the positions where semantic relations occur: sequential progression I, sequential progression II, and sequential progression III. Instead of using terms such as directly- and indirectly-related sequential progression, my categorization does not connote semantic distance between sequential progressions.

Sequential progression I (SP1) refers to successive sentences of which topics are different. Semantic difference, similar to Schneider and Connor's (1990) definition, includes word derivations (e.g., *children, childhood*), part-whole relations (e.g., *family, father, mother*), repetition of part but not all of a preceding topic (e.g., *two characteristics, the first characteristic*), contrast (e.g., *children from rich families, children from poor families*), post-modification (e.g., *contributions of scientists, contributions of artists*), pre-modification (e.g., *a flower, a red flower*). In Figure 4.13, two successive sentences, A and B, are related to each other through the semantic meaning of their topics. For example, as shown in Appendix 4 (italics are used to indicate T-unit topics), "1In recent years, *air pollution* has attracted increasing attention./2 *Our environment* is polluted by tremendous industrial emissions.". The sentence topic "*air pollution*" in Sentence 1 and the sentence topic "*Our environment*" in Sentence 2 are in semantic sets relation. Therefore, these two sentences are regarded as belonging to sequential progression I. In Schneider and Connor's (1990) definition, semantic sets are not defined but it can be implied from their examples (e.g., *Scientists, their inventions and discoveries, and the invention of radio, telephone and television*) that semantic sets here seem to be identified with Halliday and Hasan's (1976) collocation. Here in this study, semantic sets are defined as words or phrases which are in collocation. A double-sided arrow is used to indicate this relation.

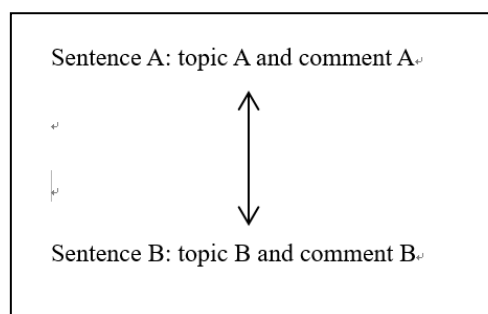


Figure 4.13 Sequential progression I

Sequential progression II (SP2) refers to successive sentences of which the comment of the preceding sentence is related to the topic of the following sentence. The relation can be semantically different or identical. Figure 4.14 shows that sentences A and B are interrelated through the comment of sentence A and the topic of sentence B. For example, as shown in Appendix 4 (italics are used to indicate T-unit topics), “4 Some people hold that *we* should shut down these factories for improving country's air./5 However, I can't agree with *it*./”. Sentence 4 and Sentence 5 are coded as belonging to sequential progression II as the comment of Sentence 5 “shutting down these factories for improving country’s air” is repeated by “*it*”, the topic of Sentence 6 (see the identification of ‘*it*’ as T-unit topic in Special cases of sentence topics below in this section). Since the repetition occurs between the topic of a sentence and the comment of a consecutive sentence, Sentence 6 can be regarded as being in a Sequential progression II with Sentence 5. The relation is indicated by a double-sided arrow.

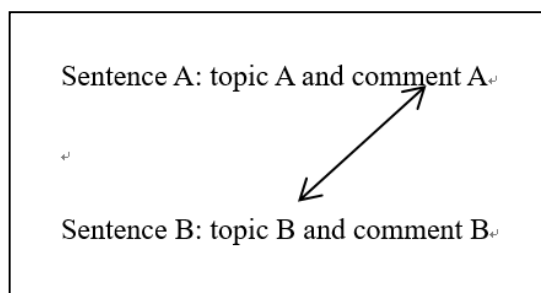


Figure 4.14 Sequential progression II

Sequential progression III (SP3) is also called discourse-related sequential progression. It is a relation in which the topic of the sentence is neither related to the topic nor the comment of the previous sentences. However, it is related to the discourse topic. It is a new sub-type that is not mentioned in Schneider and Connor (1990). For example, as shown in Appendix 4 (italics are used to indicate T-unit topics), “3 *More and more factories* have been built in China for economic development./4 Some people hold that *we* should shut down these factories for improving country's air./5 However, I can't agree with *it./*”. The topics of Sentences 3, 4 and 5 are *more and more factories*, *we*, and *it* respectively. The topic of Sentence 3 “*we*” can be interpreted here as people who have the administrative power to close factories, or the country. It is unclear which specific referent applies, therefore, its semantic relation with the previous sentence topic and the consecutive sentence topic is unclear too. However, with either relation identified it seems reasonable to regard *we* as being related to the discourse topic whether closing factories is necessary to improve air quality.

Unrelated topic progression (UTP) refers to an unrelated type of topical progression. It does not fall into any of the types of progression mentioned above. The unrelated topical progression is a relation in which the topic of a sentence is neither clearly related with the topic nor with the comment of the preceding sentence. It is also not clearly related to the discourse topic. The relation does not fall into any of the relations of parallel or sequential progression. Figure 4.15 presents this type of unrelated topical progression. The X mark is used to denote neither identical nor different relations indicated by two double-sided arrows occurring between the topic or the comment of the previous sentence A and the topic of the successive sentence B. For example, as shown in Appendix 4 (italics are used to indicate T-unit topics), “14 So in a word, *our country* need to take some measures to optimize industrial structure and develop people's awareness to protect environment./15

Try to maintain balance between human beings and those existing ecosystems./ 16 What's more, urge factories to attach great importance their producing process./17Try their best to make the manufacturing process more environmentally friendly./18 Unswervingly pursue the scientific outlook on development, and build an environmental-friendly society. /". Sentences 15, 16, 17, and 18 are grammatically-incorrect sentences as there are no subjects in these sentences. The topics of Sentences 15, 16, 17, and 18 are unknown. Therefore, these sentences are regarded as belonging to unrelated topic progressions.

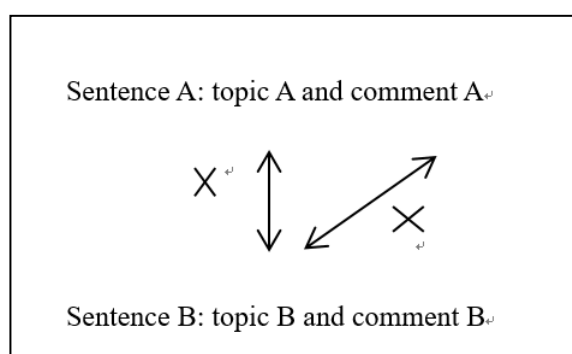


Figure 4.15 Unrelated topic progression

Extended sequential progression (ESP) is a sequential progression which extends over a number of sentences or T-units. That is, the comment or the topic of a previous sentence is taken up as the topic of a non-consecutive sentence after an interruption of a number of topics. The original definition was proposed by Simpson (2000, p. 301) but he did not address the topic-topic relation in his definition. Figure 4.16 shows that the topic or the comment of sentence A is taken up as the topic of a non-consecutive sentence X. For example, as shown in Appendix 4 (italics are used to indicate T-unit topics), “6 For one hand, *shutting down some factories* would hinder economic growth and leave many people unemployed./... 12 Besides, these factories are closed, where does *those unemployed* can go?/”. The topic of Sentence 6-*shutting down some factories* is in semantic collocation with the topic of Sentence 12-*those unemployed*. Since the relation extends over Sentences 7, 8, 9, 10, and 11, Sentence 12 is regarded as being in extended sequential

progression with Sentence 6.

The relations are indicated by two double-sided arrows.

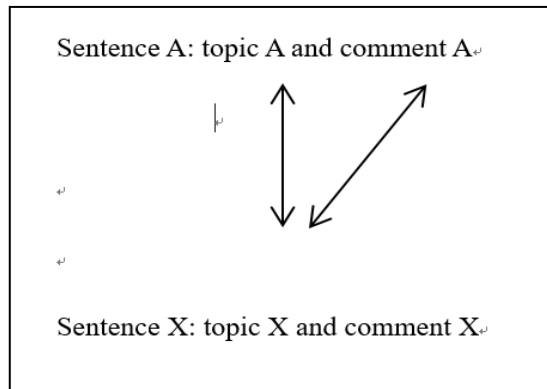


Figure 4.16 Extended sequential progression

The key to topical structure analysis is to identify the topics of sentences and discourses. Lautamatti (1978) describes the main idea discussed in the discourse as the discourse topic. The sentential topic is described as ‘what a sentence or T-unit is about’ (Connor & Farmer, 1990; Schneider & Connor, 1991). However, the descriptions prove that the rule of thumb is not sufficient to cover every case of sentence topic. Therefore, after viewing guidelines and examples of sentence topics illustrated in the above-mentioned literature, I summarize a number of cases for identifying the topic of a sentence or T-unit:

1. The topic is what the sentence or T-unit is about.
2. The topic often corresponds to the grammatical subject of the sentence or T-unit.
3. Noun or noun phrases are potential topics.
4. Special cases of sentences of which the grammatical subjects are not the topical subject (sentence topics in *italics*):

a. Cleft sentence:

It is *the scientist* who ensures that everyone reaches his office on time.

b. Anticipatory pronoun it:



It is well known that *a society* benefits from the work of its members.

c. Existential there:

There often exists in our society *a certain dichotomy* of art and science.

d. Introductory phrases:

I believe that *art and science* sustain and support each other. (Schneider and Connor, 1991, p. 424)

e. Imperative sentences:

Let us have a look for background for *Hong Kong*. (Schneider and Connor, 1991, example 3, p. 421)

The analysis of students' sample essays was operationalized in four steps: (1) identifying T-unit topics with underlines and identifying discourse topic; (2) diagramming the progression of topics indicated by indentation and labels; (3) determining the progression of T-unit topics; (4) calculating the number of topics in each type of progression. The proportion of each type of progression is the number of T-units involved in that type of progression over the total number of T-units minus one. One refers to the first T-unit which does not count in TSA. The analysis was completed manually by the researcher.

An exemplar essay coded for TSA and a diagram of different topical progression types found in the essay are presented in Appendix 4.

### *Metadiscourse markers*

As mentioned in Section 3.3.2.2, meta-discourse is defined as the ways in which writers organize text and help readers to interpret and react to the text they write. The linguistic devices used to indicate the organization of texts are categorized into textual and interpersonal meta-discourse markers. The former is mainly used by the writers to organize propositional content into a cohesive and coherent text, while the latter is mainly

used to express their attitudes towards the content and to involve readers. Although coherence in this study is mainly regarded as textual, the interpersonal meta-discourse markers are taken into account as they reflect writers' awareness of interactions between them and readers by involving readers through, for example, direct address to readers (e.g., *you must know*), thus indirectly indicating readers' perspectives or interpretations of the text.

A number of taxonomies on metadiscourse markers have been proposed since initial interest began (see Vande Kopple, 1985; Crismore et al., 1993; Beauvais, 1989; Hyland, 1998, 2005; Dafouz-Milne, 2008). The taxonomy used in Dafouz-Milne (2008) was chosen to guide the coding as this is the most comprehensive summary of metadiscourse markers. The taxonomy was adapted using examples from students' writing texts (see Table 4.15 ).

Table 4.15 Definitions and examples of metadiscourse markers,  
adapted from Dafouz-Milne (2008, p. 98, italics added)

Metadiscourse
<ol style="list-style-type: none"> <li>1. Textual markers <ol style="list-style-type: none"> <li>a. Logical Connectives: Express semantic relationships between stretches of discourse. They are additive (e.g., <i>and/furthermore/in addition</i>), adversative (e.g., <i>or; however; but</i>), consecutive (e.g., <i>so, therefore, as a consequence</i>), and conclusive (e.g., <i>finally, in any case</i>) <ul style="list-style-type: none"> <li>•Coded writing sample excerpt: Some people hold that we should shut down these factories to improve country's air. <b>However</b>, I can't agree with it.</li> </ul> </li> <li>b. Sequencers: Mark a particular position in a series. For example, <i>first, second, on the one hand,...on the other...</i> <ul style="list-style-type: none"> <li>•Coded writing sample excerpt: <b>For one hand</b>, shutting down some factories would hinder economic growth and leave many people unemployed.</li> </ul> </li> <li>c. Reminders: Reference back to the previous section of the text. For example, <i>as was mentioned above, let us return to</i> <ul style="list-style-type: none"> <li>•Coded writing sample excerpt: <b>From what has been discussed above</b>. We can conclude that tertiary education remains a good investment for people's whole life.</li> </ul> </li> <li>d. Topicalisers: Indicate shifts in topic. For example, <i>in the case of..., in regard to</i> <ul style="list-style-type: none"> <li>•Coded writing sample excerpt: And, <b>of course</b>, China has taken many manages to decrease it as much as possible.</li> </ul> </li> <li>e. Code glosses: Explain, rephrase or illustrate textual material. They are parenthesis (e.g., <i>When (as with the Tories now)</i>), punctuation devices (e.g., tax evasion: it is deplored in others but not in oneself), reformulators (e.g., <i>in other words/that is/to put</i></li> </ol> </li> </ol>

- 
- it simply*), and exemplifiers (e.g., *for example, for instance*)
    - Coded writing sample excerpt: **That is** why a heated debate was raised on whether these factories should be shut down to improve air quality.
  - f. Announcements: Reference forward to the future section in the text. For example, *there are many good reasons/as we'll see later*
    - Coded writing sample excerpt: **I have some examples.** Our country will not get the huge amount of tax if all the factories are closed down.
  - g. Illocutionary markers: the explicit naming of the act the writer performs. For example, *I hope/ I hope to persuade*
    - Coded writing sample excerpt: The another reason **I want to mention** is that people can make 'true' friends on the Internet.
2. Interpersonal markers
- a. Hedges: Express partial commitment to the truth-value of the text. They are epistemic verbs (e.g., *it may/might/must be two o'clock*), probability adverbs (e.g., *probably/perhaps/maybe*), epistemic expressions (e.g., *it is likely*)
    - Coded writing sample excerpt: And the relevant industries **may** also be influenced by this action.
  - b. Certainty markers: Express total commitment to the truth-value of the text. For example, *certainly/undoubtedly/clearly*
    - Coded writing sample excerpt: **There is no denying that** Weibo become more and more prevail over Chinese netizens and enrich our daily life.
  - c. Attributors: refer to the source of information. For example, *Smith claims that...*
    - Coded writing sample excerpt: **Some people** hold that we should shut down these factories to improve country's air.
  - d. Attitude markers: Display writer's affective values. They include deontic verbs (e.g., *have to/we must understand/ needs to*), attitudinal adverbs (e.g., *unfortunately/remarkably/ pathetically*), attitudinal adjectives (e.g., *it is absurd/ it is surprising*), and cognitive verbs (e.g., *I feel, I think, I believe*)
    - Coded writing sample excerpt: And as college students, we **should** make contribution to the protection of environment.
  - e. Commentary: Help to build a relationship with the reader. They include rhetorical questions (e.g., *what is the difference between A and B?*), direct address to reader (e.g., *you must know*), inclusive expressions which include the writer and the reader (e.g., *we all believe, let us summarize*), personalizations (e.g., *I do not want*), asides (e.g., *Diana (ironically for a Spencer) was not of the Establishment*)
    - Coded writing sample excerpt: Besides, **these factories are closed, where do those unemployed can go? How can they survive in the society without jobs?**
- 

A number of guidelines were taken into account during the coding to make the interpretation of results clearer. First, following Cheng and Stephenson (1996), only logical connectives that connect different T-units were counted. For example, logical connectives which provide cause-effect relations for a main clause and a subordinate clause, or additive relations between two subordinate clauses were not counted. Second, metadiscourse markers with varied uses were only counted once, for the major use they served. For example, 'As far as I'm concerned' was coded as *attributors*, while 'I' in the

phrase could also be coded as ‘personalization’. In this case, ‘As far as I’m concerned’ is counted as *attributors*.

The results of the analysis for coherence can be found in Table 4.16. This table shows that only discourse-related sequential progression was successful because only the discourse-related sequential progression decreased as the level of writing increased. Other types of progression and the metadiscourse markers did not show a linear relationship with the level of writing. However, it was decided that all of these progression types and metadiscourse markers would be analyzed in the main study because different types of topical progression are closely interrelated with each other and should be investigated together in describing the coherence.

Table 4.16 Descriptive analysis – Coherence

	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
Parallel progression/total progression	.18	.13	.23	.09	.23	.15
Extended parallel progression/ total progression	.13	.10	.23	.08	.16	.10
Related sequential progression/ total progression	.55	.13	.38	.11	.47	.16
Discourse-related sequential progression/ total progression	.31	.17	.22	.06	.16	.07
Unrelated topic progression/total progression	.04	.05	.01	.03	.04	.09
Extended sequential progression/total progression	.11	.08	.13	.14	.11	.02
No. Metadiscourse markers	12.80	6.98	18.20	6.54	16.80	5.22

#### 4.6.2.8 Register

The number of personal pronouns, contractions, informal metadiscourse markers, colloquial words, and ‘knowledge-base’ use of ‘because’ used by Shaw and Liu (1998) have been identified as the measure of register knowledge in this study (see Section 3.3.3). The assumption is that students with a higher writing proficiency are more aware of differences between spoken English and written English, and thus use less language typical of spoken English. I selected a list of categories where spoken and written language usually differ, from ‘Longman grammar of spoken and written English’ (Biber, Johansson, Leech, Conrad, and Finegan, 1999, p. 987–1036) and those mentioned in Shaw and Liu (1998). They include deixis (i.e., *here, now*); personal pronouns (i.e., *I, me, my, we, us, it, you, your*); hedges and vague language (i.e., *sort of*), contractions (i.e., *he’s, it’s*, etc.); forms of informal metadiscourse markers (personal pronoun+active verb, e.g., *I conclude/as a conclusion, I can see*; passives, e.g., *it can be concluded that*; non-finite, e.g., *To conclude*); colloquial word (i.e., *a bit, a lot, lots, thing, nice, big, little*); and use of because for clausal causes (e.g., ‘knowledge-base’ use of ‘because’ rather than causal use).

A sample of writing scripts was manually coded by the researcher. It was found that colloquial words and expressions were rare in students’ writing scripts. It was also found that the use of personal pronouns was common in all writing scripts. This is because all students in the pilot analysis used personal pronouns to express their opinions (e.g., *in my opinion, I think, I believe, as far as I am concerned*) and their experiences (e.g., *we as college students, my family, my hometown, us*) in making their arguments. Contractions and informal discourse markers were found to vary with the writing quality. The informal use of metadiscourse markers was limited to sequencers (e.g., *so, and*) and personal pronouns+active verb (or attitude markers) (e.g., *we know that, I can see that*). No use of

passives was found. The descriptive results are presented in Table 4.17. It can be seen from the table that both contractions and informal discourse markers did not decrease as writing quality increased. Therefore, they are not investigated in the main study.

Table 4.17 Descriptive statistics – Register

	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
No. contractions	1.00	1.00	1.60	2.61	1.40	1.14
No. informal metadiscourse markers	1.20	.45	2.80	2.17	2.20	2.95

#### 4.6.2.9 *The structure of argumentation*

Four measures of the structure of argumentation were identified in the literature: the number of opposing elements (counterarguments, counterargument reasons, rebuttals, rebuttal reasons), the number of level-2 reasons and below, the number of types of different functional elements, and the number of non-functional elements. In the current study, I converted the number measure into the proportion measure of the four structural elements mentioned above as the measure of the structure of argumentation to take the length of scripts into account. Through the standardized measures, the results are more comparable.

The coding scheme of structural elements of argumentative writing in the current study was adopted from Chase (2011) and is presented in Appendix 5. An exemplar of a coded sample is presented in Figure 4.17. Sentences were numbered for better illustration of the different coded elements. Structural elements were indicated with acronyms of the corresponding structural elements in capital letters in round brackets. In Figure 4.17, sentences 1–4 were coded as introduction elements (I) which provided background

information on the controversial issue. Although sentence 4 is a view opposing the writer's standpoint in sentence 5, indicated by 'however' in this sentence, since this view was mentioned in the writing prompt and the writer did not expand on it, sentence 4 is regarded as part of the background information rather than an opposing argument. Sentence 5 was coded as a negative standpoint using (SN) because SN is defined as expressing a negative opinion about the proposition, like 'I can't agree with it' in the sentence. Sentences 6–7, 9–13 were coded as reasons for the main standpoint. Sentence 6 was coded as two convergent reasons: 'hinder economic growth' (SN.R1) and 'leave many people unemployed' (SN.R2). Sentence 7 was coded as another convergent reason: 'reduce air pollution temporarily' (SN.R3). Convergent reasons do not depend on each other to support the main standpoint. Sentence 8 was coded as functional markers (FM), as it is used as a transition to introduce reasons. Sentences 9 to 11 were coded as coordinate reasons as they provide a fourth reason – the influence of shutting down factories on the development of industries. These reasons depend on each other to form the fourth reason to support the main standpoint. They were coded SN.R4a, SN.R4b and SN.R4c. Sentences 12 and 13 extend on the SN.R2 reason – the unemployment problem – and use two rhetorical questions – no place to go and no means to make a living – to support the influence of the unemployment caused by shutting down factories; this, in turn, strengthens the support offered by SN.R2. Since these two reasons offer support to the higher-level reason and their supports do not depend on each other, they were coded SN.R2.R1 and SN.R2.R2. Sentences 14–18 were coded as irrelevant information (Non-functional elements, NFI) because these sentences do not provide reasons to support or rebut the standpoint that factories should not be closed down, but are listing measures to be taken to reduce pollution, thus being considered irrelevant to the arguments. Sentences 19–20 were coded as rebuttals for the possible counterargument 'closing down factories is good' though the counterargument was partially mentioned in the theme position of sentence 19. These were coded as weakly developed rebuttals since counterarguments were not explicitly and well-developed. Sentence 21 is a recommendation and was coded

as conclusion (C).

The results of the analysis of argumentation structural elements can be found in Table 4.18. The table shows that the proportion of functional markers and the number of rhetorical functional repetition were successful in distinguishing between different writing levels, although both of them were not commonly seen in the sample scripts. It also shows that opposing elements (i.e., counterarguments, counterargument reasons, rebuttals, and rebuttal reasons) and the number of level-1 reasons and level-2 reasons and below did not increase as the writing level increased. Opposing elements were also not common in the sample scripts. Despite the results of the pilot study, I decided to investigate all of these elements in the main analysis because studies show that argumentative elements contribute to a fair amount of variance of essay quality indicated by overall scores (e.g., Chase, 2011). If these elements were not investigated, a student could, in principle, get a very high grade without understanding the elements of a good argument.



1. In recent years, air pollution has attract increasing attention. (I) 2. Our environment is polluted by tremendous industrial emissions. (I) 3. More and more factories have been built in China for economic development. (I) 4. Some people hold that we should shut down these factories for improve country's air. (I) 5. However, I can't agree with it. (SN)

6. For one hand, shutting down some factories would hinder economic growth (SN.R1) and leave many people unemployed. (SN.R2) 7. This step just can reduce the emission of air pollutants temporarily. (SN.R3) 8. For antoher, it also has some other terrible consequences. (FM) 9. These factories are important parts of industrial production. (SN.R4a) 10. Without them the manufacturing industry cannot continue to function. (SN.R4b) 11. And the relevant industries may also be influenced by this action. (SN.R4c) 12. Besides, these factories are closed, where does those unemployed can go? (SN.R2.R1) 13. How can they survive in the society without jobs? (SN.R2.R2) 14. So in a word, our country need to take some measures to optimize industrial structure (NFI) and develop people's awareness to protect environment. (NFI) 15. Try to maintain balance between human beings and those existing ecosystems. (NFI) 16. What's more, urge factories to attach great importance their producing process. (NFI) 17. Try their best to make the manufacturing process more environmental friendly. (NFI) 18. Unswervingly pursue the scientific outlook on development, (NFI) and build an environmental-friendly society. (NFI)

19 As far as I am concerned, closing factories cannot reduce the emission of air pollutants fundamentally. (RB) 20. Other effective measures should be adopted to solve the problem. (RB) 21. And as a college students we should make contribution to the protection of environment.

Figure 4.17 Sample text with argumentative elements

Table 4.18 Descriptive analysis – Argumentation structure elements

	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
Introduction	.44	.17	.17	.05	.17	.07
Standpoint	.19	.13	.08	.04	.09	.04
Level-1 reason	.08	.08	.28	.15	.23	.14
Level-2 reason and below	.05	.08	.23	.24	.21	.18
Non-functional elements	.10	.14	.12	.20	.06	.13
Functional markers	.00	.00	.01	.02	.06	.07
Counterargument	.03	.04	.00	.00	.00	.00
Counterargument reason	.04	.09	.00	.00	.00	.00
Rebuttal	.04	.04	.00	.00	.01	.03
Rebuttal reason	.02	.04	.00	.00	.02	.05
Rhetorical functional repetition	.00	.00	.01	.02	.03	.05
Conclusion	.01	.01	.11	.06	.11	.02

#### 4.6.2.10 Acceptability of arguments

The overall score for relevance and acceptability was established by the literature review as the measure of the relevance and acceptability of arguments, taking into account the conditions of acceptability of a premise (or reason) in an argument proposed by Govier (2013, p. 128) and van Eemeren et al. (2002). The assumption is that the higher the average acceptability score, the more acceptable a piece of argumentative writing is, and thus the higher the writing quality. The acceptability of reasoning was judged by the researcher using a three-level Likert-scale: not acceptable = 1, weak = 2, acceptable = 3.

A three-level Likert-scale was adapted from the four-level scale used by Stapleton and Wu (2015). The first level – not relevant = 0 – was removed from their scale because the

irrelevant information was not coded as reasons in the structural analysis of argumentation, therefore this irrelevant information was not taken into account in the judgment of acceptability of arguments. The conditions of acceptability of a premise (reason) in an argument include (1) when it is supported by a cogent sub-argument; (2) when it is known a priori to be true or when it can be proven to be true by logic alone (i.e., no one can steal his own property); (3) when it is a matter of common knowledge; (4) when it is supported by the experience of reliable sources; (5) when it appeals to authority; (6) when it is not known to be rationally acceptable, but can be accepted provisionally for the purpose of argument (Govier, 2013 p. 128), (7) arguments based on facts are more acceptable than those based on values or judgments, while the latter statements need more argumentation to demonstrate their acceptability (van Eemeren et al., 2002).

Exemplar acceptable, weak and unacceptable arguments are presented in Figure 4.18 and Figure 4.19. They are coded in <AC>, <WE> and <NA>. Correction of errors is presented in round brackets after the error. For example, the correction of “enrich” is in round brackets and appears after the wrong use of the word “full” Argumentative structural elements coded were retained to facilitate the judgment. Standpoints are coded as SP (positive), or SN (negative). The reasons are coded as SP.R or SN.R. In Figure 4.18, the author’s standpoint is that higher education is ‘worth going to’. The reason that tertiary education can ‘full (enrich) students’ knowledges’ offers acceptable support to the author’s standpoint as the reason states one of the benefits of receiving higher education. In contrast, the reason that higher education is good for students’ marks is not acceptable as it is not clear how students’ marks are positively influenced by higher education and how this influence can support his or her standpoint. In Figure 4.19, the author disagrees that tertiary education is worth going to. The reason that the rate of unemployment has been rising is considered as weak as it needs sub-arguments to state that a large part of the unemployed are those who have received higher education. Reasons that more professionals are needed than graduates, skills are more important than learning, and

learning is a waste of time are regarded as not acceptable because these reasons are not only unclear but also not facts. Professionals and graduates, and skills and learning are not two groups of opposites. For strong advocates of higher education, learning as a waste of time is not true. Neither is it true that we need more professionals when no proper context is given. It may well be that opponents who do not value the worth of higher education do probably think higher education is a waste of time.

... As far as I am concerned, I think that tertiary education is worth going to. (SP) The reasons is as follows: on the one hand, going to tertiary education is going to full (enrich) students' knowledges (SP.R1) <AC> and improve their profession skills. (SP.R2) <AC> I believe it is good for their mark. (SP.R3) <NA> On the other hand...

Figure 4.18 Sample text with acceptable and unacceptable arguments  
(correction of errors is presented in round brackets after the error)

.... Is tertiary education worth going to, according to this topic, different people holds different ideas. (I) Some people agree with it but others argue the topic. (I)...But I have opposite ideas to this topic (SN)...  
From our country's environment, the most important things that we need more professionals not graduates. (SN.R1) <NA> We know that the rate of unemployment having rised (risen) year to year, and have no good jobs. (SN.R2.R1) <WE> In my opinion, most college students are wastes (wasting) their time not to study more things to improve skills. (SN.R2) <NA> Compared to those junior college students, they are enter society more (much) earlier than college students...

Figure 4.19 Sample text with weak and unacceptable arguments  
(correction of errors is presented in round brackets after the error)

The results of the acceptability of reasons presented in Table 4.19 show that the average point of acceptability per reason increased as the writing quality increased. However, it remains a question as to how acceptability can be evaluated more objectively by a single human rater and the results built into a rating scale. Therefore, despite my efforts to make

the evaluation of acceptability by listing conditions for acceptability more objective and the result that shows it as a promising measure of writing quality, it is not pursued in the main analysis.

Table 4.19 Descriptive analysis – Acceptability of reasons

	Low	Middle	High
Mean	2.60	2.66	2.82
SD	0.45	.60	.29

#### 4.6.2.11 *Persuasiveness of argumentation*

The number of appeal types and the number of specific appeals were identified as the measure of persuasiveness of argumentation. According to Connor and Lauer (1985), a basic unit for the analysis of persuasive appeals is an episode, a term proposed by van Dijk (1982, cited in Connor & Lauer, 1985) to define a semantic unit of discourse in terms of ‘thematic unity’ or ‘psychological relevance’. van Dijk and Kintsch (1983, cited in Connor & Lauer, 1985, p. 319–20) operationalized an episode as meeting any of seven criteria, for example, change of possible world, change of time or period, change of perspective or point of view, change of place, introduction of new participants. An appeal or a number of appeals can be identified within an episode. Finding van Dijk and Kintsch’s episode boundary criteria hard to follow, I localized the criteria to fit the argumentative elements and operationalized in the current study an episode as a level-1 reason and its subordinate reasons. This was possible because a level-1 reason can be treated as an interface where a number of subordinate reasons are connected to support the level-1 reason and they can be treated together as a ‘thematic unity’ to support the main standpoint. To illustrate this, two examples are given in Figure 4.20 and Figure 4.21. Following Connor and Lauer (1985), appeals were identified in the following ways. In the case when an episode was identified as achieving either one type of appeal or another,

the type of appeal that played a major role was counted; in the case when reasons at lower levels in an episode supported the upper-level reason differently from how the upper-level reasons supported the main standpoint, as indicated by appealing to different appeals, appeals were counted individually.

In Figure 4.20, three episodes were identified as there are three level-1 reasons. The reason elements are indicated by (SN.R1), (SN.R2), and (SN.R3). Each episode was coded with a cause-effect appeal (R8) because each episode provides a different negative effect caused by shutting down factories. By showing the negative effects, the writer was attempting to convince readers that shutting down factories is not a good measure to reduce air pollution.

For one hand, shutting down some factories would hinder economic growth (SN.R1) (R8) and leave many people unemployed.(SN.R2) (R8) This step just can reduce the emission of air pollutants temporarily.(SN.R3) (R8)

Figure 4.20 Three level-1 reasons in convergent arguments operating as three appeals

In Figure 4.21, one episode was identified as there is one level-1 reason (SN.R1) and one level-2 reason (SN.R1.R1). The writer offered support to the main standpoint that higher education is not worthwhile by providing what the writer believes as ‘factual’ information (R12) in SN.R1. The writer further expanded on SN.R1 by showing that the rising unemployment rate in SN.R2. Both SN.R1 and SN.R1.R1 provide factual information to support upper-level statements. Since these two reasons apply to the same appeal, only one appeal is counted.

From our country's environment, the most important things that we need more professionals not graduates. (SN.R1) (R12) We know that the rate of unemployment having rised (risen) year to year, and have no good jobs. (SN.R1.R1)

Figure 4.21 Level-1 reasons and reasons below in subordination operating as one appeal

A third measure of persuasiveness was developed during the data analysis: the measure of development of appeal. It was found that the average number of reasons contained in an appeal was indicative of the development of the appeal. That is, a greater number of reasons indicates that the appeal is explained and developed in greater depth, and is thus assumed to be better in quality. Connor and Lauer (1988, p. 149) describe the development of an appeal as one of the criteria for evaluating the effectiveness of an appeal, although they did not realize it quantitatively.

The results of the analysis of appeals can be found in Table 4.20. It was found that the majority of appeals that were to be investigated were not seen in the sample writing. For those which were seen, R8 (the appeal of cause/effect-means/end-consequences) was much more commonly used than R1 (descriptive example), R4 (comparison), R11 (ideal or principle, or Values), R12 (Information), C14 (showing writer's respect for audience's interests and point of view), and A17 (appealing to the audience's views – emotional, attitudinal, moral). It was also found that only the number of R8 (the appeal of cause/effect-means/end-consequences) and reasons per episode were successful measures of persuasiveness of argumentative writing, while the number of other appeals and the number of types of appeals were not.

The assumption of the appeal analysis was that Chinese EFL college students' use of types of appeals and individual appeals may differentiate between different proficiency levels.

The results of the pilot study implied that Chinese EFL college students' use of appeals was limited to rational appeals, R8 (the appeal of cause/effect-means/end-consequences) in particular. Although the number of R8 appeals and the development of appeals were successful measures, these measures seem to show a similar nature of construct as justification and depth of justification in argumentation structure. As a result, appeal analysis is not conducted in the main analysis.

Table 4.20 Descriptive statistics – Appeals of argumentation

	Low		Middle		High	
	Mean	SD	Mean	SD	Mean	SD
No. descriptive example	.20	.45	.20	.45	.00	.00
No. narrative example	.00	.00	.00	.00	.00	.00
No. classification	.00	.00	.40	.55	.40	.55
No. comparison	.20	.45	.00	.00	.20	.45
No. contrast	.00	.00	.00	.00	.00	.00
No. degree	.00	.00	.00	.00	.00	.00
No. authority	.00	.00	.20	.45	.00	.00
No. cause/effect-means/ end-consequences	1.40	1.14	1.80	1.10	2.60	1.14
No. model	.00	.00	.20	.45	.00	.00
No. stage in process	.00	.00	.00	.00	.00	.00
No. ideal or principle	.00	.00	.60	.548	1.00	1.73
No. information	.60	.89	1.00	1.41	.60	.89
No. first-hand experience	.00	.00	.00	.00	.00	.00
No. showing writer's respect for audience's interests and point of view	.20	.45	.00	.00	.20	.45



No. showing writer-audience shared interests and points of view	.00	.00	.00	.00	.00	.00
No. showing writer's good character and/or judgment	.00	.00	.00	.00	.00	.00
No. appealing to the audience's views	.20	.45	.00	.00	.20	.45
No. vivid picture	.00	.00	.00	.00	.00	.00
No. charged language	.00	.00	.00	.00	.00	.00
No. appeal types	1.60	.89	2.80	.45	2.60	1.14
No. reasons per episode	1.40	.55	2.64	1.92	3.54	.99

#### 4.6.2.12 Top-level argumentative structure

The measure of top-level argumentative structure is the number of paragraphs per argumentative move. It was developed during the data analysis in the current study. It was found that the distribution of argumentative moves in relation to paragraphs, such as introduction, myside arguments, counterarguments, rebuttals, and conclusion, tended to vary with writing quality. For example, in essays of low quality, two out of three paragraphs were devoted to the introduction, introducing the background of the topic, leaving other elements like myside arguments and conclusion in one paragraph. This may indicate poor knowledge of the organization of paragraphs in accordance with the argument structure. On the contrary, in the essays of high quality, most paragraphs were devoted to the justification of the author's standpoint, which may indicate a more in-depth justification. It was assumed that the number of paragraphs per argumentative move would increase as the writing quality increased.

Table 4.21 shows that the measure did not increase as the writing quality increased. The ratio of paragraphs per element at all three levels roughly equals 1, though there is a slight

increase from low level to middle level. Although an interesting measure, it was not successful in distinguishing between different levels, and therefore it is not investigated in the main analysis.

Table 4.21 Descriptive statistics – Number of paragraphs per element

	Low	Middle	High
Mean	1.00	1.13	1.12
SD	.30	.30	.16

#### 4.6.2.13 Summary

Through the pilot analysis, a number of measures were found to have potential in distinguishing between different levels of writing performance for the main analysis. These measures are presented in Table 4.22. At least one measure was retained from the constructs accuracy, temporal fluency, mechanics, cohesion, coherence, and argumentation structure. However, lexical complexity, grammatical complexity, repair fluency, register, relevance and acceptability of reasons, persuasiveness of argumentation, and top-level argumentative structure were excluded as none of them distinguished between different levels of writing performance.

Table 4.22 Measures to be used in the main analysis

Constructs	Measures
Accuracy	Ratio of error-free T-units per T-unit
Temporal fluency	Number of word tokens
Mechanics	Number of paragraphs
Cohesion	Number of conjunctions

Coherence	The proportion of parallel progression
	The proportion of extended parallel progression
	The proportion of related sequential progression
	The proportion of unrelated topical progression
	The proportion of discourse-related sequential progression
	The proportion of extended sequential progression
Argumentation structure	The proportion of introduction
	The proportion of standpoints
	The proportion of level-1 reason
	The proportion of level-2 and below reason
	The proportion of non-functional elements
	The proportion of counterarguments
	The proportion of counterargument reasons
	The proportion of rebuttals
	The proportion of rebuttal reasons
	The proportion of rhetorical functional repetitions

## 4.7 Main study

The aim of the main study is to investigate those suitable discourse analytic measures identified in the pilot study using a large sample of writing scripts and to select successful measures for the development of a rating scale. The successful measures should meet two criteria: (1) They should distinguish between writing at different proficiency levels; (2) They should be easy to use by human raters.

### 4.7.1 Data selection

Before the main study was conducted, it was decided that the number of writing proficiency levels that would be used in the main analysis for identification of successful

measures should be decided beforehand. However, there is little literature to be referred to for the number of proficiency levels involved in the development of the rating scale. I decided to use five levels to describe the range of the target population's proficiency on the TEM4 scale. The rationale for this was based on Pollit's (1991) suggestion that five levels usually ensures a well-developed scale and well-trained raters (see Section 2.3.5 for details). It needs to be noted that although five levels was decided on for the main analysis, this is not necessarily the number of levels included in the new rating scale as the empirical results of the main analysis are also taken into consideration to decide the number of levels. However, it has to be admitted that the use of five levels here has an influence on it. This is acknowledged as one of the limitations of this study in Chapter 6 .

Four groups of writing scripts in the corpus were then formulated according to their scores as there were no writing scripts scoring 13–15 for level 5. The four groups are referred to as four AWA levels: 'level 1' for scripts scoring 1–3, 'level 2' for scripts scoring 4–6, 'level 3' for scripts scoring 7–9, and 'level 4' for scripts scoring 10–12.

At the time of analysis, it was found that the analysis of the 623 writing scripts in the corpus was time-consuming, and a smaller sample of four levels of writing scripts was drawn from the corpus. The main principle for the sampling is that there should be 'as large and as representative a sample as possible so as to minimize any possible impact caused by sampling error' (Green, 2013, p. 72). A 'large' sample was interpreted in the main analysis as one containing at least 30 for each score (i.e., the rule of thumb for a large sample) and a 'representative' sample was interpreted as there being a sample for each score and different sub-scores of 'ideas and arguments', 'language use' and 'mechanics'. Based on this principle, 89 writing scripts were selected from those at level 3 in the corpus, with 29 scripts scoring 7, and with 30 scripts each scoring 8 and 9. Scripts at levels 1, 2 and 4 were all retained as there were not enough to reduce at levels 1 and 2, and scripts at level 4 were roughly similar in number to those at level 3.

The distribution of levels in the sample and the corpus can be found in Table 4.23. It is clear that distribution of scripts at level 3 decreased from 454 to 89. The descriptive analysis of the score distribution in the sample can be found in Table 4.24. It shows that the mean score and standard deviation increased from 8.47 and 1.28 in the CEAW corpus to 8.71 and 1.80 in the sample respectively.

Table 4.23 Levels of writing scripts in sample and in the CEAW corpus

Writing level	Frequency		Percentage	
	Sample	Corpus	Sample	Corpus
1 (scores 1-3)	2	2	0.8%	0.3%
2 (scores 4-6)	29	29	11.2%	4.7%
3 (scores 7-9)	89	454	34.5%	72.9%
4 (scores 10-12)	138	138	53.5%	22.2%
5 (scores 13-15)	0	0	0.0%	0.0%
Total	258	623	100%	100%

Table 4.24 Descriptive statistics – Final scores awarded to writing scripts in sample and the CEAW corpus

	N	Min	Max	Mean	SD
The sample	258	1	12	8.71	1.80
The corpus	623	1	12	8.47	1.28

Thus, a corpus of 258 texts was built. The demographic information presented in Tables 4.25-4.29 is roughly representative of the 623 students taking part in the writing tests in this study in terms of the distribution of writing tasks, major, gender, university, and year. It should be noted that the demographic information is presented here for other researchers who might be interested and there was no analysis based on these

characteristics in this study. Table 4.25 shows that the written texts from three writing tasks in the sample were relatively evenly distributed, with students responding to the topic of Weibo being 35.3%, those responding to the topic of Higher education 31.8%, and those responding to the topic of Air pollution 32.9%.

Table 4.25 Writing task distribution used in sample

Writing task	Frequency	Percentage
Weibo	91	35.3%
Higher education	82	31.8%
Air pollution	85	32.9%
Total	258	100%

Several background variables for the student participants were recorded, as they were required to fill in this information on the answer sheet. Table 4.26 shows there were many more female students (89.9%) than male students (10.1%) in the sample. It may be because of the fact that fewer male students were enrolled in language related programs than female students.

Table 4.26 Gender distribution in sample

Gender	Frequency	Percentage
Female	232	89.9%
Male	26	10.1%

Table 4.27 shows that the largest group of students majored in English teaching (62%), while other groups of students were much smaller. These groups of students majored in English Foreign Affairs Management (18.6%), English International Relations (7.0%), English Public Diplomacy (7.8%), and Translation (4.7%).

Table 4.27 Major distribution in sample

Major	Frequency	Percentage
English Teaching	160	62.0%
English Foreign Affairs Management	48	18.6%
English International Relations	18	7.0%
English Public Diplomacy	20	7.8%
Translation	12	4.7%
Total	258	100%

Table 4.28 shows that two large groups of students were from the high-ranking university, A (33.3%) and the low ranking university, D (34.5%), while the other two smaller groups were from the middle ranking universities, B (16.3%) and C (15.9%).

Table 4.28 University background in sample

University	Frequency	Percentage
A	86	33.3%
B	42	16.3%
C	41	15.9%
D	89	34.5%
Total	258	100%

Table 4.29 shows that a large number of students were enrolled in 2013 (78.7%), while a relatively small number of students were enrolled in 2012 (21.3%).

Table 4.29 Year distribution in sample

Year	Frequency	Percentage
------	-----------	------------

2012	55	21.3%
2013	203	78.7%

Table 4.30 Grade distribution over tasks

Grade	Weibo	Environment	Education	Total
Mean	8.91	8.41	8.81	8.71
N	91	85	82	258
S.D.	1.69	2.04	1.61	1.80

Table 4.30 presents the grade distribution over three tasks. It can be seen from the table that the Weibo task produced the highest mean score, while the Environment task produced the lowest score. The scores produced by Education are more clustered than for the Weibo and Environment tasks. Since the equal variation assumption was violated in this case ( $F=4.21$ ,  $p=.02<.05$  in Levene's Test of Homogeneity of variances), Robust Tests of Equality of Means were conducted, showing that the three tasks produced similar results ( $F=1.67$ ,  $p=.19>.05$  for the Welsh procedure;  $F=1.92$ ,  $p=.15>.05$  for the Brown-Forsythe procedure).

## 4.7.2 Data analysis

The section below outlines the method taken for the main analysis. Since definitions for the different measures have been given in the pilot analysis in Section 4.6.2, they are not repeated in this section.

### 4.7.2.1 Accuracy

As described in the pilot study, the measure chosen for accuracy was Error-free T-unit ratio. The analysis involved identifying T-units and errors. T-units and errors were coded manually. Since error coding was laborious, three coders were involved; two of them were Chinese EFL teachers. The two coders plus myself coded 10 scripts; differences in scoring



and disagreement on the types and errors were discussed and resolved by following an adapted error taxonomy. Copies of the 258 texts were divided evenly into three folders. One folder was sent to each of the three coders. The coding was conducted on personal computers. All three coders worked in the same room for the convenience of communication during the coding and for the clarification of error types. Three sessions took place on three separate days to prevent possible fatigue. After all texts were coded, the type and the number of errors (in form of error codes) for each text were counted by Patcount. Patcount is a free text processing tool that can automatically count the frequency of lexical, syntactic, and discourse features in texts (Liang & Xiong, 2008). Since the accuracy of Patcount had not been reported, I checked its accuracy. The frequency results were 100 percent accurate on five randomly selected error-coded text samples. This was so because the error codes were combinations of letters and numbers (e.g., fm1, fm2) (see Appendix 3), which could hardly be mistaken as lexical, syntactic, and discourse features by the software. Therefore, the frequency results of error codes were not manually edited for misrepresentations. The results were automatically recorded into the Excel spreadsheet and I saved it as a separate file.

After the coding of errors, I coded all the T-units on the same texts used in the error coding. The number of T-units and the number of error-free T-units were counted and recorded into the Excel spreadsheet by myself. To ensure inter-rater reliability, a second coder was invited to double-code a subset of 15 scripts randomly selected from the sample. Pearson correlation coefficients were calculated for error coding and T-unit coding using IBM SPSS.

#### 4.7.2.2 *Temporal fluency*

The measure for temporal fluency identified in the pilot analysis was the number of word tokens. The number of word tokens was automatically calculated by Web Vocabprofile (Cobb, 2002), an online analyzer which provides word tokens automatically. Since the

number of word tokens was analyzed by a computer program, double rating was not carried out. After the analysis of word tokens in each text, the number of word tokens was entered into the Excel spreadsheet created for the main analysis.

#### 4.7.2.3 *Mechanics*

Mechanics was operationalized as the number of capitalization errors and the number of paragraphs. Since capitalization error was one of the errors coded in analysis of accuracy, the number of capitalization errors was counted by the researcher and entered into the Excel spreadsheet created for the main analysis. The same batch of folders of texts were used for the coding of paragraphs. The paragraphs were coded manually by the researcher and the number was entered into the Excel spreadsheet. Since the definitions of capitalization error and paragraph were unambiguous, the coding was easy and no double coding was carried out.

#### 4.7.2.4 *Cohesion*

The number of conjunctions was identified for the main analysis of cohesion. Following the principle that conjunctions operate between T-units, I manually coded the conjunctions which connected T-units in each script by labeling numbers after each conjunction. The manual coding was aided by checking the list of conjunctions given by Halliday and Hasan (1976, p. 242–3). Those which are identified by Halliday and Hasan (1976) as not cohesive devices (i.e., both...and..., either...or..., neither...nor...) were not coded. After each coding, the number of conjunctions was recorded on the Excel spreadsheet. A second coder was involved to double-code 15 randomly selected scripts. A Pearson correlation coefficient was calculated using SPSS.

#### 4.7.2.5 *Coherence*

Using the categories established in the pilot study, I manually coded the scripts. The manual coding included identifying T-units by labeling ‘/’ after each T-unit, underlining

the topics of each T-unit, copying topics onto a different Word file and numbering these topics, drawing diagrams (see Appendix 4) as used by Lautamatti (1978), and identifying the type of topical progression and labelling it. Appendix 4 shows different types of topical progressions occurring in a text. Since the definitions of different types of topical progressions have been given in the pilot analysis (see Section 4.6.2.7), these types shown in the diagram are not discussed here. After the coding of topical progressions, the number of different types of topical progression was recorded and entered into the Excel spreadsheet. A second coder was invited to double-code a subset of 15 texts. A Pearson correlation coefficient was calculated using SPSS.

#### *4.7.2.6 Argumentation structure*

Following Chase's (2011) coding taxonomy, I manually coded the scripts. To ensure inter-rater reliability, a second coder was involved to double-code a subset of 15 texts. A Pearson correlation coefficient was calculated using SPSS.

### **4.7.3 Data analysis – Inferential statistics**

A one-way Analysis of Variance (ANOVA) was conducted using SPSS to examine whether any differences between different writing levels occurred purely due to sampling variation. To conduct ANOVA, three assumptions should be satisfied (Green, 2013, p. 108). The first assumption is that the sample should be normally distributed. According to Wild and Seber (2000, cited in Knoch, 2009), ANOVA is robust enough to cope with departures from this assumption. However, when independent variables show a markedly non-normal distribution, nonparametric tests should be used (Green, 2013, p. 99). Green (2013, p. 82) describes a general rule to decide the 'excessiveness' of non-normal distribution, that is, if a skewness index (skewness value divided by standard error of skewness, positive or negative values) does not exceed 2, a distribution can be described as approximately normally distributed. The Kruskal Wallis test is one of the parallel nonparametric tests for one-way ANOVA when differences between 3 or more

independent variables are compared (Abbot, 2011, p. 293). In the main analysis, the Kolmogorov-Smirnov test was conducted to examine whether the sample was normally distributed. The Kolmogorov-Smirnov test was chosen as it is the most powerful for a large sample size (over 50) (Ricci, 2005, cited in Larsen-Hall, 2010) and the sample size in the current study is larger than 50 (258). A skewness index was calculated to determine the use of parametric ANOVA or Kruskal Wallis tests. Based on the general rule described by Green (2013), it was decided if the skewness index exceeded 2, the Kruskal Wallis test would be conducted.

The second assumption requires that the groups compared should have equal variances. While Green (2013, p. 108) suggests Levene's test can examine the homogeneity of variance. Wilcox (2010, cited in Larsen-Hall, 2010, p. 395) suggests that Levene's test should not be used as the sole source of information for homogeneity as it is too sensitive in the case of large samples and does not have enough power in the case of small samples. Wild and Seber (2000, cited in Knoch, 2009) suggest that variances can be regarded as equal if the largest standard deviation is no more than twice as large as the smallest standard deviation. Since there was no absolute rule to follow, both statistics were considered. If the variances were found to be unequal, a Welch test (Welch's variance-weighted ANOVA) was used. This test is sufficiently robust to overcome the situation where variances are not equal. The third assumption is the independence of samples. This assumption was satisfied because no writing script in all of the groups in the current study was repeated.

Multiple comparisons were conducted when significant differences were found between different proficiency levels using one-way ANOVA or the Kruskal Wallis test. The Games-Howell post hoc test was conducted to locate where the significant differences occur after one-way ANOVA. This post hoc test is appropriate when variances or sample sizes are unequal across groups (Larsen-Hall, 2010, p. 282). Since the groups in the

current study were of unequal sizes, the Games-Howell test was used. Stepwise step-down multiple comparisons built into the SPSS Kruskal Wallis analysis were conducted to locate where the significant differences occur when significant differences were found with the Kruskal Wallis test. While pair-wise or stepwise step-down post hoc comparisons were performed for each measure, it was not deemed important for each measure to achieve statistical significance between each adjacent level. However, post hoc comparisons between neighboring levels are reported in the results chapter.

Effect size was also calculated. It has been suggested that effect size be reported in the statistical analysis results (e.g., Larsen-Hall, 2010, personal communication with Fulcher) because it offers useful information that other statistics do not, such as *p*-value. According to Kline (2004, p. 97, cited in Larsen-Hall, 2010), effect size is ‘the magnitude of the impact of the independent variable on the dependent variable’. There are two families of effect sizes: group difference indexes and relationship indexes. Group difference indexes measure the size of the mean difference. Relationship indexes measure the closeness of two variables. In the case of the current study, effect size can be used to indicate the size of the difference in discourse measures between four writing levels. Larsen-Hall (2010) states that this statistic is more of practical value to researchers than *p*-value in that it tells researchers whether a difference found between pairwise comparison is large enough to be taken into consideration in a real-life situation. If the effect size is quite small, then it makes sense to simply discount the findings as unimportant, even if they are statistically significant. If the effect size is large, then the researcher has found something that it is important to understand. In addition, effect size is independent of group size while *p*-value is sensitive to a large group size. For one-way ANOVA, there are two types of *d* family measures: *f* for the overall ANOVA test effect size and Cohen’s *d* for group effect size. Larsen-Hall (2010, p. 118) provides a calculation formula for these two measures:

$$f^2 = \frac{\eta^2}{1-\eta^2}, \eta^2 \text{ is calculated by SPSS; } d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma}, \text{ where } \bar{x}_k = \text{the mean of group } k,$$

and  $\sigma$ =the pooled standard deviation of the groups being compared. Her book also

summarizes Cohen's guidelines for judging values of  $f$  and  $d$ : for  $f$ ,  $f = .10$  (small),  $f = .25$  (medium), and  $f = .40$  (large). For  $d$ ,  $d = .2$  (small),  $d = .5$  (medium) and  $d = .8$  (large). With Cohen's guidelines, the size of difference can be measured. Since the purpose of this measure is to investigate the size of the practical effect that the independent variable (i.e., all four AWA writing levels) has on the dependent variable (e.g., number of paragraphs), rather than that which certain groups (e.g., AWA level 1 and 2) have on the dependent variable,  $f$  for the overall ANOVA test rather than  $d$  is selected in this study. However, as noted by Cohen (1988, cited in Larsen-Hall, 2010), the importance of effect size is not only dependent on his guidelines but also on the practice of previous studies in the field.

## Chapter 5 Results and discussion – Analysis of writing scripts

This chapter presents the analysis results of each aspect of argumentative writing ability outlined in Chapter 4. The research questions to be answered are (1) Which discourse analytical measures are successful in distinguishing between argumentative writing samples from Chinese EFL college students majoring in English at different proficiency levels? (2) Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing? Then key findings for each aspect are summarized and discussed in relation to previous literature. Based on the findings and discussion, a trial scale for each aspect of argumentative writing ability is designed and presented.

Each section focuses on one aspect of argumentative writing ability. For each aspect, four results sections are provided. Firstly, box-and-whisker plots are provided to show the distribution of each variable over four different AWA levels. Box-and-whisker plots display graphically the median, interquartile range and extremes of a dataset. The box on each plot represents the interquartile range, with the bottom demarcating the first quartile and the top the third quartile. The horizontal line inside the box represents the median of the distribution. The whiskers at each end of the box indicate the minimum and the maximum data. Circles and asterisks represent outliers and data extremes respectively. Secondly, descriptive statistics for each variable including the mean, standard deviation, the minimum and the maximum are presented in a table. The third data item is the inferential analysis – ANOVA or Welch, or Kruskal Wallis results.  $F$  values, or  $\chi^2$  values (Chi-square) and  $p$  values are reported, to indicate successful measures. A  $p$  value of .05 is regarded as the statistical significance cut-off for successful measures. The fourth data item is effect size. Effect size (eta squared,  $\eta^2$ ) is presented to indicate the practical effect that the independent variable has on the dependent variable. Eta squared was chosen instead of  $f$  because it can be calculated directly by SPSS, while  $f$  needs to be calculated by following the formula  $f^2 = \frac{\eta^2}{1-\eta^2}$ . Cohen's guideline on the magnitude of

effect size in terms of eta squared is as follows: eta squared = .01, small; = .06, medium; = .14, large. Although effect size is reported, it is not considered for the selection of successful criteria because all successful measures have medium and large effect.

Analytical scales for each trait/attribute were developed based on the principles proposed by Knoch (2009, p. 169) in her development of the DELNA scale, as follows:

- Only measures that successfully distinguished between the different levels of writing were used in the rating scale (i.e., measures that were statistically significant and did not result in a u-shaped or n-shaped distribution).
- The measures selected needed to be used by raters in a rating situation.
- The differences between levels needed to be large enough to be detectable by raters.
- The measures had to be reliable (as indicated by inter-rater reliability measures). A reliability of over .80 was seen as acceptable.
- Only measures that incorporated features that were found in most scripts were included in the scale.
- If several measures were available for a certain feature of writing, the one that would be the easiest to apply in a rating situation and that had the best discrimination between levels was chosen.
- Rating scales were designed based on either the numeric value of a measure or an approximation (e.g., 50% was represented as ‘half’).

An additional principle was added to this list: A feature can be used to discriminate between one level and the rest. This helps to make that level unique, and thus can be used as a key to identify which texts are likely to be included in that level.

## **5.1 Mechanics**

Mechanics was measured through the number of paragraphs.



### 5.1.1 Results

No inter-rater reliability check was undertaken for the coding of paragraphs because there was no difficulty in identifying paragraphs. Paragraphs were all indicated by indentations in the writing samples. No unclear indentation was found during the coding for paragraphs.

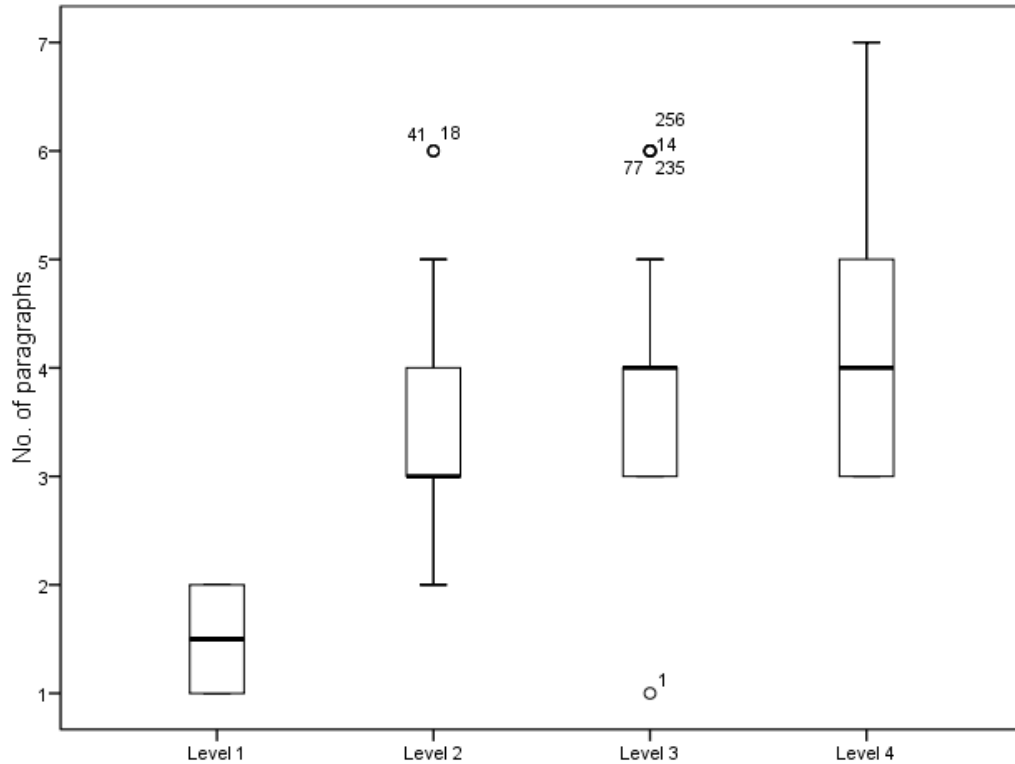


Figure 5.1 Distribution of paragraphs over the AWA levels

Figure 5.1 and Table 5.1 present the number of paragraphs for each AWA level. It can be seen that writers at level 1 (L1) produced less than two paragraphs on average while writers at level 4 (L4) produced four paragraphs on average. Students at levels 2 (L2) and 3 (L3) produced a very similar number of paragraphs on average.

Table 5.1 Descriptive statistics – Number of paragraphs

AWA level	Mean	Std. Dev.	Minimum	Maximum
-----------	------	-----------	---------	---------

1	1.50	.71	1	2
2	3.62	1.08	2	6
3	3.89	.96	1	6
4	4.05	.93	3	7

Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{\text{sewness}} = 2.43 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 13.388$ ,  $p = .004 < .05$ , with a statistically significant difference in the number of paragraphs between the four different levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 1 and 2, and levels 3 and 4. According to Cohen's guideline on the magnitude of effect size, eta squared = .01, small; = .06, medium; = .14, large; the effect size for the number of paragraphs is medium, eta squared = .07.

Paragraphing was used as the measure for the analysis of mechanics. This measure has previously been used for the assessment of writing by Knoch (2009) in the development of the DELNA rating scale. Knoch (2009) found that the number of paragraphs successfully distinguished between five DELNA levels although the differences in the mean between the five levels were not obvious (less than 1 paragraph). Similar to Knoch's (2009) results, the number of paragraphs also discriminated between the four different levels in this study and the differences were also not obvious. Although the number of paragraphs is a successful measure, Knoch (2009, p. 176) cautions against the use of this measure to account for paragraphing as being too mechanical: firstly, there was no penalization for short paragraphs; secondly, this measure did not account for the ordering of the information within a paragraph. Therefore, although this measure was used to assess mechanics in the AWA rating scale, the shortcomings of this measure should be noted.

### 5.1.2 Trial scale for mechanics

Since only paragraphing was investigated for the analysis of mechanics in the main analysis, the scale for mechanics is also the scale for paragraphing. The trial scale for paragraphing can be found in Table 5.2. It was decided to follow Knoch (2009) in checking the histogram while developing the descriptors, as the box plots were unable to provide detailed information on the distribution of paragraphs. The histogram showed that very small percentages of students produced 1, 2, 6, and 7 paragraphs (< 8% in total), while the majority of students produced 3, 4, and 5 paragraphs (38%, 32%, and 22% respectively). Based on these distributions and the findings above, it was decided that a range of 1–7 paragraphs should be appropriate for levels 1 to 5 in the scale for paragraphing, with three middle levels assigned to 3, 4, and 5 paragraphs, and the lowest and highest levels assigned to 1 and 2, and 6 and 7 paragraphs respectively.

Table 5.2 Trial scale – Paragraphing by level

Level	1	2	3	4	5
Description (paragraphs)	1-2	3	4	5	6-7

## 5.2 Fluency

Fluency was measured through the number of word tokens per script.

### 5.2.1 Results

Since the number of word tokens was counted using a computer program, double rating was not carried out, and hence an inter-rater reliability check was not undertaken. The average number of words of all the writing scripts was 275.93. Figure 5.2 and Table 5.3 show that the average number of words increased as the writing level rose. Even though there was a large increase in the average number of words from levels 1 to 2 in the sample,

there was much overlap between levels 3 and 4. A close check on the individual cases at each level showed that over half of the scripts at L2 produced a number of words < 200.

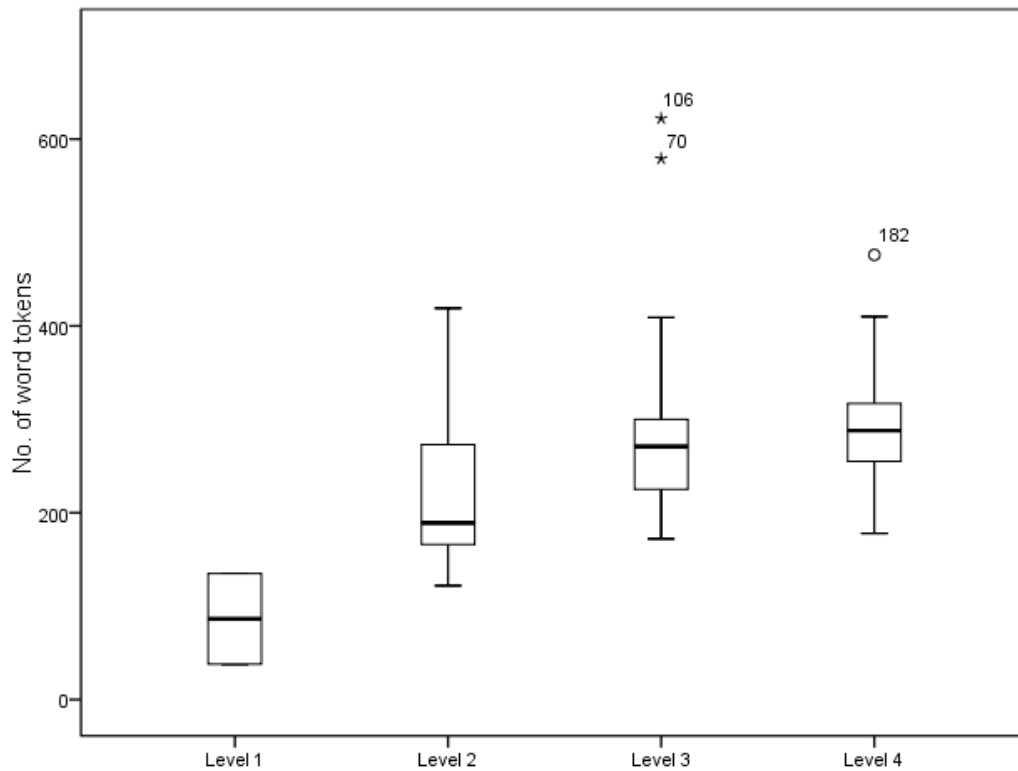


Figure 5.2 Distribution of word tokens over the AWA levels

Table 5.3 Descriptive statistics – Number of word tokens

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	86.50	68.59	38	135
2	222.14	82.31	122	419
3	275.38	71.27	172	622
4	290.33	46.59	178	476

Because the Kolmogorov-Smirnov test showed  $p = .005 < .05$ ,  $z_{skewness} = 4.57 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2 (3) = 33.047$ ,  $p = .000 < .05$ , with a statistically significant difference in the number of tokens between the different writing

proficiency levels. Stepwise step down multiple comparisons showed significant differences between all adjacent levels: levels 1 and 2, levels 2 and 3, and levels 3 and 4. Following Cohen's guidelines, eta squared = .01, small; = .06, medium; = .14, large, the effect size for the number of tokens is large, eta squared = .16.

Previous studies show that the number of words per script is a good discriminator of different proficiency levels (Hirano 1991; Homburg, 1984; Linnarud, 1986, cited in Wolfe-Quintero et al., 1998). Furthermore, studies also show that this variable has a ceiling effect around the higher level where it decreases at the advanced level (e.g., Knoch, 2009). This study found that this variable was a good indicator of four different AWA levels of Chinese EFL learners majoring in English, while it did not find the ceiling effect shown in other studies. The reason could be that there were no samples collected at level 5 (scores from 13–15), the possible advanced level of the target population.

### 5.2.2 Trial scale for fluency

The scale for fluency was largely based on the findings from the analysis. However, the levels were slightly adjusted to allow for better distinction between bands. For example, band 4 was designed to include 301–400 words although the analysis of band 4 resulted in a mean of 290.33. Since there was no sample for level 5, a number of words > 400 was added to acknowledge the potentially more fluent writers. The trial scale for fluency is presented in Table 5.4.

Table 5.4 Trial scale –Fluency by level

Level	1	2	3	4	5
Description (words)	1-100	101-200	201-300	301-400	401-

Since the scale for fluency is connected with the time limit (50 minutes in this study), this

scale should be used with caution.

### **5.3 Accuracy**

Accuracy was measured through Error-free T-unit ratio.

#### **5.3.1 Results**

An inter-rater reliability analysis was undertaken by a second coder who double coded a subset of 15 scripts for errors and T-units. A correlation analysis of the coding results of errors between two coders showed a strong correlation,  $r = .852$ ,  $n = 15$ ,  $p = .000$ . A correlation analysis of the coding results of T-units between two coders showed a strong correlation,  $r = .95$ ,  $n = 15$ ,  $p = .000$ .

The overall mean of the ratio of error-free T-units is .44. Figure 5.3 depicts the distribution of the ratio of error-free T-units. Table 5.5 shows that writers at higher levels tended to commit fewer errors than writers at lower levels. As shown in Table 5.5 and Figure 5.3, the ratio of error-free T-units successfully distinguished between the different levels, though with some overlaps between the levels.

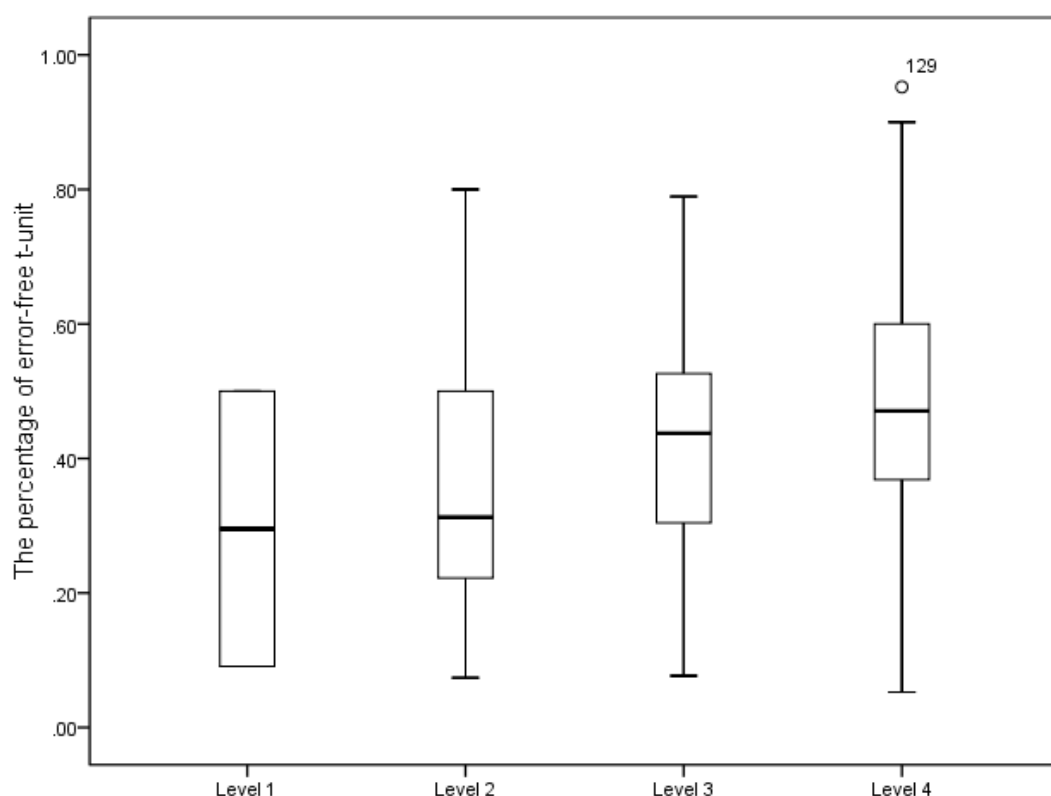


Figure 5.3 Distribution of error-free T-unit ratio over the AWA levels

Table 5.5 Descriptive statistics – Error-free T-unit ratio

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.30	.29	.09	.50
2	.36	.20	.07	.80
3	.43	.16	.08	.79
4	.47	.18	.05	.95

Because the Kolmogorov-Smirnov test showed a normal distribution for the variable ( $p = .200 > .05$ ), and the homogeneity of variance was verified ( $p > .05$ ), an analysis of variance (ANOVA) test was conducted. The one-way ANOVA test revealed significant differences between the four different proficiency levels,  $F(3, 254) = 3.99, p = .008$ . The Games-Howell post-hoc test showed no significant difference between adjacent levels. Following Cohen's guidelines, the effect size for the error-free T-unit ratio is small,  $\eta^2 = .01$ .

squared = .05.

Four measures of accuracy identified in Chapter 3 were explored in the pilot study and all were shown to successfully distinguish between different AWA levels. However, only the ratio of error-free T-units was investigated in the main analysis because it is easier to apply than other measures. As has been shown in Wolfe-Quintero et al. (1998) and the pilot analysis, the ratio of error-free T-units successfully distinguished between the four AWA levels in the main analysis, though with considerable overlap. The overlap was possibly because accuracy for some learners is non-linearly correlated with proficiency levels (e.g., Neuman, 1977, cited in Wolfe-Quintero et al., 1998), and the severity of an error and the number of errors which are sensitive to change of proficiency levels are not taken into account in the more general error-free measure (e.g., Gaies, 1980, cited in Wolfe-Quintero et al., 1998).

### **5.3.2 Trial scale for accuracy**

It was decided to check the histogram of the distribution of the ratio of error-free T-units, as the differences between different levels were not obvious and there was a considerable overlap and a wide data range (from .05–.95). The histogram showed a normal distribution for the variable (mean = .44, Std. Deviation = .18). Essays which contained nearly half of total number of T-units as error-free (Error-free T-units ratio = .50) were the most ( $n > 25$ ), while essays which did not contain error-free T-units (Error-free T-units ratio = .05) were close to none ( $n = 4$ ) and those which contained almost no errors (Error-free T-units ratio = .95) were also close to none ( $n = 2$ ).

Based on the above findings, it was decided that five levels were sufficient to scale the range of error-free T-units. Levels 1 and 5 describe the minimum and maximum values, while L3 describes the middle 50%. Levels 2 and 4 were added to account for one third and two thirds respectively. It was also decided to convert the T-units into sentences as



these are easier to apply. This was possible because a brief analysis showed that the number of T-units was nearly equal to that of sentences (with a slightly less than 90% overlap). The analytical scale for accuracy is presented in Table 5.6. The rating scale required raters to estimate the proportion of error-free sentences.

Table 5.6 Trial scale – Accuracy by level

Level	1	2	3	4	5
Description	Nearly all sentences contain one or more errors	About two thirds of sentences contain one or more errors	About half of sentences contain one or more errors	About one third of sentences contain one or more errors	Almost no sentences contain one or more errors

#### 5.4 Cohesion

Cohesion was measured through the number of conjunctions. To control for essays of different lengths, the number of conjunctions was adjusted by dividing the number of devices by the number of T-units in the essay and multiplying that number by 10 T-units to yield a frequency of cohesive devices per 10 T-units.

Inter-rater reliability for the variable was investigated by a second coder who double-coded a subset of 15 scripts. A Pearson correlation coefficient showed a strong correlation,  $r = .890$ ,  $n = 15$ ,  $p = .000$ .

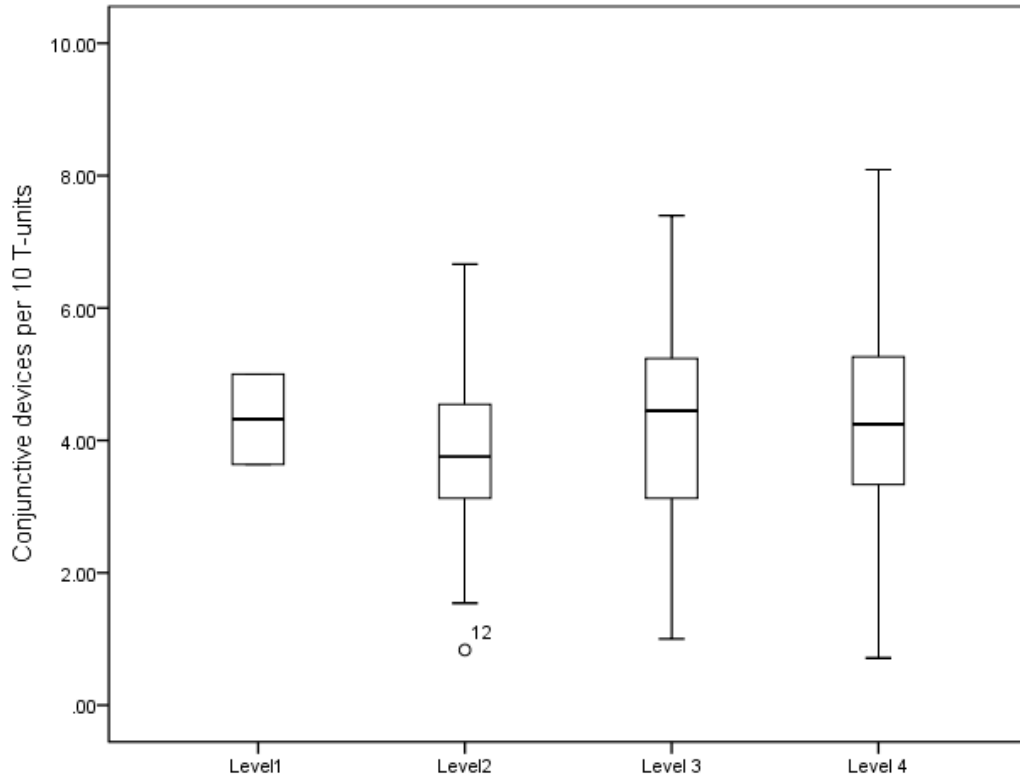


Figure 5.4 Distribution of conjunctive devices over the AWA levels

Table 5.7 Descriptive statistics – Number of conjunctive devices per 10 T-units

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	4.32	2.12	3.64	5.00
2	3.86	2.77	0.83	6.67
3	4.28	3.70	1.00	7.39
4	4.21	3.25	0.71	8.10

Figure 5.4 depicts the distribution of the number of conjunctive devices at four different AWA levels. Table 5.7 shows the descriptive statistics of the ratio of conjunctive devices per 10 T-units between four AWA levels. As shown in Table 5.7 and Figure 5.4, the number of conjunctive devices did not increase as AWA levels increased and there was a sizeable overlap between the four AWA levels.

Because the Kolmogorov-Smirnov test showed a normally distributed variable ( $p = .71 > .05$ ), and the homogeneity of variance was verified ( $p = .55 > .05$ ), an ANOVA test was conducted. The one-way ANOVA test revealed no significant difference between four different proficiency levels,  $F(3, 254) = .677, p = .567 > .05$ . Following Cohen's guidelines, the effect size for conjunctive devices is small,  $\eta^2 = .008$ .

The numbers of lexical chains, conjunctions, references and incorrect use of cohesive devices have been previously identified in Chapter 3 and trialled in the pilot study as measures for cohesion (see Section 4.6.2.6). Among these measures, only the number of conjunctions was investigated in the main analysis. References and incorrect use of cohesive devices were not investigated as they were not common in the written scripts. Lexical cohesion, though very promising, was not further investigated in the main analysis because it was difficult for our raters to apply a common rating system. The findings show that the number of conjunctions was not able to discriminate between the four AWA levels. These findings are in line with those of Yang and Sun (2012), and Neuner (1987) regarding the use of conjunctions in good and poor essays, but are in contrast with those of Witte and Faigley (1981). However, no significant finding was reported in Witte and Faigley (1981). It is thus possible that the number of conjunctions may not be a good indicator of writing quality when essay length is taken into account.

Since no measure investigated in this study was successful in distinguishing between the four AWA levels, a rating scale for cohesion was not developed.

## **5.5 Coherence**

Coherence was measured through the proportion of different types of topical progression. Following previous studies on the use of topical progression by language learners of different proficiency levels (e.g., Schneider and Connor, 1990; Knoch, 2007), I interpret the proportion of each type of topical progression (e.g., parallel progression) as the

number of T-units whose topic is coded as belonging to that type of topical progression (see Section 4.6.2.7) divided by the total number of T-units minus one T-unit. The first T-unit topic is not categorized as it is the starting point of the coding and there is no preceding T-unit topic to be related with. Statistical analysis results for the proportion of different types of topical progression are presented first in Section 5.5.1 and then a trial scale is developed based on statistical analysis in Section 5.5.2. The Results section presents coding agreement results, descriptive and inferential analysis results for different types of topical progression, and the discussion of the results in relation to previous research. At the end of the section, a summary table of the proportion of all types of topical progression is provided to better understand their distribution across four levels. It should be noted that, although only SP3 discourse-related sequential progression was found to be successful in the pilot study, it was decided that all different types of topical progression would be investigated in the main study as different types of topical progressions are related (see inconclusive selection principle in Section 4.6.2).

### 5.5.1 Results

Before an analysis of coherence was undertaken, an inter-rater reliability analysis was conducted. The results for each type are shown in Table 5.8. Since this variable is highly inferential, reliability coefficients for SP3 discourse-related sequential progression and unrelated sequential progression  $< .80$  (the commonly acceptable cut-off level) were regarded as acceptable.

Table 5.8 Inter-rater reliability for topical progression types

Topical progression type	Correlation coefficients
Parallel progression	$r = .944, N = 15, p = .000$
Extended parallel progression	$r = .804, N = 15, p = .000$
Related sequential progression	$r = .784, N = 15, p = .001$

SP1 sequential progression	$r = .873, N = 15, p = .001$
SP2 sequential progression	$r = .823, N = 15, p = .001$
SP3 discourse-related sequential progression	$r = .750, N = 15, p = .000$
Unrelated topical progression	$r = .776, N = 15, p = .000$
Extended sequential progression	$r = .710, N = 15, p = .000$

Among different types of topical progression investigated in this study, parallel progression, extended parallel progression, and sequential progression were selected from Lautamatti (1978) and Schneider and Connor (1990); extended sequential progression and unrelated topical progression taken from Simpson (2000) and Schneider and Connor (1990) respectively but adapted to suit the data in this study. SP1–SP3 were developed to facilitate the coding of sequential progression, as the subtypes of direct and indirect sequential progression proposed by Schneider and Connor (1990) are ambiguous (see Section 4.6.2.7). The investigation of three refined sequential progression subtypes (SP1–SP3) was expected to interpret how these componential subtypes of sequential progression might distinguish between the four levels of progression and provide more details for the development of descriptors for sequential progression types.

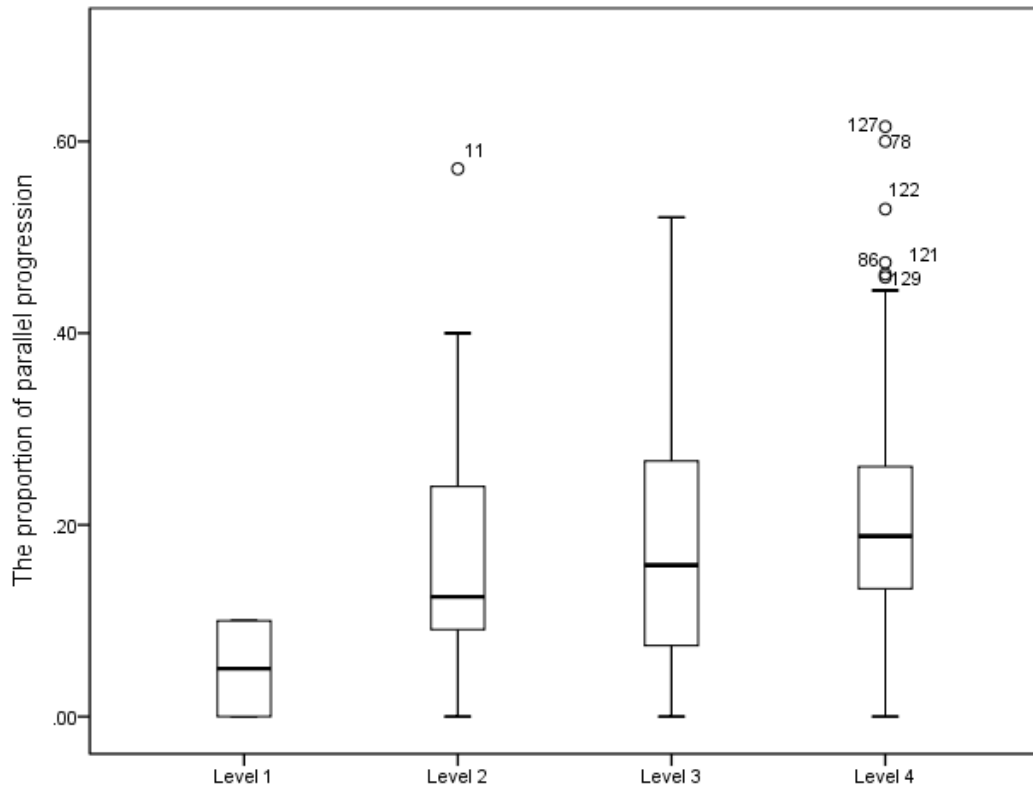


Figure 5.5 Distribution of parallel progression over the AWA levels

Table 5.9 Descriptive statistics – Proportion of parallel progression (number of T-units whose topic is coded as belonging to parallel progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.05	.07	.00	.10
2	.17	.14	.00	.57
3	.18	.13	.00	.52
4	.21	.12	.00	.62

Figure 5.5 shows that the use of parallel progression increased as writing quality increased, though with some overlap. The proportion of parallel progression is the number of T-units whose topic is coded as belonging to parallel progression, divided by (total number of T-units minus one). Table 5.9 lists the descriptive statistics of the proportion of parallel progression between four AWA levels. A large increase in the proportion of

parallel progression occurred from L1 to L2.

Because the Kolmogorov-Smirnov test showed  $p = .001 < .05$ ,  $z_{skewness} = 4.81 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 10.503$ ,  $p = .015 < .05$ , with a statistically significant difference in the proportion of parallel progression between the four different proficiency levels. Stepwise step down multiple comparisons showed that significant differences were found between adjacent levels 3 and 4. According to Cohen's guideline, the effect size investigated for the parallel progression is small,  $\eta^2 = .03$ .

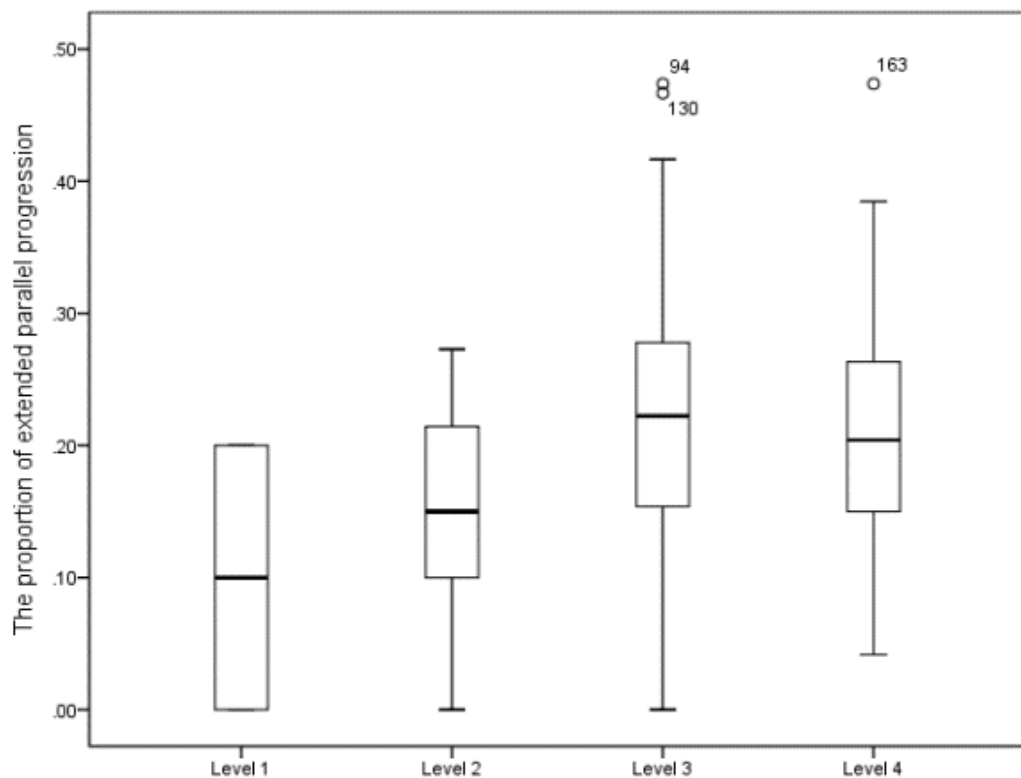


Figure 5.6 Distribution of extended parallel progression over the AWA levels

Table 5.10 Descriptive statistics – Proportion of extended parallel progression (number of T-units whose topic is coded as belonging to extended parallel progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.10	.14	.00	.20
2	.15	.09	.00	.27
3	.22	.10	.00	.47
4	.21	.09	.04	.47

In Figure 5.6 it can be seen that the use of the extended parallel progression steadily increased from L1 to L3, and decreased slightly from L3 to L4. The proportion of extended parallel progression is the number of T-units whose topic is coded as belonging to extended parallel progression, divided by (total number of T-units minus one). Table 5.10 shows the descriptive statistics of the proportion of extended parallel progression between four different AWA levels.

Because the Kolmogorov-Smirnov test showed a normal distribution for the variable ( $p = .188 > .05$ ), and the homogeneity of variance was verified ( $p = .459 > .05$ ), an ANOVA was conducted. The one-way ANOVA test revealed significant differences between four different proficiency levels,  $F(3, 254) = 4.923$ ,  $p = .002$ . The Games-Howell post hoc test shows statistically significant difference between adjacent levels 2 and 3 ( $p = .004$ ). According to Cohen's guideline, the effect size investigated for the extended parallel progression is medium,  $\eta^2 = .06$ .

The related sequential progression includes SP1, SP2, and SP3 sequential progressions.



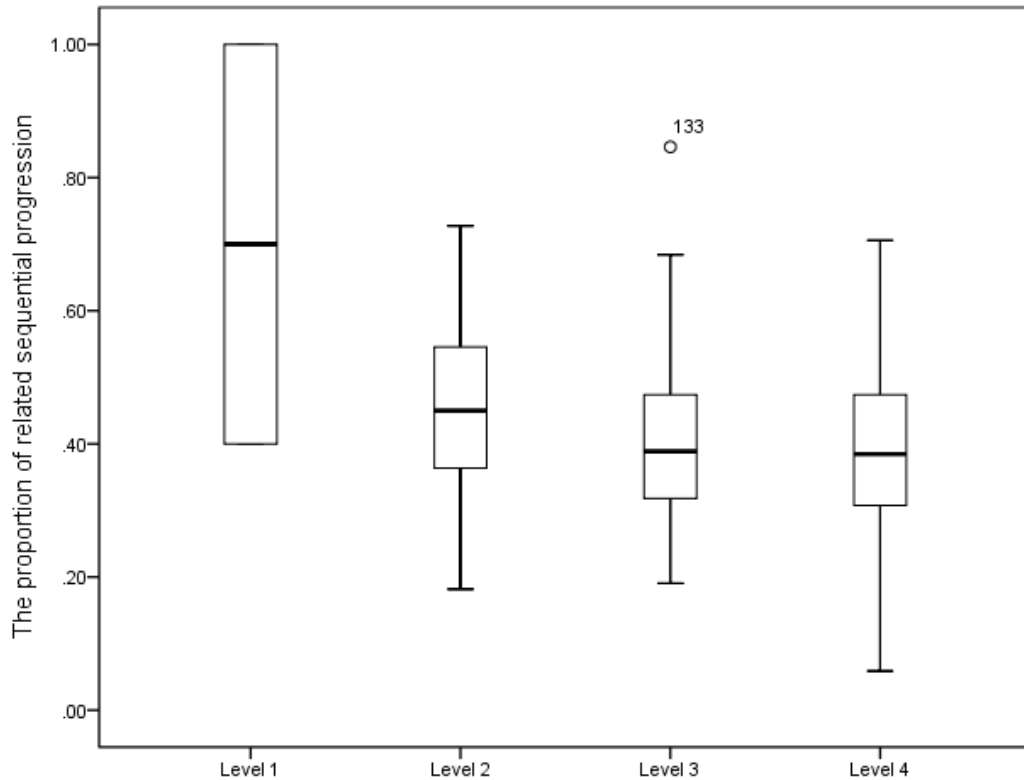


Figure 5.7 Distribution of related sequential progression over the AWA levels

Table 5.11 Descriptive statistics – Proportion of related sequential progression (number of T-units whose topic is coded as belonging to sequential progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.70	.42	.40	1.00
2	.45	.13	.18	.73
3	.41	.12	.19	.85
4	.40	.13	.06	.71

Table 5.11 lists the descriptive statistics of the proportion of related sequential progression across four different AWA levels. The proportion of related sequential progression is the number of T-units whose topic is coded as belonging to sequential progression, divided by (total number of T-units minus one). As can be seen in Table 5.11 and Figure 5.7, the proportion of related sequential progression decreased as writing levels increased. For

writers at L1 and L2, related sequential progression decreased dramatically from 70% to 45% at L2. The variable decreased from 45% at L2 to 41% at L3, and further decreased to 40% at L4.

Because the Kolmogorov-Smirnov test showed  $p = .029 < .05$ ,  $z_{skewness} = 3.52 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2 (3) = 5.721$ ,  $p = .126 > .05$ , with no statistically significant difference in the proportion of related sequential progression between the different proficiency levels. According to Cohen's guideline, the effect size investigated for the related sequential progression is small, eta squared = .05.

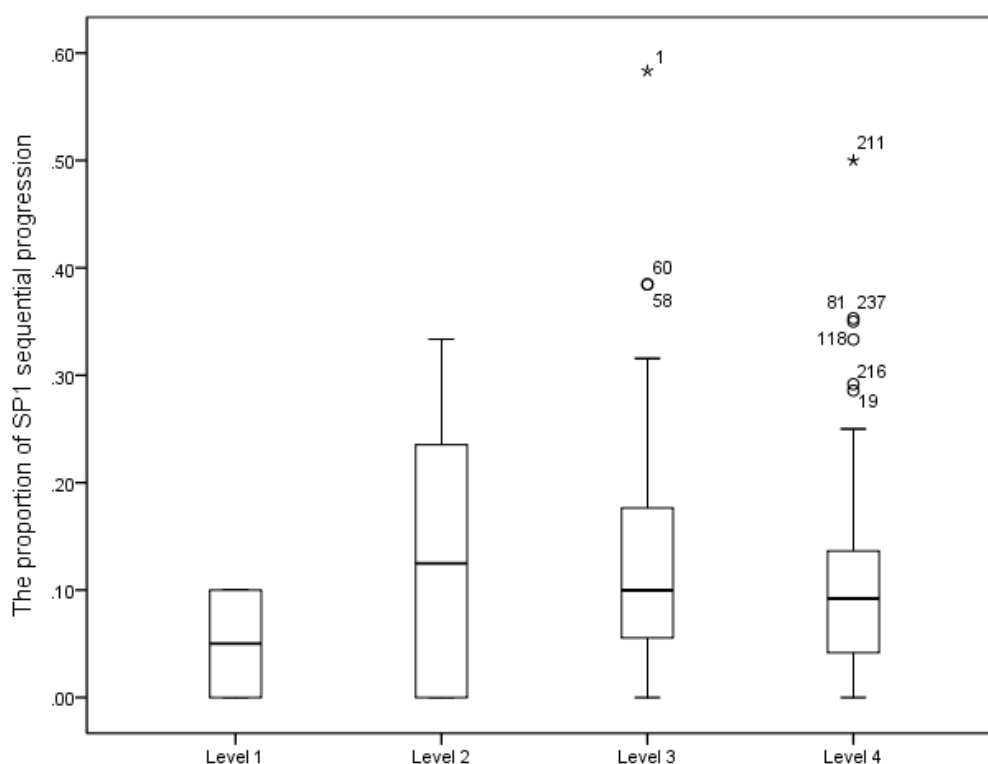


Figure 5.8 Distribution of the SP1 sequential progression over the AWA levels

Table 5.12 Descriptive statistics – Proportion of SP1 sequential progression (number of T-units whose topic is coded as belonging to SP1 sequential progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.05	.07	.00	.10
2	.14	.11	.00	.33
3	.12	.10	.00	.58
4	.10	.09	.00	.50

Figure 5.8 shows that SP1 sequential progression was least used by writers at L1. The proportion of SP1 sequential progression is the number of T-units whose topic is coded as belonging to SP1 sequential progression, divided by (total number of T-units minus one). Table 5.12 lists the descriptive statistics of the proportion of SP1 sequential progression at all four levels. As can be seen in Table 5.12, the proportion of SP1 sequential progression used by writers at L1 is as low as 5%, while that used by writers at levels 2, 3 and 4 are similar. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 8.66 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 5.471$ ,  $p = .140 > .05$ , with no statistically significant difference in the proportion of SP1 sequential progression between the different proficiency levels. According to Cohen's guideline, the effect size investigated for the SP1 sequential progression is small,  $\eta^2 = .02$ .

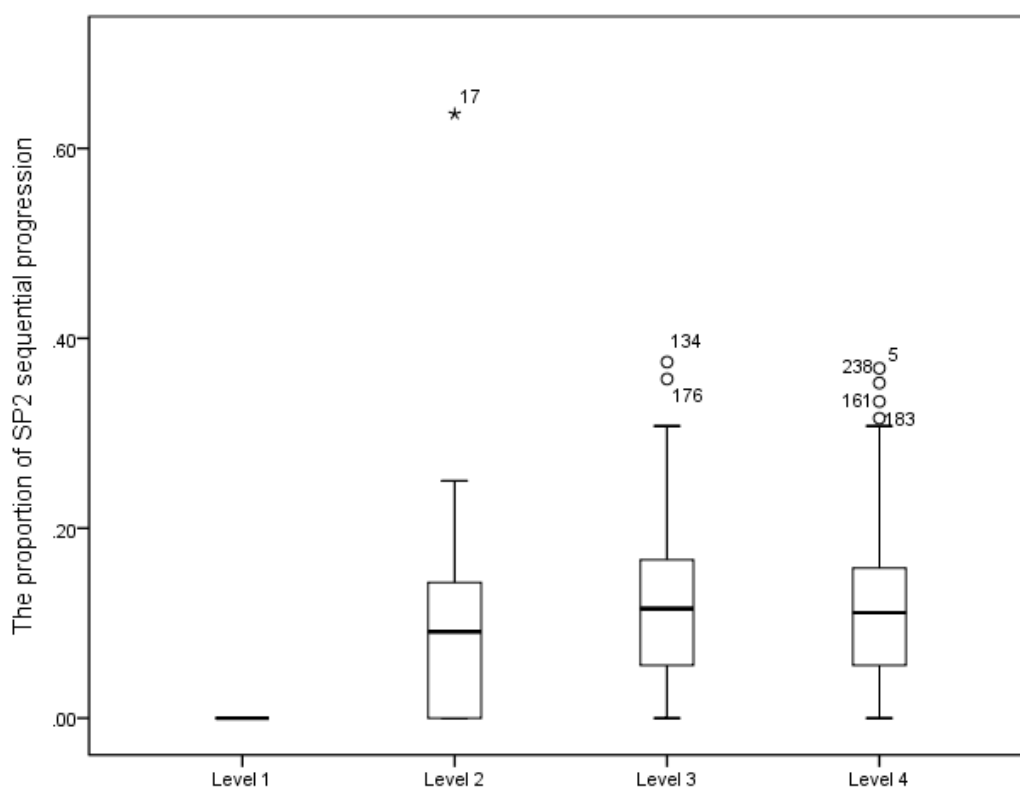


Figure 5.9 Distribution of the SP2 sequential progression over the AWA levels

Table 5.13 Descriptive statistics – Proportion of SP2 sequential progression (number of T-units whose topic is coded as belonging to SP2 sequential progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.11	.13	.00	.64
3	.12	.08	.00	.38
4	.12	.08	.00	.37

Figure 5.9 and Table 5.13 show that writers at L1 did not use SP2 sequential progression, while writers at levels 2, 3, and 4 used a similar proportion of SP2 sequential progression. The proportion of SP2 sequential progression is the number of T-units whose topic is coded as belonging to SP2 sequential progression, divided by (total number of T-units minus one). Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} =$

7.86 > 2, and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 5.597$ ,  $p = .133 > .05$ , and no statistically significant difference was found in SP2 sequential progression between the different writing proficiency levels. According to Cohen's guideline, the effect size investigated for the SP2 sequential progression is small,  $\eta^2 = .01$ .

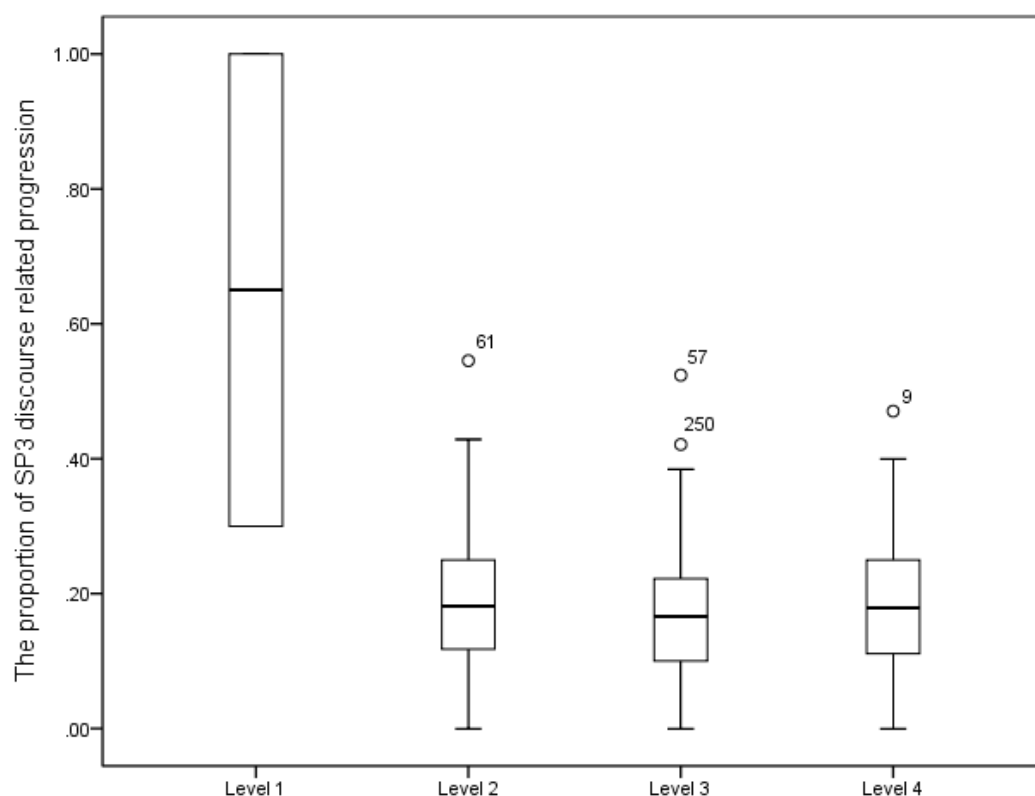


Figure 5.10 Distribution of SP3 discourse-related sequential progression over the AWA levels

Table 5.14 Descriptive statistics – Proportion of SP3 discourse-related sequential progression (number of T-units whose topic is coded as belonging to SP3 discourse-related sequential progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.65	.49	.30	1.00
2	.20	.13	.00	.55
3	.17	.10	.00	.52

4	.18	.10	.00	.47
---	-----	-----	-----	-----

Table 5.14 and Figure 5.10 show an upside-down u-shape distribution. The proportion of SP3 discourse related sequential progression was mostly used (> 60%) by writers at L1. The variable decreased to a relatively low value at levels 3 and 4 (17% and 18% respectively). The proportion of SP3 discourse-related sequential progression is the number of T-units whose topic is coded as belonging to SP3 discourse-related sequential progression, divided by (total number of T-units minus one). Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 12.14 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 5.878$ ,  $p = .118 > .05$ , and no statistically significant difference was found in discourse-related sequential progression between the different proficiency levels. According to Cohen's guideline, the effect size investigated for the discourse-related progression is large,  $\eta^2 = .14$ .

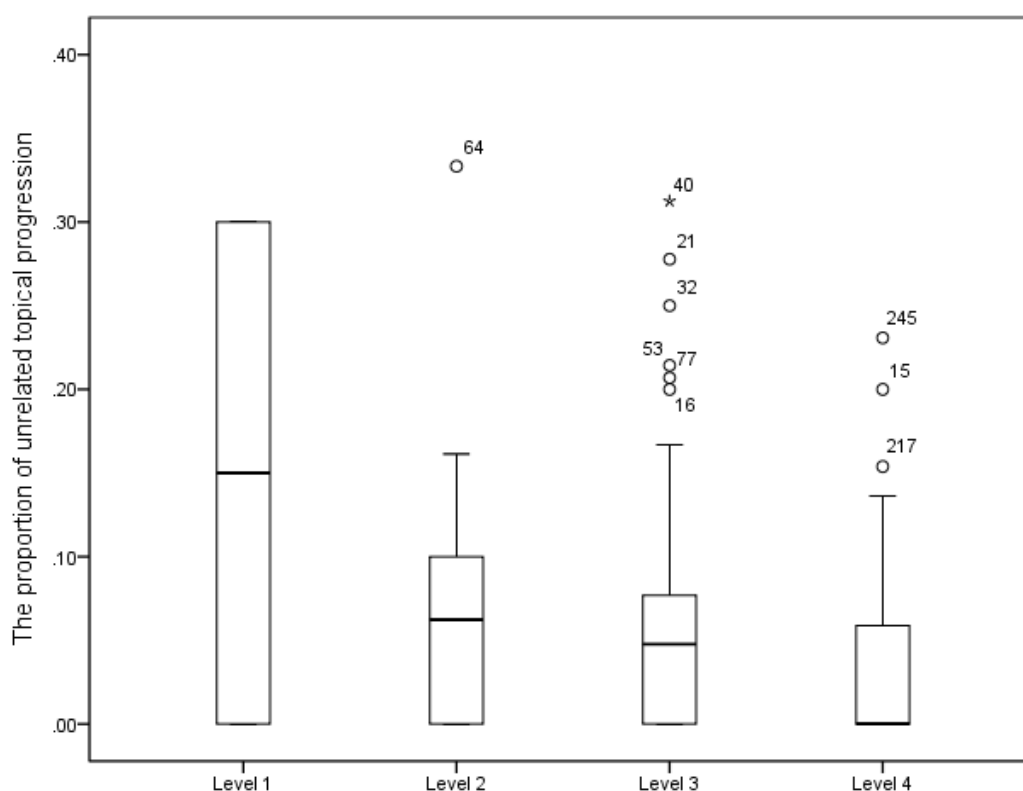


Figure 5.11 Distribution of unrelated topical progression over the AWA levels

Table 5.15 Descriptive statistics – Proportion of unrelated topical progression (number of T-units whose topic is coded as belonging to unrelated topical progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.15	.21	.00	.30
2	.06	.07	.00	.33
3	.05	.07	.00	.31
4	.03	.05	.00	.23

As shown in Table 5.15 and Figure 5.11, the proportion of unrelated topical progression decreased as the four AWA levels increased. The proportion of unrelated topical progression is the number of T-units whose topic is coded as belonging to unrelated topical progression, divided by (total number of T-units minus one). Higher level writers produced less unrelated sequential progression, and lower level writers produced more unrelated topical progression. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 12.74 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 8.070$ ,  $p = .045 < .05$ , and there was a statistically significant difference in the proportion of unrelated topical progression between the different AWA levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 3 and 4. According to Cohen's guideline, the effect size investigated for the unrelated topical progression is medium,  $\eta^2 = .06$ .

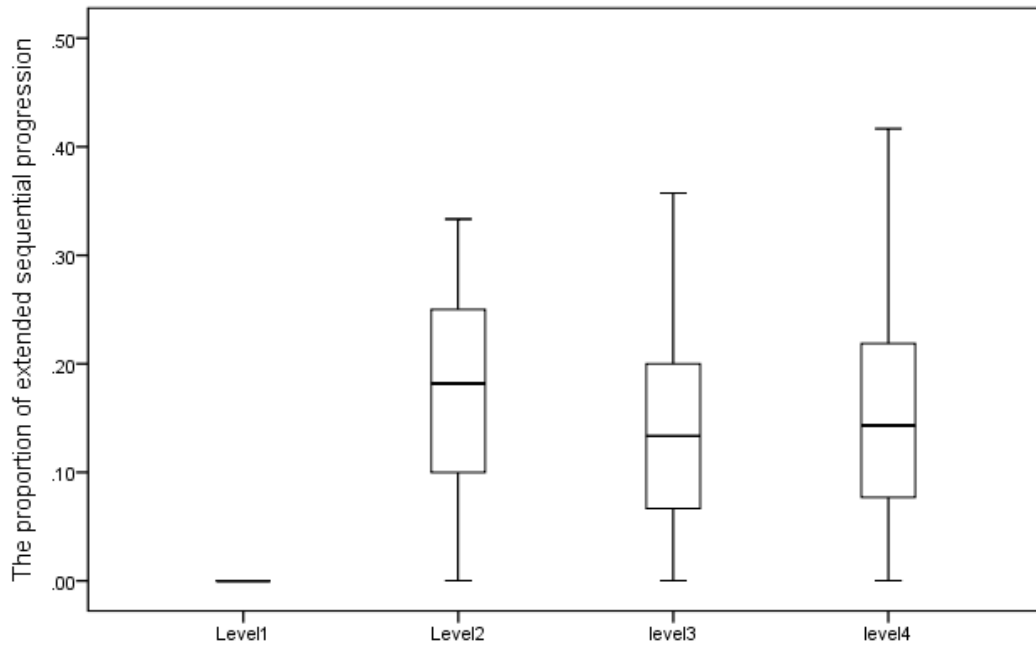


Figure 5.12 Distribution of extended sequential progression over the AWA levels

Table 5.16 Descriptive statistics – Proportion of extended sequential progression (number of T-units whose topic is coded as belonging to extended sequential progression, divided by (total number of T-units minus one))

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.17	.09	.00	.33
3	.14	.10	.00	.36
4	.15	.09	.00	.42

As shown in Table 5.16 and Figure 5.12, writers at L1 did not use extended sequential progression while writers at L2 used on average 17% of T-units on extended sequential progression. Writers at levels 3 and 4 used a similar proportion of extended sequential progression as those at L2. The proportion of extended sequential progression is the number of T-units whose topic is coded as belonging to extended sequential progression, divided by (total number of T-units minus one). Because the Kolmogorov-Smirnov test showed  $p = .023 < .05$ ,  $z_{skewness} = 2.17 > 2$ , and as the variable was clearly not normally



distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 8.487$ ,  $p = .037 < .05$ , with a statistically significant difference in the proportion of extended sequential progression between the different AWA levels. Stepwise step down multiple comparisons showed that significant differences were found between adjacent levels 1 and 2. According to Cohen's guideline, the effect size investigated for the extended sequential progression is small,  $\eta^2 = .03$ .

The analysis showed that linear relationship and statistical significance were both found for parallel progression and unrelated topical progression; statistical significance was found for extended parallel progression and extended parallel progression, but no linear relationship and no salient level was found for them; statistical significance and a salient level 1 was found for SP3 progression while no linear relation was found. Table 5.17 presents the mean of all the different topical progression types at four levels. As is clearly demonstrated in the table, writers at L1 produced more unrelated topical progression (i.e., topics that are neither related to the discourse topic or comments made in previous sentences) than higher level writers. This finding is in line with Knoch (2007) who also found that lower level writers use more unrelated topical progression. However, it was surprising to see that higher level writers used more parallel progression than lower level writers, which was contrary to the findings in previous research (e.g., Knoch, 2007, Schneider and Connor, 1990). The underuse of parallel progression at lower level writing can be attributed to the overuse of sequential progression, especially discourse related sequential progression. Discourse related sequential progression, a new subcategory within related sequential progression, is largely comprised of the use of 'we', 'you', and 'our' as the topic of the T-units. The use of this subcategory does not seem to contribute to high writing quality, possibly because these topics, though related to discourse, might not be close to the discourse topics. Witte (1983) also describes how sequential progression which did not contribute to the discourse topic might be the reason for low level writing. As with Burneikaite and Zabaliute (2003), high level writers tended to strike a

balance between the parallel and extended parallel progression, though no significant difference was found in the use of extended parallel progression between the four different levels.

Table 5.17 Summary table – Proportion of different topical progression types

Topical progression type	Level 1	Level 2	Level 3	Level 4
Parallel progression	.05	.17	.18	.21
Extended parallel progression	.10	.15	.22	.21
Related sequential progression	.70	.45	.41	.40
SP1 sequential progression	.05	.14	.12	.10
SP2 sequential progression	.00	.11	.12	.12
SP3 discourse-related sequential progression	.65	.20	.17	.18
Unrelated topical progression	.15	.06	.05	.03
Extended sequential progression	.00	.17	.14	.15

### 5.5.2 Trial scale for coherence

The rating scale based on these findings is presented in Table 5.18.

Table 5.18 Trial scale – Coherence

Level	Description
4	Unrelated topic progression is rarely seen, but when present, it is 1-2 unrelated topical progression.  Mixture of parallel, sequential and extended parallel and extended sequential progression. Discourse-related topics such as ‘we’, ‘you’, ‘our...’, ‘everyone’ are rarely seen.
3	Unrelated topic use is infrequent; 1 or 2 occurrences only.

	Mixture of parallel, sequential and extended parallel and extended sequential progression. Discourse-related topics such as 'we', 'you', 'our...', 'everyone' are sometimes seen.
2	<p>Unrelated topic progression is frequent; 3 or more occurrences are occasionally found.</p> <p>Mixture of parallel, sequential and extended parallel and extended sequential progression. Discourse-related topics such as 'we', 'you', 'our...', 'everyone' are more frequent than other discourse-related topics.</p>
1	<p>Unrelated topic progression is frequent; 3 or more occurrences are often found.</p> <p>Approximately 2/3 sentence topics are discourse-related, most are 'we', 'you', 'our...', 'everyone'.</p> <p>There is no use of extended sequential progression.</p>

The design of the trait scale for coherence was more difficult because the results for a number of categories needed to be considered and synthesized. It was decided that only four levels should be included in the rating scale because the analysis of scripts was based only on levels 1 to 4, and it was difficult to envisage a synthesized use of any particular categories of topical progression at level 5. For example, it is possible to expect no errors at an ideally high level, while it is very difficult to imagine how categories of topical progression can be synthesized at an ideally high level. Next, it was decided to check the case summaries of the use of successful measures and discourse-related sequential topics using SPSS (IBM Corp, 2013) to include the frequency of these features commonly or not commonly expected by the raters. Four levels were subsequently scaled based on the findings of the analysis and case summaries of these features.

## 5.6 Argument structure

### 5.6.1 Results

Before the analysis of the argument structure, an inter-rater reliability analysis was conducted. The results for each structure are presented in the Table 5.19.

Table 5.19 Inter-rater reliability for argument structural elements

Argument structural element	Correlation coefficients
Introduction	$r = 1.000$ , $N = 15$ , $p = .000$
Level-1 reasons	$r = .774$ , $N = 15$ , $p = .000$
Level-2 reasons and below	$r = .942$ , $N = 15$ , $p = .000$
Standpoint	$r = 1.000$ , $N = 15$ , $p = .000$
Yourside argument	$r = .747$ , $N = 15$ , $p = .000$
Functional markers	$r = .886$ , $N = 15$ , $p = .000$
Non-functional elements	$r = .896$ , $N = 15$ , $p = .000$
Conclusion	$r = .832$ , $N = 15$ , $p = .000$

Since this variable is highly inferential, reliability coefficients for level-1 reasons and yourside argument  $< .80$  (the commonly acceptable cut-off level) were regarded as acceptable.

Argumentation structure was measured by the proportion of its various components. That is, the number of different types of argumentative structural elements divided by the total number of all argumentative structural elements. The elements are essentially T-units except for reason elements which can be a clause, a phrase or a T-unit. The argumentative structural types include: introduction, standpoints, level-1 reasons, level-2 reasons and below, yourside arguments (including counterarguments, reasons for counterarguments, alternative standpoints, rebuttals, reasons for rebuttals), functional markers, non-

functional elements, and conclusion. The basic assumption is that the proportions of level-2 reasons and below, and yourside argument increase as writing quality increases, as level-2 reason and below is indicative of the depth of the reasoning, and yourside arguments is indicative of a more balanced argument, thus being more convincing. Specific types of argumentation structure included in yourside argument were investigated to provide detailed information as to how yourside arguments are able to distinguish between different writing levels. Moreover, the proportions of introduction and non-functional elements were shown in the pilot analysis to be disproportionate at level one than other levels, with introduction being exceptionally larger in proportion at low level than middle and high levels, and non-functional elements smaller in proportion at high level than the other two levels. These two elements were expected to distinguish one level from the other levels in the main analysis. These components were investigated in this study.

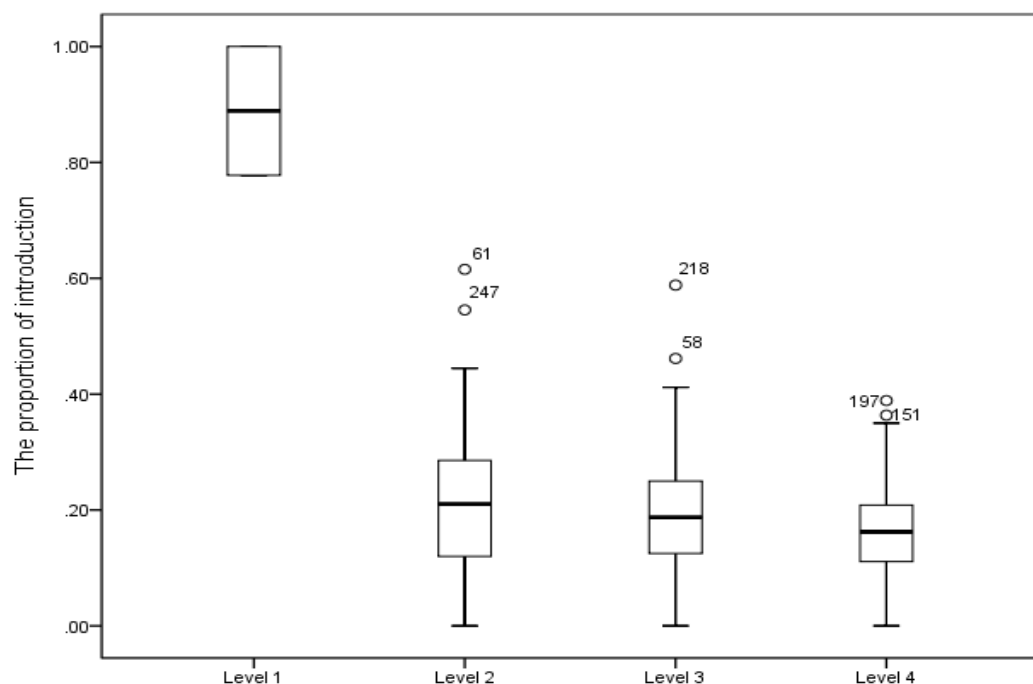


Figure 5.13 Distribution of introduction

Table 5.20 Descriptive statistics – Proportion of introduction (number of introduction elements divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.89	.16	.78	1.00
2	.23	.14	.00	.62
3	.20	.10	.00	.59
4	.16	.08	.00	.39

Figure 5.13 depicts the distribution of introduction. As the level increases, the proportion of introduction decreases. In particular, writers at L1 allocated 89% of the writing to the introduction. Writers at levels 2, 3 and 4 used a much smaller proportion of writing for the introduction. Table 5.20 shows the descriptive statistics of the proportion of introduction across four different AWA levels.

Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 15.66 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 17.743$ ,  $p = .000 < .05$ , with a statistically significant difference in the proportion of introduction between the different AWA levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 1 and 2, and levels 3 and 4. The effect size for introduction is large,  $\eta^2 = .33 > .14$ .

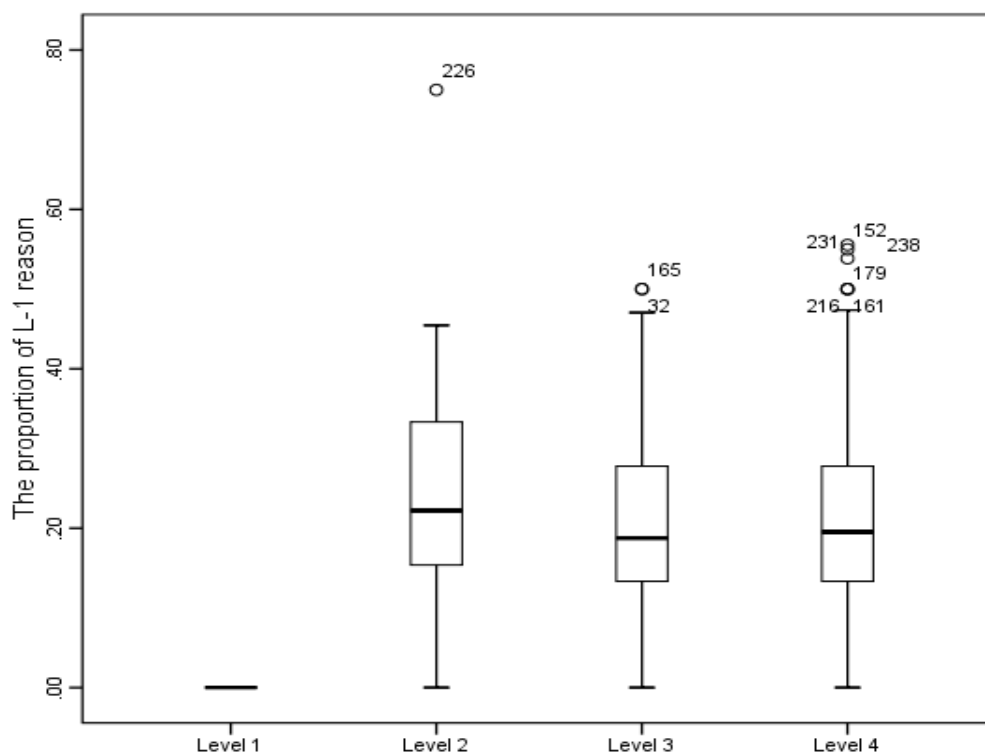


Figure 5.14 Distribution of level-1 reasons

Figure 5.14 depicts the distribution of level-1 reasons. Table 5.21 shows the descriptive statistics of the proportion of level-1 reasons across four different AWA levels. There is no clear linear relationship between the four AWA levels. Writers at L1 did not produce level-1 reasons. A scrutiny of the L1 scripts showed that writers at this level were keener on introducing background of a controversial topic, while they left little time to elaborate on their standpoints. A sample text of L1 scripts is presented in Figure 5.15. In the first paragraph of the sample text (coded “I”- introduction), the student writer expands on the background to how seriously the air is polluted and how alternative measures are proposed to resolve the problem. Then in the second paragraph, the student writer provides his or her standpoint, which is coded “ST”-standpoint as it is hard to identify whether it is positive (i.e., closing down factories) or negative (i.e., not closing down factories). It is not clear why the student writer failed to elaborate on his or her standpoint

and left the essay unfinished.

Factories be closed down?

1.As times goes on, it is more and more thorny and noticed that air pollution have been influenced on humanity. (I) 2. It is luxury for human being to sightsee the blue sky and take fresh air. (I) 3. Since the less trees are planted, the more factories are built in our planet. (I) 4. What we see is the sky filled with brown, (I) 5. and human are busy with other issues./ (I) 6. However, some people reckon that we can do something to solve it and make our air more clean and we can be harmony with our planet. (I) 7. But how? 8. Just closed down factories? (I)

9. In my opinion, it is a complex and systematic issues, (ST) 10. we must find out ways to achieve win-win.(ST) 11. If you just close down factories or plant more trees, which we do not support and accomplemant it (NFU).

Figure 5.15 Sample text at level 1 with a disproportionate introduction

Since the sample size for L1 was small, the proportion of level-1 reasons could potentially be higher with a larger sample size.

Table 5.21 Descriptive statistics – Proportion of level-1 reasons (number of level-1 reason elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.25	.15	.00	.75
3	.21	.11	.00	.50
4	.22	.12	.00	.56

Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 5.66 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 7.232$ ,  $p = .065 > .05$ , and no statistically



significant difference was found in the proportion of level-1 reasons between the different AWA levels. The effect size for level-1 reasons is small, eta squared = .04. This is possible as writers at different levels are able to provide at least level-1 reasons to defend their standpoints, leaving few standpoints unjustified.

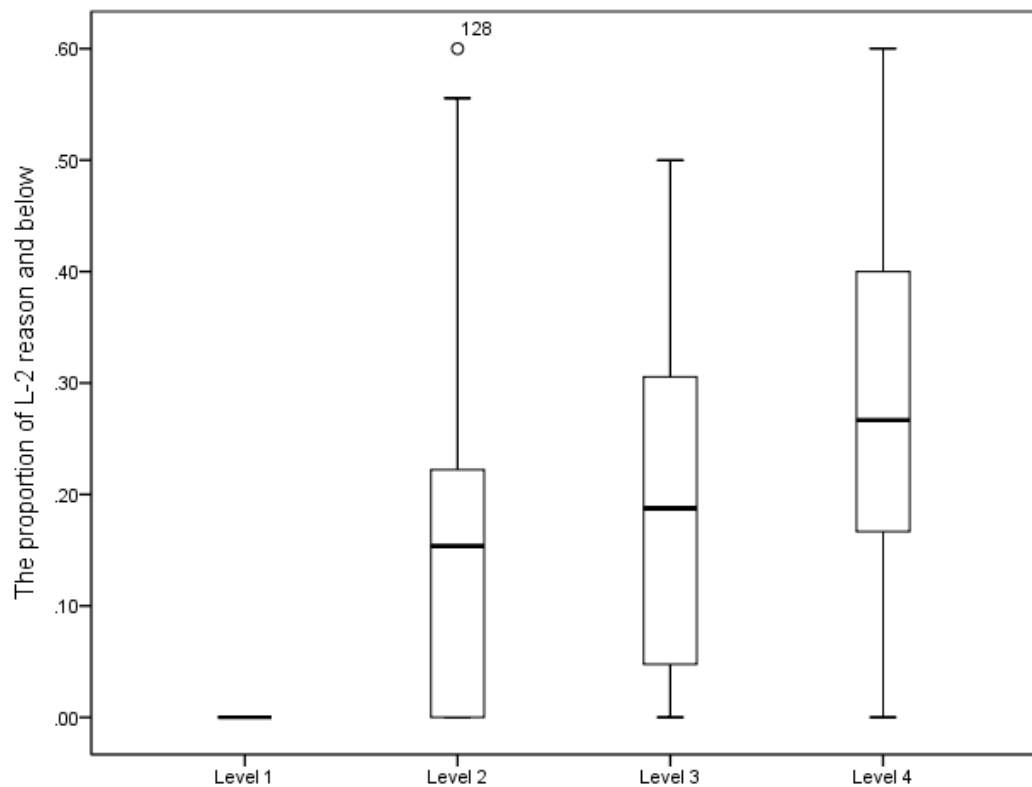


Figure 5.16 Distribution of level-2 reasons and below

Figure 5.16 depicts the distribution of level-2 reasons and below which shows a clear increase in the use of level-2 reasons and below as the level of writing increases.

Table 5.22 Descriptive statistics – Proportion of level-2 reasons and below (number of level-2 reason elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.17	.20	.00	.60

3	.20	.16	.00	.50
4	.27	.17	.00	.60

Table 5.22 shows the descriptive statistics of the proportion of level-2 reasons and below between the four different AWA levels. As shown in Table 5.22, the proportion of level-2 reasons and below distinguishes between the four AWA levels. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 1.29 > 2$ , and the variable was normally distributed with an equality of variance assumption ( $p = .093 < .05$ ) that could be verified in this case, an ANOVA test was conducted. The one-way ANOVA test revealed significant differences between the four different proficiency levels,  $F(3, 254) = 6.671$ ,  $p = .000$ . The Games-Howell post-hoc test shows no statistically significant differences between adjacent levels. The effect size for level-2 reasons and below is medium,  $\eta^2 = .07$ . This finding is consistent with previous research (e.g., Crammond, 1997) which shows that the depth of argument structure is a good indicator of writing proficiency.

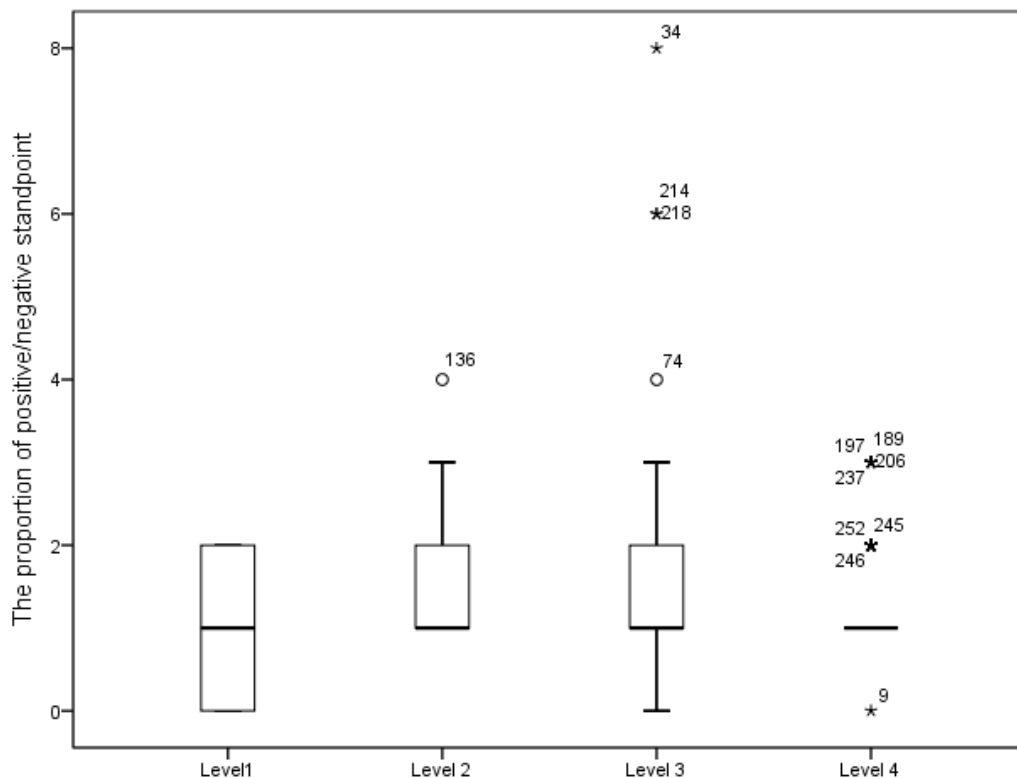


Figure 5.17 Distribution of standpoints

Table 5.23 Descriptive statistics – Proportion of standpoints (number of standpoint elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.11	.16	.00	.22
2	.12	.09	.04	.40
3	.09	.07	.00	.38
4	.08	.04	.00	.27

Table 5.23 shows the descriptive statistics of the proportion of standpoints across the four different AWA levels. As shown in Table 5.23 and Figure 5.17, the proportion of standpoints was not able to distinguish between the four AWA levels as the group mean was roughly similar. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 18.70 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 13.446$ ,  $p = .004 < .05$ , with a statistically significant difference in the proportion of standpoints between the different AWA levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 1 and 2, and levels 3 and 4. The effect size for the standpoints is medium, eta squared = .06. The analysis shows that writers at different levels devote a similar proportion of text to standpoints. It is noted that in the analysis of scripts some writers were found to expand on and compare the influences of closing factories or not closing factories before presenting their standpoints in the conclusion part. In the coding of argument structure, standpoints which are presented in this way were coded as conclusions rather than standpoints. This is why zero occurrences of standpoints are found in some scripts at levels 1, 3 and 4 (see Table 5.23).

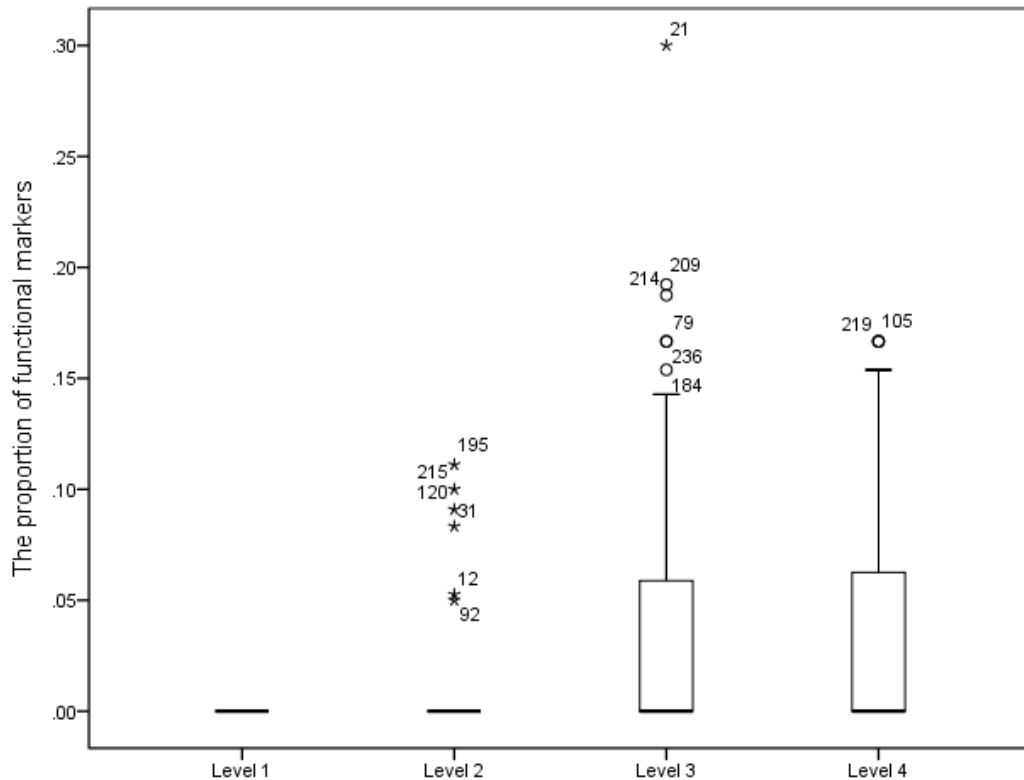


Figure 5.18 Distribution of functional markers

As is visually depicted in Figure 5.18, the proportion of functional markers (information used as a transition to introduce reasons, arguments, and standpoints) is very small and similar at the different levels.

Table 5.24 Descriptive statistics – Proportion of functional markers (number of functional marker elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.02	.04	.00	.11
3	.04	.06	.00	.30
4	.03	.04	.00	.17

Figure 5.18 and Table 5.24 show the descriptive statistics of the proportion of functional markers across the four different AWA levels. The proportion of functional markers was

not able to distinguish between the four AWA levels. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 11.53 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 6.105$ ,  $p = .107 > .05$ , with no statistically significant difference in the proportion of functional markers between the different AWA levels. The effect size on functional markers is small, eta squared = .02. The analysis shows that functional markers were not common in students' scripts and no salient occurrence of functional markers was found at any level.

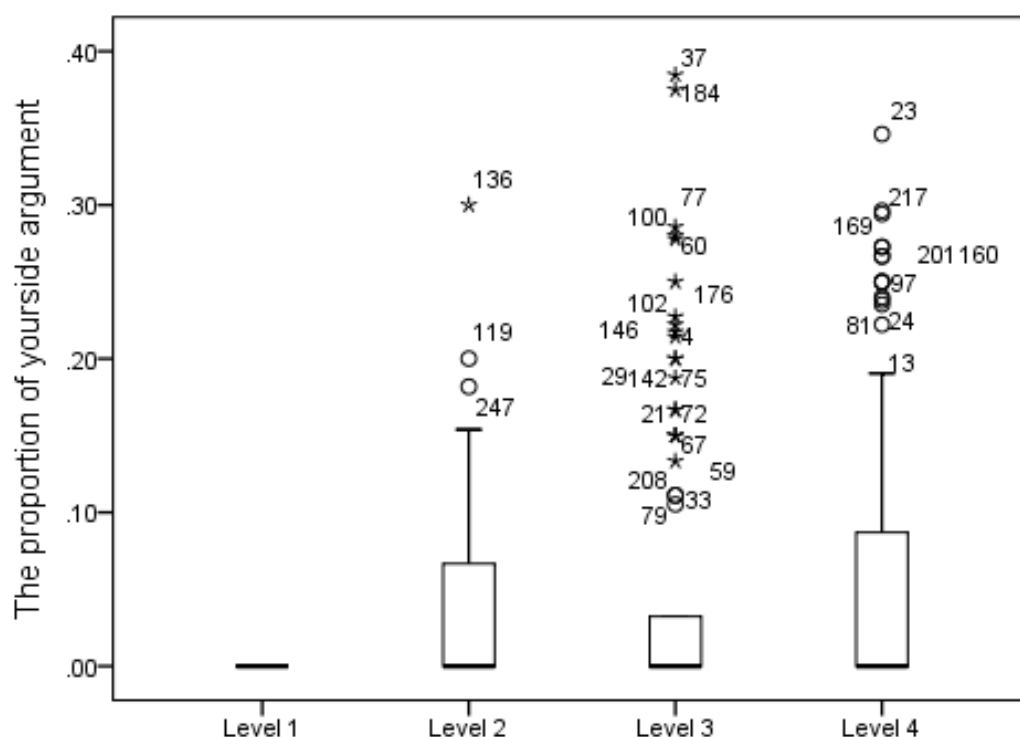


Figure 5.19 Distribution of yourside arguments

Table 5.25 Descriptive statistics – Proportion of yourside arguments (number of yourside argument elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.04	.08	.00	.30

3	.05	.10	.00	.38
4	.05	.09	.00	.35

---

Yourside argument includes counterarguments, reasons for counterarguments, alternative standpoints, rebuttals, and reasons for rebuttals. Figure 5.19 and Table 5.25 indicate that the proportion of yourside arguments is very small ( $< 10\%$ ), and level-1 writers did not produce otherside arguments. Though the mean at each level increased, the difference between the levels was very small. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 11.43 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = .703$ ,  $p = .873 > .05$ , and no statistically significant difference was found in the proportion of yourside arguments between the different AWA levels. The effect size on otherside arguments is small, eta squared = .003. There were also quite a number of outliers at levels 3 and 4, and outliers at level 3 were more than and more distant than those at level 4 as are indicated by asterisks and circles.

The analysis shows that yourside arguments were not common in students' essays. A close scrutiny of individual scripts showed that the majority of students did not write yourside arguments at all. A typical writing script is where writers put forward their standpoint in the first paragraph, defend their standpoint with reasons in two or three paragraphs, and then conclude their arguments in the last paragraph. This is probably why the proportion was very low at four levels. A close scrutiny of individual scripts also showed that writing scripts at level 3 with a similar proportion of yourside arguments were poorer in language use than writing scripts at level 4. For those who wrote yourside arguments, the yourside arguments were immature and not properly defended. In Figure 5.20, for example, the student writer argues that shutting down factories is not the cure for the air pollution because of a series of social and economic problems that may arise due to shutting down factories. In the last reason, the writer acknowledges a counterargument '*shut (shutting)*

*down these factories is useful for the air pollution'* that undermines the main standpoint, but immediately after it he or she attempts to rebut the counterargument by stating “*but there are also many (other) source(s) of pollution, such as cars and others*”. The writer attempts to imply that factories are only one of many sources of air pollution, therefore, shutting them down cannot fundamentally solve the problem. These yoursides arguments were underdeveloped, as their meaning needed to be inferred, or were poorly constructed such as with the wrong use of the linking device “*at the last*”. Note that words in round brackets were error corrections added to facilitate the understanding of the script.

In my opinion, shut (shutting) down them can't solve the problem of air population, and also can cause some social problems.

At first, also is the most important point, there are hundred and thousand people working in these factories, they rely on this job to afford their family and themselves. If they lost this job, what they can do?...

Secondly, close (closing) these factories will have the effect on economic growth. With the world economic globalization, a country want to improve its statue on international...

At the last, shut (shutting) down these factories is useful for the air pollution, but there are also many (other) source of pollution, such as cars and others.

So, in my opinion, we should not shut down these factories.

Figure 5.20 Sample text at level 3, with underdeveloped yoursides arguments

This poor use of language may cause difficulty in interpreting the meaning of the arguments by different raters consistently. This may explain why outliers at level 3 were more than and more distant than those at level 4. This may also imply that raters may not be accustomed to attend to yoursides arguments as there is no mention of balanced argumentation in the current TEM4 rating scale (see Figure 2.6).

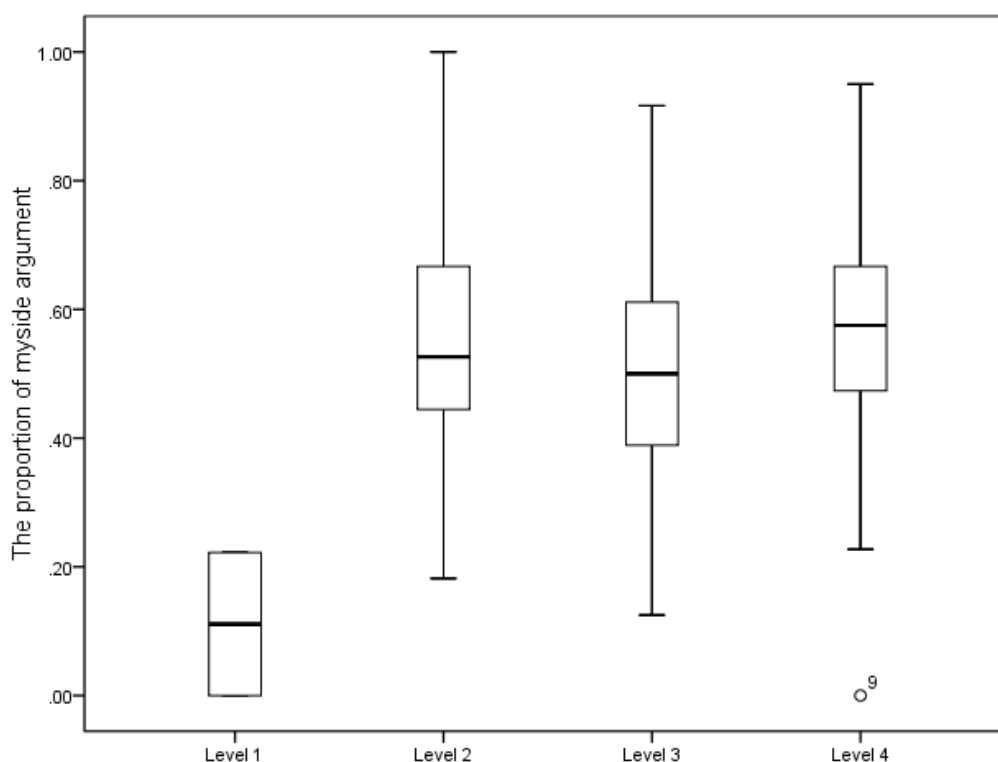


Figure 5.21 Distribution of myside arguments

Myside arguments include the author's standpoint, level-1 reasons, and level-2 reasons and below. Figure 5.21 depicts the distribution of the proportion of myside arguments. The proportion of myside arguments was not able to distinguish between the four AWA levels linearly. The variable shows a marked increase from L1 to L2, but decreased at L3 and then increased again at L4.

Table 5.26 Descriptive statistics – Proportion of myside arguments (number of myside argument elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.11	.16	.00	.22
2	.54	.20	.18	1.00
3	.50	.15	.13	.92
4	.57	.15	.00	.95



Table 5.26 shows the descriptive statistics of the proportion of myside arguments between the four different AWA levels. Writers at L4 produced the largest amount of myside arguments, comprising just over half of the text. Because the Kolmogorov-Smirnov test showed  $p = .008 < .05$ ,  $z_{skewness} = 2.40 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 16.955$ ,  $p = .001 < .05$ , with a statistically significant difference in the proportion of myside arguments between the different AWA levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 3 and 4. The effect size for myside arguments is medium, eta squared = .09. There was no particular assumption for myside arguments but the purpose was to provide detailed information of the distribution of argument structure for the development of scale descriptors.

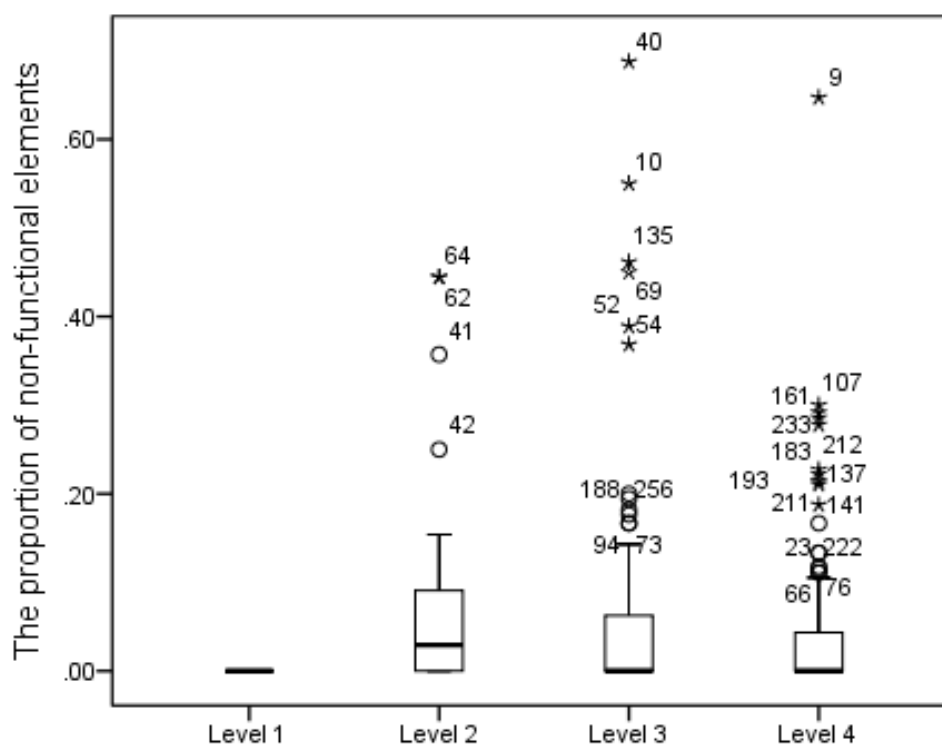


Figure 5.22 Distribution of non-functional elements

Figure 5.22 depicts the distribution of non-functional elements. None of the writers, regardless of the level, used many non-functional elements. Meanwhile, there are quite a

number of outliers. Non-functional elements include repetitions, other information that does not appear to be relevant to the topic, and illegible or nonsensical information. The assumption for this measure is that as the proportion of non-functional elements decreases the writing quality increases.

Table 5.27 Descriptive statistics – Proportion of non-functional elements (number of non-functional elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.08	.13	.00	.44
3	.06	.13	.00	.69
4	.04	.09	.00	.65

Table 5.27 shows the descriptive statistics of proportion of non-functional elements across the four different AWA levels. The proportion of non-functional elements was not able to distinguish between the four AWA levels. Because the Kolmogorov-Smirnov test showed  $p = .000 < .05$ ,  $z_{skewness} = 21.24 > 2$ , and as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 7.033$ ,  $p = .071 > .05$ , with no statistically significant difference in the proportion of non-functional elements between the different AWA levels. The effect size for non-functional elements is small,  $\eta^2 = .02$ .

A close scrutiny of writing scripts with a fair amount of non-functional elements at levels 3 and 4 (i.e., outliers) showed that these scripts, though with a higher amount of non-functional elements, were still rated high because they were good in other features, such as strong arguments, accurate language or raters did not attend to this because non-functional elements were not accounted for in the current TEM4 rating scale (see Figure 2.6). In Figure 5.23, for example, the student writer extensively describes a place that is

affected by factory pollution (see the second paragraph). After that the student writer puts

Air pollution is considered as a big environmental disaster. Our daily life is connected with air pollution ... However, compared with (the) factory, they (fire, gas, car) just a dust.

Can you image that when a factory move into a beautiful, quiet town? Dusts everywhere, we can't see a little green. Leaves fall, trees die, rivers stop, and grass turn black. Some people in town are ill or dead, parents never put wet clothes in the yard, because they will bring dust. Every morning, some men and women must go to another safe place to get some clean water.

I don't mean factories should be closed down to improve air quality. I just hope they can improve their technology-How to deal with "the rest (waste)"? Before you let the dirty water into the clean one, please think more. Why not collect together, and destroy them together with a safe, healthy way? Before you ..., why not .... Before you ..., please think more and find a better way.

Figure 5.23 Sample text with non-functional elements

forward the main standpoint "I don't mean factories should be closed down to improve air quality. I just hope they can improve their technology-How to deal with "the rest (waste)"?" The writer then makes proposals to deal with the waste in a more environment-friendly way. The extensive description of the affected town and the measures to take to reduce the pollution (or air pollution) were coded as irrelevant elements as I think they did not offer direct support to his or her standpoint. That is, they did not provide reasons for improving technology as a good way to improve air quality, but provide specific measures to improve environment. However, raters tended to award high scores to these scripts because it seems that writers holding this point of view (i.e., argue for taking measures to reduce air pollution rather than simply closing or not closing factories down) is novel, and the description of the affected town seems to appeal to the raters emotionally,

thus, these writing scripts were rated high. Note that words in round brackets were error corrections added to facilitate the understanding of the script.

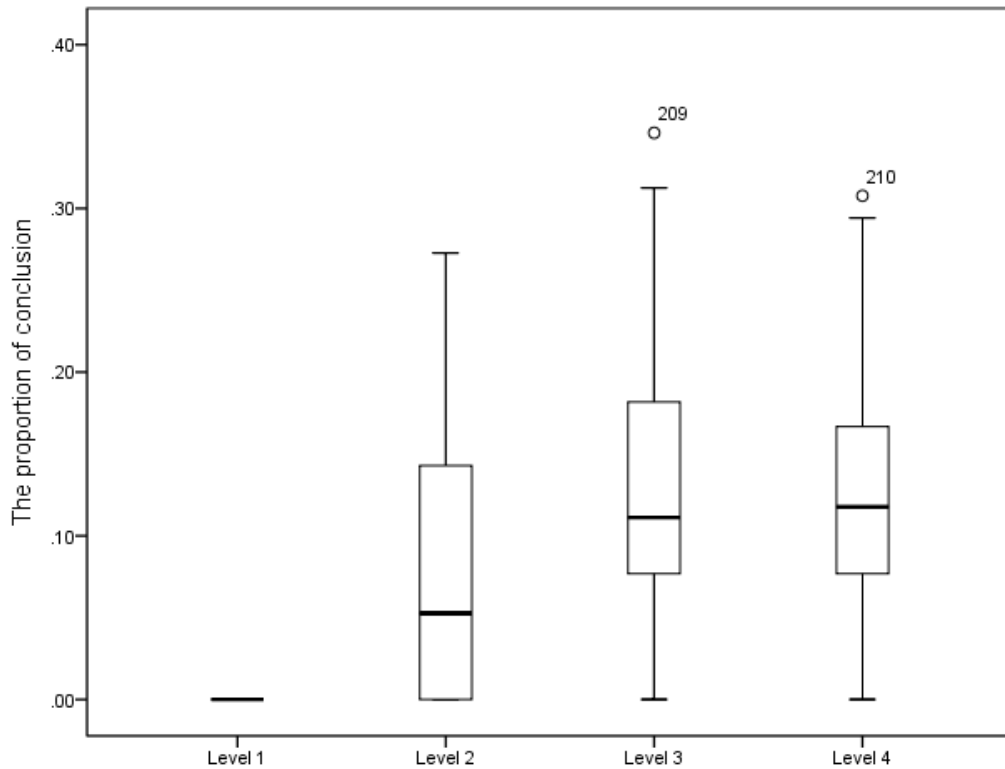


Figure 5.24 Distribution of conclusion

Table 5.28 Descriptive statistics – Proportion of conclusion (number of conclusion elements, divided by total number of argumentative structural elements)

AWA level	Mean	Std. Dev.	Minimum	Maximum
1	.00	.00	.00	.00
2	.08	.08	.00	.27
3	.13	.08	.00	.35
4	.13	.06	.00	.31

Figure 5.24 and Table 5.28 show the descriptive statistics of proportion of conclusion across the four different AWA levels. The proportion of conclusion increases as the writing level increases from L1 to L3. The variable remained the same at levels 3 and 4. Because the Kolmogorov-Smirnov test showed  $p = .001 < .05$ ,  $z_{skewness} = 2.68 > 2$ , and

as the variable was clearly not normally distributed, a Kruskal-Wallis Test was conducted. The Kruskal-Wallis test showed  $\chi^2(3) = 15.228, p = .002 < .05$ , with a statistically significant difference in the proportion of conclusion between the different AWA levels. Stepwise step down multiple comparisons showed significant differences between adjacent levels 2 and 3. The effect size for conclusion is medium,  $\eta^2 = .07$ . There was also no particular assumption for conclusion but the purpose was to provide detailed information of the distribution of argument structure for the development of scale descriptors.

It was found that only the proportion of introduction and level-2 reasons and below significantly distinguished between the different levels and showed a linear relationship with writing levels; the proportion of standpoints, conclusion, and myside argument each significantly distinguished between the different levels but did not show a linear relationship with writing levels and these former two elements were not common; the proportion of level-1 reasons, functional markers, yourside argument, and non-functional elements did not significantly distinguish between the different levels and no linear relationship was found and no salient level was found for any of these elements. Table 5.29 presents the mean of all the structural components at all four levels. As is clearly demonstrated in the table, writers at L1 were unable to produce any argumentative elements other than introduction and standpoint, while writers at level 4 displayed a more balanced rhetorical organization by allocating different proportions to different argumentative elements or components. Writers at levels 2, 3 and 4 produced little yourside arguments (< 5%) compared with myside argument (> 50%). This result was in line with previous findings (e.g., Qin and Karabacak, 2010) which showed that counterarguments and rebuttals were few in Chinese EFL writing. Although non-functional elements, which include repetitions, irrelevant information to the topic, and illegible or nonsensical information, did not distinguish between the different levels, this variable clearly decreased as the writing levels increased from 2 to 4. Functional markers

which include transitional devices to introduce reasons, arguments and standpoints did not distinguish between the different levels and were not common.

Table 5.29 Summary table – The proportion of different argument structural component

Argument structure	Level 1	Level 2	Level 3	Level 4
Positive/Negative standpoint	.11	.12	.09	.08
Level-1 reasons	.00	.25	.21	.22
Level-2 reasons and below	.00	.17	.20	.27
Non-functional elements	.00	.08	.06	.04
Functional markers	.00	.02	.04	.03
Yourside argument	.00	.04	.05	.05
Myside argument	.11	.54	.50	.57
Conclusion	.00	.09	.13	.13

### 5.6.2 Trial scale for argumentation

The design of the trial scale for argumentation was the most difficult as the difference between introduction and level-2 reasons is not obvious enough to transfer to scale descriptors. A close scrutiny of individual cases of the sample's essays was conducted. Qualitative analysis of individual cases found that poor language use was another factor in influencing the strength of argument in addition to argument structure. In Figure 5.25, for example, the writer argues that shutting down factories is not a good idea. The writer seems to defend his standpoint with two L1 reasons: the harm brought by shutting down factories to people and lack of sources for replacement of factories. He or she provides a number of L2 reasons to support the first L1 reason (e.g., *largest population, depending on them to live*) and the second L1 reason (e.g., *a giant program, (no) funds, technical workers, other requires to change the present situations*). Although L1 reasons and L2 reasons are well provided, language use errors that occur impede the exact meaning of

the reasons. For example, wrong use of “after all”, nonsensical clause “other requires to change the present situations”, inaccurate statements “China has not the capability to decide to shut down”.

Some experts firmly believe shutting down these factories can improve the country’s air. Others believe that shutting down them would curb the economic growth and leave people unemployed, which I can’t agree more. The reasons as following:

If we shut down all these factories, what will happen? China has the largest population over the world, the minority aristocracy control the majority wealth, after all, the peasants and the middle estate are depending on them to live. Closed these factories is equal to kill some people. Therefore, can we do this?

Additionally, shutting down these factories mean that China need to explore new energy to replace them, it is a giant program. Does China have funds, technical workers, and other requires to change the present situations? Undoubtedly, China has not the capability to decide to shut down. If we have a better solution, why don’t use it?

Therefore, it is not wise to shut down these factories ....

Figure 5.25 Sample text with weak reasoning due to poor language use

It was therefore decided that argument structural difference, and poor use of language should be included in the descriptors for the different levels. Based on the findings and a summary of individual cases at each level, a trial scale for argumentation structure was developed and is presented in Table 5.30.

Table 5.30 Trial scale – Argumentation

Level	Description
5	All level-1 reasons are convergent (e.g., R1, R2, R3) and supported by at

	<p>least two level-2 reasons and below or one very good level-2 reason. Level-2 reasons and below are a mixture of convergent, subordinate and coordinative reasons. Occasionally, rebuttals are added to rebut alternative standpoints or counterargument to reinforce writer's arguments.</p> <p>Reasons are generally accurately expressed and relevant to the topic. There are very few irrelevant reasons and inaccurately expressed reasons, and their existence does not influence the strength of the whole argument.</p>
4	<p>There are more attempts to support each level-1 reason adequately by at least two level-2 reasons in a mixture of convergent, subordinate and coordinate relations.</p> <p>There are less frequent occurrences of four argument structures: only unsupported convergent (R1, R2, R3) or a mixture of convergent and coordinate level-1 reasons (e.g., R1A, R1B, R2A, R2B), only one or two unsupported level-1 reasons and only one L1 reasons, which is extensively supported by level-2 reasons and below (e.g., R1, R1.R1A, R1.R1B, R1.R1.R1).</p> <p>Reasons are generally acceptable, and there are less frequent weak reasons caused by poor language use than that at L3.</p>
3	<p>There are mainly four typical argument structures: only unsupported convergent (R1, R2, R3) or a mixture of convergent and coordinate level-1 reasons (e.g., R1A, R1B, R2A, R2B), only one or two unsupported level-1 reasons and only one L1 reason, which is extensively supported by level-2 reasons and below (e.g., R1, R1.R1A, R1.R1B, R1.R1.R1). A typical structure, in which both opposing sides are expanded and the author's standpoint is provided at the concluding part, or where the author has a</p>



	<p>neutral standpoint regarding the controversial topic can be found at this level.</p> <p>There are more convincing and relevant reasons than those in L2 essays, however, there is still considerable weak reasoning and some unintelligible reasons.</p> <p>There are occasional attempts to support each level-1 reason adequately by at least two level-2 reasons in a mixture of convergent, subordinate and coordinate relations. However, language is used poorly to discount this strong argument structure.</p>
2	<p>There is a small amount of reasoning, including convergent and coordinate reasons (2-4) (e.g., R1, R1.R1, R2) and a general lack of support of level-1 reasons by level-2 reasons. In most texts, L1 reasons are unsupported and are in coordination (e.g., R1A, R1B, R2A, R2B). When level-1 reasons are supported, only one or two level-1 reasons are supported by level-2 reasons (e.g., R1, R1.R1A, R2). No reason below level-2 is found.</p> <p>Reasoning is generally weak because of the frequent wrong or poor use of words or phrases, or is poorly expanded or supported. Occasionally there is a large amount of irrelevant information.</p>
1	<p>Writer's standpoint or claim is put forward after an elaborated introduction of background. No reason is given to support the standpoint or claim.</p>

## 5.7 Conclusion

This chapter presented the results of the analysis of argumentative writing responses of 258 Chinese EFL college students majoring in English, and discussion of the results in

relation to previous literature. Based on the findings, successful measures in discriminating between different proficiency levels were selected and descriptors created. Modifications were made to the descriptors and levels based on the limitations imposed by human rating of transcripts.

## **Chapter 6    Methodology – Analysis of ratings and questionnaire data**

The following chapters provide the methodology, results and discussion chapters of the second phase – the usability of the new rating scale. As mentioned in the overall research design in Section 4.1, a mixed research design was adopted in the second phase: a quantitative methodology was adopted for the inter-rater reliability analysis, while a qualitative methodology was adopted for the analysis of questionnaire feedback.

This chapter presents the methodology of the second phase of the study, and details the research design, participants, instruments and procedures.

### **6.1    Research design**

The second phase was designed to answer the overall Research question 2: Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing? Two subsidiary questions were developed to explicate the second research question: 2a.How reliable are the ratings produced by Chinese EFL argumentative writing teachers using the new rating scale? 2b.What are raters/Chinese EFL argumentative writing teachers' perceptions of the new rating scale? The quantitative analysis of rating data was designed to answer the first subsidiary question. The qualitative analysis of questionnaire data was designed to answer the second subsidiary question. The usability study of the new scale involved several stages: recruitment of raters, selection of writing scripts, preparation of training materials, trial of new scale, training, data collection, and data analysis.

Three Chinese EFL writing teachers who are the target users of the new rating scale in classroom assessment were recruited for the study. They were recruited because they varied in their experience of teaching and rating of argumentative writing and were

expected to represent as closely as possible the target population of teacher users of the new rating scale. After the three raters were recruited, 32 scripts were selected from the CEAW corpus to represent as closely as possible the target population of student writers, especially the spread of scores. The purpose was to see if raters could consistently rate writing scripts with a wide range of proficiency levels and written by student writers from wide variety of demographics. Next, training materials were prepared. These included a training manual (see Appendix 8), three coding manuals (see Appendix 3, Appendix 4, and Appendix 5), a rating sheet (see Appendix 7), a writing task (see Appendix 11), and two writing scripts (scoring 5.5. and 7). The training manual was developed to provide detailed instructions and exercises to raters in using the rating scale. After the preparation of the training materials, three teachers received these materials through email. They were asked to read the training manual and complete the exercises, and trial-rate the two scripts using the new rating scale. The trial rating was conducted before a training session. This was because the training would take a long time as the terms in the scale descriptors were new to the teachers (Personal communication with teachers, 2019) and teachers had a tight schedule to meet for training. It was then decided to let the teachers familiarize themselves with the new rating scale and trial-rate two scripts first, and then meet for discussion and training with some knowledge and understanding of the new rating scale acquired in the trial rating. After I received the trial rating results, teachers met for discussion and training. At end of the meeting, rating materials were handed out to each teacher, including a rating sheet (see Appendix 7), 30 writing scripts, the writing task and the new rating scale. They were asked to hand in the rating results immediately through email or on paper and prepare to fill in questionnaires afterwards. The three raters returned the rating results two to three weeks later. They were given questionnaires and filled in the electronic or paper versions of questionnaires at home and returned their feedback as required. Finally, statistical analysis of the rating results and content analysis of questionnaire feedback were both conducted.

## 6.2 Participants

The participants who took part in the study of usability of rating scales included student writers and raters/writing teachers. Student writers produced the writing scripts that were rated by the raters. Raters took part in the trial of the rating scale, the training exercise, and completed a questionnaire.

### 6.2.1 Students

Thirty students' writing scripts were selected from the CEAW corpus for the investigation of the usability of the new scale. These students were not recruited separately but from those who took part in the development phase. These students were chosen to represent the CEAW corpus as closely as possible. University, major, and gender are presented in the following tables. As mentioned in main study (see Section 4.7.1), the demographic information presented here is for other researchers who might be interested. There was no analysis based on these characteristics in this study. Score distribution of the writing scripts was also important in selection of students, but is reported in Section 6.3.1.

Table 6.1 Gender distribution

Gender	Male	Female	Total
Student	4	26	30
Percentage	13.3%	86.7%	100%

Table 6.1 shows that there were many more female students (86.7%) than male students (13.3%). This is consistent with the gender distribution of the CEAW corpus, as described in Section 4.4.2.

Table 6.2 Major distribution

Major	English education	Foreign affairs management	International relationship	Public policy	Translation	Total
Student	16	0	3	4	13	30
Percentage	53.3%	0	10%	13.3%	43.3%	100%

Table 6.2 shows that the majority of the student writers were majoring in English education (53.3%) –10 percent more than the percentage of student writers majoring in English translation (43.3%). This distribution is different from the major distribution of the CEAW corpus and the main study sample. Students majoring in International relations and Public policy take up 10% and 13.3% respectively. There were no student writers majoring in Foreign affairs management because none of them wrote for Writing Task III-Air pollution.

Table 6.3Year distribution

Year	2012	2013	Total
Student	9	21	30
Percentage	30%	70%	100%

In Table 6.3, it can be seen that 70% students were enrolled in 2013, while 30% of the students were enrolled in 2012, which is roughly consistent with the sample in the main study (see Table 4.29) and the sample in the CEAW corpus (see Table 4.4).

Table 6.4 University background

University	A	B	C	D	Total
Student	2	13	3	12	30

Percentage	6.7%	43.3%	10%	40%	100%
------------	------	-------	-----	-----	------

Table 6.4 presents the distribution of students in terms of universities they were enrolled in. The majority of students were from University B (43.3%), which ranks the second highest among the four universities according to the national university ranking system (see Section 4.4.2). This is slightly more than the number of students from University D (40%), which ranks the lowest among the four universities. The number of students from Universities A and C was much lower, taking up 6.7% and 10% respectively.

### 6.2.2 Raters

I contacted three writing teachers in the university where I was working and asked them if they would take part in the validation phase through email. They showed interest in the new scale and all agreed to take part. They were selected as they varied in their teaching and rating experience and were expected to be representative of college teachers of argumentative writing who may use the new rating scale in actual classroom assessment. Two of the three teachers had taught foundation and intermediate writing courses for university English majors for more than 5 years, while the other one had taught English reading and TEM4 writing preparation classes for 10 years. All of them had experience of rating writing assignments, however, three teachers varied in the number of times they had rated high-stakes writing tests (e.g., TEM4 writing tests). Their teaching and rating experience background is presented in Table 6.5.

Table 6.5 Raters' background

Rater	Argumentative writing teaching (years)	Argumentative writing rating experience (times)
A	7 years	No TEM4 rating experience writing assignments rating: multiple times

B	10 years	TEM4: once
		Writing assignments rating: multiple times
C	10 years of TEM4 writing preparation course	TEM4 writing: four
		TEM4 writing rating: multiple times

### 6.3 Instruments

Five instruments were used in the usability study: thirty writing scripts, the new rating scale, a rating sheet, training materials, and a questionnaire. Each of these is described in detail in the following sections.

#### 6.3.1 Writing scripts

Thirty scripts were selected for the study of usability of rating scales to represent the AWA as closely as possible. All thirty writing scripts were written on Writing Task III-Air pollution (see Figure 4.4). At the time of selecting the writing scripts following a discussion with Anthony Green (personal communication, 2019), I understood that writing tasks could be a key factor in influencing the distribution of text features, thus influencing formulation of descriptors. It was then decided to use one writing task in this phase. Writing Task III-Air pollution was chosen as it has the widest spread of scores (from 2 to 12). Scripts that were used in the development phase were avoided as much as possible because it was intended that the new scale would be as useful as possible in rating different samples of writing scripts with a similar score range. Three copies were printed for each writing script.

Thirty writing scripts were first selected from the CEAW corpus in which they were stored in PDF files. The background information was then encrypted using the PDF encryption function, in order to avoid student writers from being identified and raters from being influenced by the university's ranking. The length of the writing scripts ranged from 140 to 382, with a mean of 239 words.



### **6.3.2 New rating scale**

The new rating scale consists of five trait scales: the scale of mechanics, the scale of fluency, the scale of accuracy, the scale of coherence, and the scale of argumentation structure. They can be found in Sections 5.1.2, 5.2.2, 5.3.2, 5.5.2, and 5.6.2 respectively. They are therefore not reproduced here.

### **6.3.3 Rating sheets**

Rating sheets (see Appendix 7) were developed to record both raters' background information and their ratings using five trait scales. They consisted of two sections: raters' background information, including raters' names, the length of argumentative writing teaching experience, and the length of argumentative writing rating experience, and the ratings sheet. The ratings sheet is a grid with the traits laid out as columns and the scripts as rows.

### **6.3.4 Training materials**

Since raters were busy at the time of the usability study and could not spare a large amount of time, training materials were produced and provided to three raters before training to help raters familiarize themselves with how the new rating scale was to be used. The training materials provided were a training manual (see Appendix 8), two sample texts scored as 5.5 and 7, the new rating scale, the error coding manual (see Appendix 3), the topical progression coding manual (see Appendix 4), the argumentation coding manual (see Appendix 5), and Writing Task III-Air pollution (see Figure 2.6) and the rating sheet (see Appendix 7). In the training manual, instructions were given on how each trait was to be rated. For example, in the manual, raters could understand the definitions of different argument categories and practice identifying them in a sample text. Answers were provided at a formal meeting held before actual training. The meeting was to provide a forum in which raters who had problems understanding the manual and doing the exercises could seek help and discuss their trial rating results.

### 6.3.5 Questionnaire

A questionnaire was administered after the actual rating of thirty writing scripts. The purpose of the questionnaire was to elicit raters' perceptions of the usability of the new rating scale. Raters were asked to consider the adequacy of the categories, the levels of each category and the wording of the descriptors, and their rating behavior. These questionnaire questions were adapted from Knoch (2009) and Li (2010), and can be found in Table 6.6. These questionnaire questions were reviewed by a doctoral student for wording of descriptors and for grammar. No suggestions for revision were made. All questions were written in English because I found in the formal meeting that the terms that were mentioned in the training manual, as well as those included in the questionnaire questions, were unfamiliar to the raters in both English and Chinese. Therefore, to avoid unnecessary confusion caused by translation, I used the English version of the terms.

A hard copy and an electronic copy of the questionnaire were created, so that each rater could choose the medium which was preferable to them.

Table 6.6 Questionnaire questions

---

1. What do you think of the new rating scale consisting five trait scales?
2. Does the rating scales cover all categories of argumentative writing? If not, please say what the missing categories are.
3. Does each category have the right number of levels? If not, please suggest the right number of levels.
4. Do you think the descriptors are clear? If not, please say what they are.
5. Are there any categories that you found difficult to apply? If yes, please say what they are and the reason?
6. Do you find the new rating scale time-consuming?
7. Do you use any rating scale in the classroom assessment of argumentative

---

---

writing? If yes, please say what they are.

8. Did the new rating scale influence your rating behaviour?
  9. Did you at times use a holistic (overall) score to arrive at the scores for the different categories when using the new rating scale?
  10. Do you think you will use the new rating scale in your classroom assessment?
  11. Do you think the new rating scale is useful for you to write feedback to students?
  12. Do you think the new rating scale is useful for you to better understand the evaluation of argumentative writing?
  13. Do you think the new rating scale will influence your teaching of argumentative writing? If yes, please say what the influences are.
  14. Do you have any comments on the new rating scale that are not mentioned in the previous questions? Please write specific comments that you have about each of the scale categories below. You could for example write how you used them, any problems that you encountered that you haven't mentioned above; you can draw comparisons to anything else that you want to mention.
- 

## **6.4 Procedures**

Before the three raters rated the thirty writing scripts using the new rating scale, a trial of the new rating scale and rater training were conducted. The following sections describe in detail the trial of the new rating scale, rater training, rating using the new rating scale, administration of questionnaires and the quantitative and qualitative analyses.

### **6.4.1 Trial of new rating scale**

#### *6.4.1.1 Trial procedure*

After obtaining the raters' spoken consent to take part in the study, I brought teacher participant information sheets (see Appendix 16) and teacher informed consent forms (see

Appendix 17) to the last bi-monthly college staff meeting that semester for the raters to sign. I collected the signed consent forms. After that, there was a one-month winter break for university teachers, during which there were two holidays. In order to avoid these two holidays, I sent to them training materials by email 10 days after two holidays and before the winter break ended. The purpose of the training materials was to provide ample time to let the raters to familiarize themselves with how the traits were to be rated. They were told to read the instructions in the training manual and could ask any questions concerning the training materials. After all their concerns were resolved, they were required to trial-rate two writing scripts.

At the time the training materials were sent, the teachers' consensus on a fixed date for a face-to-face formal meeting was obtained. They agreed that they would meet on the second Friday of the first month of the new semester, and send their trial rating results back to me before the meeting. The primary purpose of the formal meeting was to answer queries that might exist about the training materials and to provide further explanation if needed, for example, about the definitions of different argument categories in the coding manuals, as well as to hear any suggestions on the description and layout of the new rating scale. Another purpose of the formal meeting was to discuss their trial rating results in relation to how they applied descriptors of the new rating scale and why they gave certain scores to the sample texts. To provide answers to exercises in the training manual, I also rated writing script (ID 517).

#### *6.4.1.2 Trial feedback*

During the trialing, all three of them contacted me about the use of the training materials. Rater A asked me about the use of level points, the decimal point, and overall score as in the TEM4 rating scale (see Figure 2.6). I confirmed the use of level numbers as level points, no use of decimal point and overall score. Rater A also requested me to share my ratings and explanations for the ratings. I explained the concern that their ratings might

be influenced by my judgement and interpretation and told him/her that we would discuss the ratings in the formal meeting. Rater B asked me if there were other features of mechanics. I explained that the number of paragraphs is the only feature. Rater C conveyed that she was unsure of the understanding of sequential progression and asked if she needed to label the error type in the writing script. I replied that there was no need to label error type and explained that any misunderstanding would be discussed and resolved in the formal meeting.

#### 6.4.1.3 Trial rating outcome

Two writing scripts: ID 184 (score 7) and ID 517 (score 5.5) were trial-rated. Ratings were sent back to me immediately after they were completed. They are presented in Table 6.7 and Table 6.8 respectively. Table 6.7 shows the ratings of the writing script of ID 184 in terms of mechanics, fluency, accuracy, coherence, and argumentation. In Table 6.7, it can be seen that the ratings are consistent across three raters, with one point difference for fluency, accuracy and coherence, while there is no point difference for mechanics and argumentation. According to White (1984, cited in Weigle, 2002), the smaller the proportion of ratings more than one point apart, the ‘better’ the rating. Therefore, their ratings of ID 184 can be regarded as consistent as there are no ratings more than one point apart.

Table 6.7 Ratings results of a good writing script (ID 184)

Rater	Mechanics	Fluency	Accuracy	Coherence	Argumentation
A	4	3	4	4	3
B	4	4	3	4	3
C	4	3	4	3	3

In Table 6.8, it can be seen that the ratings are not consistent except those for fluency. For mechanics and accuracy, the point difference is one, while for coherence and

argumentation, the difference is as large as 2 points and there is no exact agreement of scores for argumentation. I decided to ask each of the raters to give the reason for their ratings for coherence and argumentation in the formal meeting. This was to discover the reason for the differences, for example, a difference in the way they applied a particular scale descriptor. For example, in rating ID 517 using the scale of accuracy, I found that if I estimated, I gave the rating 3; if I counted the error-free sentences, the rating was 2.

Table 6.8 Rating results of a poor writing script (ID517)

Rater	Mechanics	Fluency	Accuracy	Coherence	Argumentation
A	3	2	3	4	4
B	4	2	2	2	2
C	3	2	3	2	3

#### 6.4.1.4 Formal meeting outcome

In the formal meeting, the questions asked during the trial rating were asked again. Their responses were noted down by me and they were briefly summarized as: “mechanics” was too broad as it only addressed the number of paragraphs; the use of ‘we’, ‘you’ should or should not be penalized; rating was time consuming. Since the use of ‘mechanics’ as the name of the scale would not influence their rating results, it was decided that ‘mechanics’ would remain as the scale name and their critiques would be reported in more details in questionnaire feedback for the sake of readability. Although one rater disagreed on the penalization of students for the use of ‘we’ and ‘you’, I suggested he or she follow the scale of coherence. The reason is discussed in Section 7.3.1.4, in which their trial rating response to the scale of coherence is also discussed in more details.

The formal meeting also focused on the reasons for differences in scores for rating coherence and argumentation in sample text ID 517 (scored as 5.5). I worked as a

coordinator and an experienced rater who knew the new rating scale. They first presented their reasons for each rating for coherence and argumentation and discussed their disagreements. I provided my ratings of sample text ID 517 using five trait scales and demonstrated how the sample text was analyzed using the coding manuals of coherence and argumentation and then explained their ratings. They were convinced by the demonstration and explanation.

#### **6.4.2 Rater training**

The training was held immediately after the formal meeting. Another writing script, scored as 6, was assigned to each of the three raters. The procedure employed in the trial rating was repeated. The three raters and I rated the scripts separately. Then their ratings were discussed and compared with mine. Any differences caused by misuse of the scales were resolved, but any differences caused by different yet reasonable interpretations of scale descriptors were retained.

#### **6.4.3 Data collection**

##### *6.4.3.1 Ratings of scripts using the new scale*

The thirty student writing scripts, with writers' information encrypted, were stored in PDF files in a separate folder. They were printed for each of the three raters and given a random ID number from one to thirty. A rating pack was given to each rater that comprised the thirty writing scripts, a copy of the new rating scale, a copy of the training manual, a copy of Writing task III, a copy of the rating sheet, a copy of the training manual, and a copy of the coding manuals for error taxonomy, coherence, and argumentation.

Since the three raters had tight teaching schedules, it was not possible for them to attend another rating session. Therefore, the raters rated the thirty scripts separately at home. In order to avoid fatigue, raters were asked to take a break when they felt it necessary. They

were also asked to return the rating sheet to me immediately after they finished their ratings. All the three raters returned the ratings as they promised. Raters A and B agreed to return their ratings in two weeks and rater C in three weeks – rater C submitted one week later because she had an extra supervision workload on the thesis writing of graduate students. All ratings were entered into an Excel spreadsheet. Payment for their time was given to each rater when they handed their rating sheets back to the researcher.

#### *6.4.3.2 Administration of questionnaires*

Each of the three raters were given a paper version of the questionnaire immediately after they handed in the rating sheets. They were asked to fill in the questionnaire as soon as possible in case they forgot their rating process and their opinions on the new rating scale. Raters were asked to retain other rating materials in the rating pack to help them remember details of their rating. An electronic version was sent to rater B through email as rater B preferred to respond to an electronic version. All raters completed and returned the questionnaires within two days of receiving the questionnaire. Two raters returned the paper versions and one rater returned the electronic version. There was no specification on the language to use but all responded in Chinese. Their feedback was then translated and stored in separate files entitled by each rater's name.

### **6.4.4 Data analysis**

#### *6.4.4.1 Inter-rater reliability analysis*

Cronbach's coefficient alpha was calculated to investigate how reliable the ratings were when using the new rating scale (the first subsidiary research question). Cronbach's coefficient alpha was developed to provide a measure of the internal consistency of a test or scale (Cronbach, 1951). The internal consistency or reliability of a test is interpreted as the extent to which the scores produced by the test are consistent or stable (e.g., Larsen-Hall, 2010). Despite its initial use in measuring internal consistency of a test or scale, it



has been widely used to estimate the reliability or consistency of scores produced in test-retest, parallel test, and interrater conditions (Larsen-Hall, 2010; Zou, 2017). Following the researchers mentioned above, the rater reliability to be investigated in the current study is interpreted as the extent to which the scores produced by different raters using the new rating scale are consistent, or to put it another way, the consistency of raters in ranking the same set of subjects (i.e., test takers). The higher the coefficient alpha, the more consistent the scores are.

There are a number of assumptions to meet if the coefficient alpha is to be accurately interpreted, and these assumptions are met in the current study. First, Cronbach's alpha is applicable to interval data or questionnaire data "where there is an implied interval scale" (Green, 2013, p. 30). Raw scores for education or psychological measurements are often considered as interval data (Zou, 2017), therefore, the ratings using the new rating scale comprising different trait scales can be roughly regarded as interval data as these ratings of writing performance measure language development or ability (i.e., educational or psychological measurement). Second, the items (on a test, or for our purpose, raters) should be homogeneous or unidimensional (i.e., measure one construct), otherwise the coefficient alpha is underestimated (e.g., Carr, 2011). Researchers suggest using factor analysis to identify dimensions of a test or questionnaire, or breaking the test into parts and measuring a different construct or concept with each part (e.g., Carr, 2011). In this study, the focus is on the inter-rater reliability using five different subscales, with each scale measuring one construct (e.g., fluency, accuracy). Therefore this assumption is met. Third, the items (on a test, or for our purpose, raters) should be independent (i.e., performance on one item is not related to performance on another). In the current study, raters marked the writing scripts separately and independently, thus this assumption is met. Fourth, there is some variation in ability levels of the test takers, otherwise, alpha will be underestimated with only high-ability or only low-ability test takers (Carr, 2011).

This assumption is also met as the writing scripts represent a wide spread of language ability indicated by scores ranging from two to twelve.

The interpretation of Cronbach's alpha should be cautious as its size is affected by the number of items (raters in my case), the number of dimensions in the data, and variability among sample subjects (test takers in my case) (Cortina, 1993, cited in Larsen-Hall, 2011; Shrout and Fleiss, 1979; Green, 2013, p.39). Koo and Li (2016) suggest that for the conditions of 30 heterogeneous samples (variability among test takers in my case) rated by at least 3 raters, intraclass correlation coefficient values (ICC) less than 0.5 (Cronbach's alpha is a measurement of intraclass correlation) are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.

Cronbach's Alpha If Item Deleted (CAID) refers to the internal consistency or reliability of a test or scale with one item deleted (Green, 2013). CAID was calculated to locate any rater who might have trouble in using the scale consistently compared with other raters. The figure was calculated in order to find out whether the item deleted contributes positively or negatively to the overall internal consistency of the test or scale. If the figure for the deleted item is lower than the overall alpha, it means that that item contributes something positive to the overall interval reliability. If the figure for the deleted item is higher than the overall alpha, it means that that item contributes something negative to the overall interval reliability. In that case, the item should be deleted. Green (2013, p. 38) suggests that a negative item alpha (i.e., CAID) would indicate wrong answers being keyed in during data entry (e.g., "a negatively worded questionnaire which has not been reversed") or a flawed item. Larsen-Hall (2011) suggests that if the CAID for a rater differs drastically from the overall alpha then that rater's scores can be discounted. Interpreted in the current study, a negative CAID and a drastic CAID in contrast with the coefficient alpha for all raters would indicate wrong keyed scores or a rater being unable

to use the rating scale consistently. Similar CAIDs for each rater would indicate raters being consistent in using the rating scale.

#### *6.4.4.2 Analysis of questionnaires*

Questionnaires were saved as word files and then coded manually by the researcher. Two broad themes were devised a priori based on the questions: positive and negative feedback. Under each broad theme, sub-themes relating to the five trait were devised: mechanics, fluency, accuracy, coherence and argumentation. Then data were grouped according to these themes and subthemes. Since the data were short and themes were straightforward, no double coding was involved.

## Chapter 7 Results and discussion – Analysis of ratings and questionnaire data

In this chapter, quantitative, and qualitative findings are summarized and presented in Sections 7.1 and 7.2 respectively. After that, a discussion of the usability of the new rating scale is presented.

### 7.1 Results of inter-rater reliability

Table 7.1 shows that the mean scores for the 30 writing scripts given by the three raters using the scale of mechanics were: rater A ( $M=2.60$ ,  $SD=.86$ ); rater B ( $M=2.63$ ,  $SD=.81$ ); rater C ( $M=2.70$ ,  $SD=.84$ ). The minimum and maximum scores given by the three raters were the same: from 1 to 4.

Table 7.1 Descriptive statistics – Scale of Mechanics

Rater	Mean	Std. Dev.	Min	Max
A	2.60	.86	1	4
B	2.63	.81	1	4
C	2.70	.84	1	4

The coefficient alpha for the three raters using the scale of mechanics was .97. Koo and Li (2016) suggest ICC values (Cronbach alpha) less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability. Accordingly, the ratings for the scale of mechanics is seen as highly reliable. Since this scale is based on the length of the scripts and is mechanical, it was expected that the three raters using this scale would be consistent. Cronbach's Alpha If Item Deleted (CAIDs) for raters A, B, C were .95, .94, and .99 respectively. Raters A and B seemed consistent in using the scale of mechanics as the CAIDs were similar, while the CAID for rater C

was higher than the overall alpha ( $=.97$ ), which means that rater C contributed negatively to the overall reliability of their ratings.

Table 7.2 Descriptive statistics – Scale of Fluency

Rater	Mean	Std. Dev.	Min	Max
A	3.03	.62	2	4
B	2.63	.77	1	4
C	2.83	.60	2	4

Table 7.2 shows that the mean score for 30 writing scripts given by rater A ( $M=3.03$ ,  $SD=.62$ ) using the scale of fluency was higher than those given by raters B ( $M=2.63$ ,  $SD=.77$ ) and C ( $M=2.83$ ,  $SD=.60$ ). The minimum and maximum scores given by raters A and C were both from 2 to 4, while the score range for rater B was from 1 to 4.

The coefficient alpha for the three raters for the scale of mechanics was .88. Following Koo and Li (2016), the ratings for the scale of fluency are reasonably reliable. Since this scale is based on the length of the scripts, the three raters using this scale were consistent. Cronbach's Alpha If Item Deleted (CAIDs) for raters A, B, C were .81, .87, and .80 respectively. Raters A, B and C were consistent in using the scale of Fluency as the CAIDs were similar. All three raters contributed positively to the overall reliability of their ratings as the CAIDs for each rater was lower than the overall alpha ( $=.88$ ).

Table 7.3 Descriptive statistics – Scale of Accuracy

Rater	Mean	Std. Dev.	Min	Max
A	3.40	.56	2	4
B	3.87	.51	3	5
C	3.53	.63	2	4

Table 7.3 shows that the mean score for the 30 writing scripts given by rater B ( $M=3.87$ ,  $SD=.51$ ) using the scale of accuracy was higher than those given by raters A ( $M=3.40$ ,  $SD=.56$ ) and C ( $M=3.53$ ,  $SD=.63$ ). The minimum and maximum scores given by raters A and C were both from 2 to 4, while the score range for rater B was from 3 to 5.

The coefficient alpha for the three raters using the scale of accuracy was .29. In accordance with Koo and Li (2016), the ratings for the scale of accuracy were not reliable. It is surprising to see that the accuracy scale did not produce reliable ratings as this scale is based on the frequency of errors occurring in each of five TEM4 levels (i.e., is data-based). However, the low reliability could be attributed either to raters not being able to interpret estimates of quantifiers (e.g., *nearly all*, *two thirds*, *half*, *one third*, *almost no*) consistently, or to raters not being used to assessing the features (e.g., spelling, punctuation use) in the new rating scale that are normally addressed in the category of mechanics in the existing rating scales (e.g., The rating scale of the TEM4 writing section (Li, 2010)), Jacobs et al. (1981)'s ESL Composition Profile: content (Jacobs et al., 1981)). This was brought up by raters in the training exercise and, although in the training exercise I emphasized that all errors including those mentioned above were accounted for in the scale of accuracy, they did not follow the new rating scale consistently. However, this could only be established by investigating the raters' rating decision behavior (i.e., how they were marking based on their interpretation of these estimates). Cronbach's Alpha If Item Deleted (CAIDs) for raters A, B, C were .21, .27, and .13 respectively. All three raters contributed positively to the overall reliability of their ratings as the CAID for each rater was lower than the overall alpha ( $=.29$ ).

Table 7.4 Descriptive statistics – Scale of Coherence

Rater	Mean	Std. Dev.	Min	Max
A	2.70	.86	2	4
B	3.00	.81	1	5

C	2.60	.84	2	3
---	------	-----	---	---

Table 7.4 shows that the mean score for the 30 writing scripts given by rater B ( $M=3.00$ ,  $SD=.81$ ) using the scale of coherence was higher than those given by raters A ( $M=2.70$ ,  $SD=.86$ ) and C ( $M=2.60$ ,  $SD=.84$ ). The range of scores given by rater C (2) was narrower than those given by raters A (3) and B (5), which shows that rater C tended to use middle levels while raters A and B tended to use all five or three levels. Since the 30 writing scripts are a representative sample of writing scripts scoring from very low (i.e., 1.5) to very high (i.e., 12) based on the existing 15-point TEM4 rating scale, it was expected that the range explored by rater C would have been at least greater than 2. This could either point to rater C being unable to use the scale or the scale not being properly developed, causing confusion for rater C.

The coefficient alpha for the three raters using the scale of coherence was .31. Following Koo and Li (2016), the ratings for the scale of coherence were not reliable. Cronbach's Alpha If Item Deleted (CAIDs) for raters A, B, C were -.17, .47, and .37 respectively. According to Green (2013, p. 38), if a negative CAID for one item was found, it could mean that the answers for that item (i.e., the scores by that rater in my study) were wrongly keyed into SPSS. However, double checking the data showed that rater A tended to give lower scores for good performance and higher scores for poor performance. This could be due to rater A's inability to use the scale (e.g., failing to identify unrelated sequential progression or discourse-related sequential progression, failing to follow the criteria in rating) or the rating scale itself failing to provide reliable rating criteria to rater A. However, determining whether it was the rater's behavior or a problem with the rating scale can only be done by further interviewing rater A.

Table 7.5 Descriptive statistics – Scale of Argumentation

Raters	Mean	Std. Dev.	Min	Max
--------	------	-----------	-----	-----

A	2.80	.66	2	4
B	2.97	1.03	1	5
C	2.57	.57	2	4

Table 7.5 shows that the mean score for the 30 writing scripts given by the three raters using the scale of argumentation were: rater A (M=2.80, SD=.66), rater B (M=2.97, SD=1.03) and rater C (M=2.57, SD=.57). The range of scores given by rater B (5) was wider than those given by raters A (3) and C (3), which shows that rater B used a relatively wider range (all five levels) than raters A and C did (three levels).

The coefficient alpha for the three raters using the scale of argument was .64. Following Koo and Li (2016), the ratings for the scale of argumentation were moderately reliable. The argumentation scale was more reliable than the scale of accuracy and the scale of coherence, while it was less reliable than the scale of mechanics and the scale of fluency. It is not surprising to see that it performed worse than the scale of mechanics and the scale of fluency, since the scale of argumentation contains more complicated descriptors which entails two factors: both conveyance of the arguments in language (e.g., clearness ) and also different argumentative structures (e.g., the frequency of occurrences of different structures). Cronbach's Alpha If Item Deleted (CAIDs) for raters A, B, C was .69, .47, and .42 respectively. The CAID for rater A was lower than the overall alpha (i.e., .64), which means that that rater contributed negatively to the reliability of the three raters' ratings. Following Larsen-Hall (2011), rater A's scores should be discounted or taken cautiously in the actual test rating. In the current study, it can be both interpreted as the rating scale needing to be improved or revised and as rater A's inconsistency of rating. Qualitative analysis of rater A's feedback and the rating process could provide more evidence for these interpretations.



## 7.2 Raters' perceptions of the new rating scale

Since only three raters were involved and their answers to questionnaire questions were short, themes were not easy to find. Thus, the raters' perceptions of the new scale will be reported around each questionnaire question, while in each question their answers will be reported following the broad themes (i.e., positive and negative) and subthemes (i.e., mechanics, fluency, accuracy, coherence and argumentation) where the themes emerged.

Question 1: *What do you think of the new rating scale consisting five trait scales?*

The first question asked for the raters' general views about the new rating scale. All three raters responded positively and thought the scale descriptors were relatively "concrete" and "easy to use". Rater A found the scales of coherence and argumentation especially useful because "the coding manuals provided more concrete analysis for coherence and argumentation". Rater B found the scale of mechanics, fluency, and accuracy especially easy to use. Rater C thought the rating scales "were relatively comprehensive and consider different aspects of the writing scripts."

Question 2: *Does the rating scales cover all categories of argumentative writing? If not, please say what the missing categories are.*

The second question asked raters if the rating scales covered all categories of argumentative writing. Two raters found there were missing categories. Rater A found two categories were missing: the quality of reasons in terms of whether the reasons or examples were "typical or not typical" "in-depth or one-sided", and the appropriateness of language, for example "the distinction between spoken English and written English". Rater C found that word-for-word English translations of Chinese formulaic expressions were missing and should be accounted for in the scale of accuracy, although this inaccurate language use was acknowledged in coding manuals of error taxonomy. Rater B seemed to be not sure about her response by writing "Maybe yes" because "The practicality would be reduced if there were too many categories."

Question 3: *Does each category have the right number of levels? If not, please suggest the right number of levels.*

The third question asked raters if they found the number of levels for each category enough to distinguish the argumentative writing scripts they rated. Raters B and C found the number of levels for each category were appropriate, while rater A suggested that there should be five levels for coherence, which would make the number of levels for each trait scale consistent.

Question 4: *Do you think the descriptors are clear? If not, please say what they are.*

The fourth question asked raters if they found the descriptors were clear. All three raters found the descriptors were clear, but rater A thought it would be better if “different levels of reasons such as R1. R1a”, “the logic of reasons such as convergent, coordination” could be simplified. Rater B suggested that it would be better if a Chinese version of the rating scales could be provided as raters “might not be familiar with terminologies such as convergent, sequential progression”. Rater C found the descriptors clear.

Question 5: *Are there any categories that you found difficult to apply? If yes, please say what they are and the reason?*

The fifth question asked raters if there were any categories that they found difficult to apply. Raters responded consistently to this question. All raters found the scale of argumentation difficult to apply. Rater A found that “identifying the logic between leveled reasons (e.g., L1 reasons, L2 reasons) was time-consuming”, and “the identification of reasons and levels of reasons varies from person to person”. Rater B found that “it was hard to identify the meaning of reasons because of language problem, thus being difficult to identify the levels and relations between reasons (e.g., convergent relation)”. Rater C also found it difficult by saying “it was not easy to figure out the levels of reasoning”.

Question 6: *Do you find the new rating scale time-consuming?*

The sixth question asked raters if they found the new rating scales time-consuming. All raters found the new rating scale time-consuming, but they differed in the degree of time consumed. Rater A found it a bit time-consuming. Rater B thought that it was time-consuming because it “involves many details” and “the descriptors invite the analysis of texts that were not familiar to raters”. Rater C thought the new rating scale “very time-consuming” based on comparisons with other rating systems. She wrote that it took “30 seconds to rate each writing script by machine scoring in large-scale tests”, while it took “one minute or so to rate each writing assignment”. In the current rating, it took “two to three minutes” to rate each one.

Question 7: *Do you use any rating scale in the classroom assessment of argumentative writing? If yes, please say what they are.*

The seventh question asked raters if they used any rating scale in the classroom assessment of argumentative writing. All raters mentioned the TEM4 rating scale of writing, but they varied in their actual use of the rating scale. Rater A did not answer as to how she used the scale. Rater B tended to give a holistic score based on the categories/aspects of the TEM4 rating scale. Rater C adopted a more analytic way of assessing by attending to “word counts”, “avoiding use of first and second person pronoun”, and “labelling inaccurate language problems”.

Question 8: *Did the new rating scale influence you rating behavior?*

The eighth question asked raters if the new rating scale influenced their rating behaviors. All the raters answered that the new rating scale influenced their behavior, but they differed in the way that their rating behaviors were influenced. Rater A started to attend to the aspect of coherence in rating which they had not done before. Rater B used to adopt “holistic rating” before, while she attended to “different aspects using the new scale”. Rater C used to focus more on “language expressions” than on other features, but using

the new rating scale, she had a balanced focus.

Question 9: *Did you at times use a holistic (overall) score to arrive at the scores for the different categories when using the new rating scale?*

The ninth question asked raters if they used a holistic score to arrive at the scores for the different categories when using the new rating scale. Raters A and B answered no. Rater C answered that she used “global scoring” when she found it difficult to identify levels and relations between reasons using the scale of argumentation.

Question 10: *Do you think you will use the new rating scale in your classroom assessment?*

The tenth question asked raters if they would use the new rating scale in their classroom assessment. Raters A, B, and C responded positively. Rater A answered that she would use the scale of accuracy, the scale of coherence, and the scale of argumentation. Rater C responded that she found that with the new rating scale she felt “more confident” in identifying problems in students’ writing and that it was “easier” to guide students to write more logically. Rater B answered that she would use the scale of argumentation.

Question 11: *Do you think the new rating scale is useful for you to write feedback to students?*

The eleventh question asked raters if the new rating scale was useful for them to write feedback to students. All three raters responded positively. Rater A did not further specify her response. Rater B responded that she would use it when she “demonstrates to students how writing script is scored”. Rater C responded that she had already used it when writing feedback to students in her TEM4 preparation class.

Question 12: *Do you think the new rating scale is useful for you to better understand the evaluation of argumentative writing?*

The twelfth question asked raters if the new rating scale was useful in facilitating their

understanding of the evaluation of argumentative writing. All three raters responded positively. Raters A and B found it provided “more details” in evaluating students’ writing. Rater C found that the evaluation of argumentative writing used to be “casual and disorganized” and marking criteria used to be relatively “general”. Recently, she had often used the scale of argumentation, identifying “big and empty reasons”, “paraphrases” of reasons, and “examples”.

Question 13: *Do you think the new rating scale will influence your teaching of argumentative writing? If yes, please say what the influences are.*

The thirteenth question asked raters if the new rating scale would influence their teaching of argumentative writing. All three raters said that the scale of argumentation would influence their teaching of argumentative writing. However, raters A and B did not specify the influences. Rater C responded that the influence was “huge”, and that she paid more attention to “the logic of writing” in her teaching in the TEM4 preparation course, for example, “demonstrating how reasons are related”, and she used the scale of coherence by guiding students to “repeat or relate to previous mentioned topics, or examples”.

Question 14: *Do you have any comments on the new rating scale that are not mentioned in the previous questions? Please write specific comments that you have about each of the scale categories below. You could for example write how you used them, any problems that you encountered that you haven’t mentioned above; you can draw comparisons to anything else that you want to mention.*

The fourteenth question asked raters if they had any comments on the new rating scale that were not mentioned in the previous questions. Most of their answers are summarized here, including those which have been mentioned previously in other raters’ responses to other questions, but not answers which have been mentioned previously in their responses to other questions, as raters occasionally repeated their responses to other questions in responses to this question. Their answers are summarized below in terms of the different

scales.

#### *The scale of mechanics*

All three raters found the term “mechanics” problematic. Rater A found that “the scale of mechanics was too simplified”, “there are overlaps between mechanics and accuracy in terms of, for example, spelling, punctuations”. Rater B responded that “it would be better to use paragraphing to replace mechanics”. Rater C found that the scale was “mechanical”, and did not take into account “the size of each paragraph”, and “More paragraphs does not necessarily mean a good writing. Three-paragraph writing should be scored 3 to 4 if the overall word account reaches 250.”

#### *The scale of fluency*

Rater A commented that, if properly used, conjunctions could be added to the scale of fluency. The other two raters did not comment on the scale of fluency for this question.

#### *The scale of accuracy*

Rater A found the error taxonomy with which the scale of accuracy was used to identify errors “very comprehensive” and “facilitate labeling errors”. Rater B commented that she recently would “try online scoring system for accuracy of language use”. Rater C commented that she found scoring accuracy is “more quantitatively justified”. For example, she found “It is quite common to see one third of sentences containing errors in TEM4 writing scripts.” She said so because she was teaching a TEM4 writing preparation course.

#### *The scale of coherence*

Rater C further commented that she found theoretical explanations for overusing ‘we’ and ‘you’ through topical progression analysis and could provide more convincing feedback on coherence. Other two raters did not comment on the scale of coherence for this

question.

In conclusion, raters generally responded positively to all five trait scales, although their responses to specific scales varied. This was possibly because of their background and rating behaviors.

### **7.3 Discussion**

This section mainly discusses how quantitative and qualitative findings together answer the second overall research question:

Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing?

The discussion is conducted based on the test usefulness framework (Bachman and Palmer, 1996). The framework of test usefulness offers a structure in which the findings and results are discussed. Bachman and Palmer's (1996) framework of test usefulness was selected as it is easier to operationalize by those who are responsible for evaluation of the validity, or usefulness, of a test than other validity frameworks (see Chapelle, 2012, p. 25 for a comprehensive view of the history of validity conceptualization). The discussion consists of five sections and, in each section, the definition of a facet of the test usefulness framework is provided, followed by a discussion on how it relates to the evaluation of validity or quality of rating scales (e.g., Weigle, 2002; Knoch, 2009), then the findings from Chapter 6 are discussed in relation to the evaluation of validity or usability of the new rating scale. At end of the discussion, the extent to which the new rating scale is usable is given.

#### **7.3.1 Construct validity**

Bachman and Palmer (1996) define construct validity as “the meaningfulness and appropriateness” of the inferences about test takers’ ability made on the basis of test scores. Weigle (2002, p.121) puts forward a number of statements on how a rating scale can be

discussed in relation to the validity of a test, in comparing holistic scales and analytic scales on the basis of the six facets of test usefulness proposed by Bachman and Palmer (1996). She views construct validity of a rating scale as whether, or to what extent, different aspects of writing ability of test takers and different rates of language development are captured by the rating scale. Knoch (2009) conceptualizes the validity of a rating scale through a number of warrants that she established to support a validity argument for the DELNA rating scale (a formal process proposed by Bachman (2005) in building an assessment use argument through a chain of warrants, claims, backings and rebuttals). The warrants she established represent an ideal situation that a rating scale should achieve in relation to each facet of Bachman and Palmer's (1996) usefulness framework. Three warrants are proposed as being relevant to construct validity of a rating scale:

Warrant 1: The scale provides the intended assessment outcome appropriate to purpose and context, and the raters perceive the scale as representing the construct adequately.

Warrant 2: The trait scales successfully discriminate between test takers and raters report that scale is functioning adequately.

Warrant 3: The rating scale descriptors reflect current applied linguistics theory as well as research.

(Knoch, 2009, p.65)

It can be inferred from Knoch's (2009) warrants that construct validity of a rating scale concerns the appropriateness of assessment outcome in relation to test purpose, adequacy of the construct's representation in rating scales in raters' views, its discrimination, and reflection of theories and research. Knoch's (2009) conceptualization of construct validity of a rating scale is more operational. Therefore, in the following, my findings will be discussed in relation to the aspects of construct validity conceptualized by Knoch (2009).

The purpose of the classroom assessment of argumentative writing is to identify strengths



and weaknesses in college students' argumentative writing and provide a detailed feedback. The new rating scale provides five trait scales. Each scale include concrete and objective descriptors (e.g., *Nearly all sentences contain one or more errors* at level 1), and each scale involves a detailed analysis of writing scripts. It can be expected that the clear scale descriptors and detailed analysis of writing scripts can provide a detailed feedback on strengths and weaknesses of students' performance than the current TEM4 rating scale with vague descriptors, if the raters can rate analytically and follow the scale descriptors. All raters responded that they rated analytically and rated five times following one scale at each time, although one rater responded that she/he rated holistically when having difficulty using the scale of argumentation. It is unclear whether the detailed feedback would be understood by students as students were not investigated in the current study. However, it would be possible that students could understand the feedback if terms (e.g., convergent arguments, discourse-related sequential progression) could be defined. Therefore, it could be argued that the new rating scale could provide assessment outcome appropriate to test purpose.

In the current study, the new rating scale was viewed by raters as relatively adequate in representing different aspects of the construct of Chinese EFL college students' argumentative writing ability and the range of writing abilities of the target population (i.e., the number of levels of the rating scale). Two raters responded that the new scale as a whole was adequate in terms of construct representation. One rater found that quality of reasoning and appropriateness of language use were missing. However, these two aspects were evaluated in the development phase. Appropriateness of language use was removed as no successful measure was found. The quality of reasoning was not included as no objective measures were available at the time of the study. It is surprising to find that the complexity of language use was not mentioned by raters although it is an important aspect of construct of language ability (e.g., Skehan, 1998). Two raters responded that the number of levels was adequate to represent the range of proficiency

levels of the target population. One rater suggested using five levels for the scale of coherence, which makes the scale of coherence consistent with other four scales. It seemed that the rater's suggestion did not have much empirical evidence.

The discrimination of a rating scale is measured by candidate separation ratio (i.e., a measure of the spread of candidates' performance relative to their measurement precision acquired by a multi-facet Rasch analysis, Linacre, 1989) and the functioning of a rating scale raters' perceptions (e.g., Knoch, 2009; Li, 2010). A higher candidate separation ratio is indicative of more power a rating scale has in being able to discriminate between more levels of candidate ability, thus a more 'superior' or more 'valid' rating scale. Two factors contribute to a higher candidate separation ratio: raters rate more similarly to each other and they use more levels on a rating scale. Since the candidate separation ratio could not be obtained on a small sample size (i.e., the multi-facet Rasch analysis requires a large sample), the number of levels raters use was used to indicate the discrimination power of the new rating scale. The underlying rationale is that the rating scale is intended to account for a wide variety of levels of ability of target population. If a restricted number of levels are used, it means that the rating scale fails to fulfill the purpose of discriminating between candidate ability levels. However, this measure needs to be interpreted with caution as rating difference between raters was not taken into account. Knoch (2009, p.256) stated that ratings with large rater difference could "cancel each other out", which reduced the candidate separation ratio, even if most levels on a rating scale were used.

In the current study, the candidate separation ratio is replaced with the range of scores used by raters. In the following sections, each trait scale will be evaluated in terms of the range of scores used by raters using each trait scale and how raters perceived each trait scale. Based on the evaluation, recommendations for further revisions of the new rating scale will be made.

### 7.3.1.1 *Mechanics*

Raters rated similarly using the scale of mechanics in terms of the number of levels employed in the rating, as is evidenced by the range of scores used by the three raters (i.e., L1 to L4 by three raters) and by most of the levels being used (e.g., four out of five levels) (see Table 7.1).

Raters responded positively to this scale as it is clear and precise. However, all raters remarked that the name of the scale should be changed to the scale of paragraphing, as ‘mechanics’ implies other aspects such as spelling and punctuation use. This suggestion reflects how current raters still hold the traditional view of mechanics as comprising spelling, punctuation use, etc. (e.g., see Grabe and Kaplan’s (1996) taxonomy of writing). This is in contrast with current language assessment, for example IELTS, in which these features of mechanics are either not accounted (e.g., punctuation use) or are included in accuracy (e.g., spelling). One rater found that the scale was mechanical which reflects similar criticisms in North (2003) who views a rating scale that relies on counting of features as mechanical. This rater also commented that those who wrote a lesser number of paragraphs and used large paragraph sizes were penalized. Overall, it could be said that the scale of mechanics was perceived well by the rater. It could be adopted in any further use of the rating scale, however, cautions need to be made as students might trick the number of paragraphs in their writing to score high. It might be helpful to change the name of scale as ‘the scale of paragraphing’ as ‘mechanics’ implied other aspects.

### 7.3.1.2 *Fluency*

Raters rated relatively similarly using the scale of fluency in terms of the number of levels employed in the rating, as is evidenced by the range of scores used by the three raters (i.e., L2 to L4 by two raters, L1 to L4 by one rater) and by most of the levels being used (e.g., three or four out of five levels) (see Table 7.2). Since this scale involves counting the

number of words, it was possible that raters did not actually count word-by-word, but rather estimated the number. When the length of the writing script was around 100 words, raters could produce unreliable ratings as the difference between L1 (e.g., the upper bound is 100 words) and L2 (e.g., the lower bound is 101 words) is not obvious when the length is around this boundary. Raters responded positively to this scale, finding this scale precise. However, one rater suggested the inclusion of conjunction use in the scale of fluency. This seemed to show that fluency, as an aspect to describe quality of writing, is complex. It could be interpreted differently from how it is defined in the current study in Section 3.3.1.4 as rater A believed that the use of conjunctions, which contributes to cohesion in the current study, could be interpreted in terms of fluency (i.e., the ‘fluent’ text that is felt by readers during reading). However, it seems that the rater’s suggestion did not influence how he/she use the scale. Therefore, no revision is recommended for this scale. This scale is recommended to be adopted in any further use.

#### *7.3.1.3 Accuracy*

The three raters generally rated similarly using the scale of accuracy as the number of levels employed in the rating is similar (i.e., L2 to L4 by two raters, L3 to L5 by one rater) and most of the levels were covered (e.g., three out of five levels) (see Table 7.3). This was possible as the rating scale consisted of approximations of error-free sentences, and there were measurement errors due to different rating behaviors. A difference in rating was found in the rating training exercise when one rater first counted the number of error-free sentences and the overall number of sentences and another made a rough estimation and did not count sentence-by-sentence. However, the difference was small and only one level of difference was found. Raters responded positively to this scale, especially the error taxonomy attached to the rating scale. Two raters found the error taxonomy was comprehensive. One rater found the approximation of error-free sentences very useful in writing feedback or remarks to students when she taught the TEM4 writing preparation course. One rater, though not making any direct criticism or suggestion to the scale of

accuracy, commented in the rating training exercise that she considered asking students to use an online scoring system (i.e., Pigaiwang, 2019) for assessing accuracy. This indirectly corresponded to the prevalence of automatic scoring in the field of language assessment (e.g., E-rater used in TOEFL assessment) and infers that the scale can be used interchangeably with online automatic scoring. It was surprising that no rater mentioned the unfairness of the error-free measure that was discussed in Section 3.3.1.3 (e.g., unfair to those who commit fewer errors per T-unit). No revision was recommended for this scale. It was recommended to be adopted in any further use. However, it might be beneficial to be aware of the rating difference caused by different rating behaviors (e.g., counting or estimation) in training of raters using this scale.

#### *7.3.1.4 Coherence*

Raters rated differentially using the scale of coherence in terms of the number of levels employed in the rating, as is evidenced by the range of scores used by the three raters (see Table 7.4Table 2.1). One rater used L2 to L4, one rater used L1 and L5, and the third rater used L2 and L3. This seem to show that the third rater tended to award central level points to writing scripts. Central tendency should be avoided as it fails to provide more useful information to student writers as diagnostic information (Knoch, 2009). Central tendency can be caused by either rater being unable to use the scale consistently or raters have difficulty in using the scale. However, exact reasons why central tendency occurred are inconclusive until the rating process is probed. Raters did not provide comments on how this scale was used and their comments were restricted to the discourse-related sequential progression. Raters' opinions varied on the use of 'we' and 'you'. Raters who had rich TEM4 training and rating experience (e.g., more than five years) commented that the use of 'we' and 'you' was penalized in TEM4 rating, which was consistent with the scale of coherence on the new rating scale; another rater commented that the use of 'we' and 'you' should not be penalized as it encouraged stereotypes in writing. The third rater commented that it depended on the topic of the writing task and in the writing task about

public topics, such as air pollution in the usability study, the use of these words was indicative of poor writing quality. However, this also reflects how the new rating scale was influenced by the TEM4 rating scale in terms of the use of ‘we’ and ‘you’ as the overuse of these words was scored as poor in quality using the TEM4 rating scale and these features were captured by the new rating scale which was developed on the basis of occurrences of these words. Since raters did not provide much information on this scale and they disagreed on the use of ‘we’ and ‘you’, a detailed analysis of raters’ rating process using this scale (e.g., interviews with raters about the rating process) was recommended before any further use of this scale is adopted.

#### *7.3.1.5 Argumentation*

Raters generally rated similarly using the scale of argumentation in terms of the number of levels employed in the rating as is evidenced by the range of scores (e.g., L2 to L4 by two raters, L1 to L5 by one rater) and by most of the levels being covered (e.g., three or five out of five levels) (see Table 7.5). Raters’ responses were mixed. Raters responded that the scale of argumentation was very useful in facilitating their understanding and evaluation of the argumentation ability of writers, and commented that they would use them in their teaching of argumentative writing. However, raters also responded that they found it relatively difficult to identify levels of reasons and their relationships, especially when there were errors in sentences that confounded their meaning. One rater reported that she occasionally rated this category holistically when difficulty of identification occurred. These findings were not reported in the literature (e.g., Chase, 2011). It could be possible that raters lacked training in using this scale and more training could result in less difficulty in using the scale. This was mentioned by one rater who commented that the scale invited an analysis of the scripts which were not familiar to raters. Therefore, more training of this scale (i.e., the text analysis involved in the use of this scale) is recommended before it is adopted in any further use. It might be useful if terms could be explained in a more accessible way to writing teachers or raters.

Knoch (2009) also argues that evidence should be found to support the view that rating scales reflect current applied linguistics theory as well as research in terms of the construct validity of rating scales. Her argument is in line with those who argue for theoretically-based rating scales (e.g., North, 2003). In view of this feature, the evidence can only be found in the development phase. As has been described in Chapters 3 and 4, the categories were selected from an adapted model of argumentative ability based on current models of writing, and measures of these categories were selected from a comprehensive literature review of writing studies employing these measures. In this sense, the new rating scale could be regarded as reflecting current applied linguistics theory and research.

### **7.3.2 Reliability**

Bachman and Palmer (1996) define reliability as consistency of test scores across different testing situations (e.g., prompts, raters). Historically, reliability was considered to be separate from validity, but since Mesick's unitary conceptualization of validity reliability has been seen as one aspect of validity (Chapelle, 2012). Different methods are employed to assess the reliability of a test (e.g., test-retest, parallel forms, rater reliability). The reliability has not been explicitly defined for rating scales but can be inferred from the methods that are used to establish the reliability for rating scales in literature (e.g., Knoch, 2009; Li, 2010). That is, rating scales are a factor that influences ratings, thus influencing the reliability of a test. This is possible as the reliability of a test and the reliability of a rating scale for that test are both investigated through the comparison of test scores, although the focus is shifted. That is, the investigation of variations of test scores is more focused on factors influencing rating (e.g., types of rating scales, scale descriptors, rater background, and rating behaviors) in establishing the reliability of a rating scale than on factors influencing testing (e.g., test forms, administrations). Researchers on the reliability of rating scales include those who are focused on the

consistency of scores (e.g., Li, 2010) and those who are focused on all possible factors (e.g., Knoch, 2009; Li, 2010). The statistics for consistency explored in my research are Cronbach's Alpha, and Cronbach's Alpha If Item (rater). Cronbach's Alpha is indicative of the overall reliability of the three raters using the new rating scale, and Cronbach's Alpha If Item (rater) is Deleted is indicative of the contribution of a rater to the overall rater reliability, thus identifying raters who have problems in using a rating scale. Since the scores from separate trait scales are not combined, the discussion only focuses on scales individually, rather than the overall scale.

The Cronbach's Alphas for scales of mechanics and fluency show that these two scales have good reliability, which indicate that these two scales can provide reliable criteria for ratings. Cronbach's Alphas If Item is Deleted show that raters using these two scales all contributed positively to the overall reliability, which indicate that raters did not have problem using the two rating scales. It is possible that this is because these two scales are fairly mechanical and precise in their descriptors. An exceptional case is shown by rater C. Cronbach's Alphas If Item is Deleted shows that rater C contributes negatively to the overall reliability of the three raters using the scale of mechanics. It could be because rater C was inconsistent in using the scale of mechanics. This can only be known if rater C's rating process is studied.

The Cronbach's Alphas for scales of accuracy and coherence show that these two scales have poor reliability. Cronbach's Alphas If Item is deleted for the scale of accuracy show that raters all contributed positively to the overall reliability. It is surprising to see that the scale of accuracy is poor in reliability since its descriptors are precise and clear. Cronbach's Alphas If Item is Deleted for the scale of coherence indicate that raters either rated conversely to other raters, or contributed negatively to the overall reliability. These measures of reliability show that this scale is highly unreliable. However, raters did not provide any suggestions or criticism on this scale (more details can be seen in Section



7.2). It could be possible that the features of unrelated topical progressions and the discourse-related topical progressions, characterized by the use of ‘we’ and ‘you’ at different levels, were contradictory for these raters. It would be useful if the raters could have been interviewed to find out how they used these two scales, however, it was not until the latter stage of data analysis that I found there was no mention in the questionnaire feedback of any difficulty the raters had in using these two scales, and it was rather late to conduct interviews then as raters might have forgotten details by that stage.

The Cronbach’s Alphas for scales of argumentation show that this scale is moderate in reliability. It is consistent with the raters’ perceptions that they experienced difficulty in identifying levels and relationships between different reasons. Cronbach’s Alphas If Item is Deleted for the scale of argumentation show that two raters contribute positively to the overall reliability, while rater A contributes negatively. It can be shown from raters’ perceptions that, given more training, raters could be expected to rate more reliably. However, it should be noted that increased training time could influence the practicality of the scale.

### **7.3.3 Authenticity**

Bachman and Palmer (1996) define authenticity as the degree of correspondence between a test task and tasks students engage in outside of the testing context. Weigle (2002) sees authenticity as concerning whether the rating process initiated by rating scales resembles a reading process. She argues that rating writing performance holistically is a more natural form of reading than rating analytically. Along the same lines, Knoch (2009) argues that the authenticity of a rating scale concerns the extent to which the rating process is representative of how readers would approach a piece of writing outside of the test context. The new rating scale consists of five trait scales. Raters needed to read the piece five times and award a score each time using these scales. Raters responded that they did not rate holistically, except for one rater who occasionally rated globally when she had

difficulty identifying reasons and their relations. In this sense, the new rating scale is not authentic as it does not resemble a real reading process. However, it should be noted that authenticity is not the main focus of classroom-based assessment, but detailed information on strengths and drawbacks of test performances are.

#### **7.3.4 Impact**

Bachman and Palmer (1996) define impact as the effect that tests have on individuals (e.g., test takers, teachers), educational systems, and society. Weigle (2002) specifies the impact of a rating scale as the completeness of information provided by holistic scales and analytic scales (e.g., single score, more scores) and the positive consequence to rater training. Knoch (2009) specifies the impact of rating scales as relevance, completeness, meaningfulness of feedback provided by the rating scale to test takers and teachers, and the positive consequence for raters in the diagnostic assessment context. Turner (2010) describes the main purpose of classroom-based assessment as providing feedback to teaching and learning and facilitating teaching and learning. She also points out that diagnostic assessment is another term for classroom-based assessment. In the following, I will discuss the impact of the rating scale in terms of relevance, completeness, meaningfulness of feedback to test takers, and the positive consequences of the new rating scale for teachers and raters.

No data were collected to establish whether feedback was relevant, complete, and meaningful to test takers, as no test takers were interviewed and no questionnaire was conducted for this purpose. However, it can be speculated as to how test takers would respond if they were interviewed. Literature shows that test takers are more inclined to act upon feedback when it contains detailed information on their strengths and weaknesses and concrete descriptions of their performance (e.g., Alderson, 2005). The new rating scale consists of five trait scales and each trait scale has concrete descriptors and includes a detailed analysis of the texts. Therefore, the new rating scale and the textual

analysis attached to it can be expected to provide a source of more meaningful feedback (e.g., concrete descriptions) than the feedback from other scoring systems (e.g., TEM4).

Teachers (also raters in the case of the usability study) found the rating scale, and the textual analysis involved, useful in providing more relevant, specific, meaningful feedback to students. One teacher responded that she had already used the scales of argumentation and coherence in demonstrating the relations between arguments and discourse-related topics and providing more specific feedback to students. Other teachers responded that they would use the scales in providing more detailed and specific feedback to students than using the current practice.

Raters (also teachers in the case of the usability study) found that their awareness of evaluation of argumentative writing was increased. They found the concrete descriptors useful in guiding their rating, although descriptors were occasionally found difficult to interpret or apply. However, one rater suggested more training than was conducted in the current study would result in more convenience in using the scales.

### **7.3.5 Practicality**

Bachman and Palmer (1996) define practicality as the extent to which resources support test development and administration (e.g., human resources, material resources, designing tasks, administering tests, scoring and score reporting). Weigle (2002) compares the practicality of holistic and analytic scales in terms of the difficulty in applying a rating scale, time spent on scoring, and expenses paid for scoring. Knoch (2009) adapted Bachman and Palmer's (1996) view of practicality and investigated the practicality of the DELNA scale in terms of its development and use. The practicality of the new rating scale is next discussed in terms of the extent to which the resources available for its development and use are adequate.

The new rating scale can be considered very time-consuming as it was based on the analysis of a large sample of writing scripts and all these writing scripts need to be scored load. However, it could be argued that since it is very time-consuming in developing a rating scale using this development method, the current rating scale did not include features that are specific to individual argumentative writing tasks, but rather features specific to a type of discourse-argumentative writing. That is, teachers could use this scale to the assessment of argumentative writing irrespective of individual argumentative writing tasks while they do not have to be concerned about development of a new one. As for the practicality of the new rating scale in terms of its use, it could be also argued that although the current educational system (i.e., large classes, no classroom rating payment) does not quite allow a more detailed rating scale, it does not necessarily mean that this scale is not useful especially when raters responded positively on its usefulness in teaching and writing feedback to students.

Interactiveness is defined as the extent and type of involvement of the test taker's individual characteristics in a test. The facet of interactiveness in the framework of test usefulness is not discussed in the current study as Weigle (2002) argues that the interactiveness of the rating scale can only be established through investigating the involvement of test takers when they know how they would be evaluated by the rating scale, which remains an empirical question and could be answered in a different study.

### **7.3.6 Conclusion**

The usability of the new rating scale was investigated in terms of construct validity, reliability, authenticity, impact and practicality. Quantitative analysis of the scoring given by the three raters using the new rating scale shows that the scales of mechanics and fluency are excellently reliable, the scales of accuracy and coherence are poorly reliable, and the scale of argumentation is moderately reliable. Since no follow-up interviews were conducted in this study, exact reasons were unknown about the low inter-rater reliability

of the scales of accuracy and coherence. However, it could be expected that with more training, the reliability of these scales could be improved. Qualitative analysis of the raters' responses to the questionnaire questions supports the following arguments: the new rating scale is generally adequate in representing the construct of argumentative writing ability and number of levels of proficiency for the target population; the new rating scale is not authentic; the impact on the test takers, teachers, and raters is positive; the new rating scale is not practical. Although the new rating scale has been shown to be not practical, and not authentic, as the primary aim of this scale is to provide students with detailed feedback about their strengths and weaknesses, which is time-consuming in itself, the practicality and authenticity (identified with holistic rating) could be regarded as less important than other quality facets in the context of this study. Therefore, it could be argued that the new rating scale is generally usable and suitable for classroom assessment, although efforts need to be made to investigate raters' use of the scale of coherence and improve the training of raters using the new rating scale, especially the scales of accuracy, coherence and argumentation.

## **Chapter 8 Conclusion**

### **8.1 Introduction**

The overall aim of the thesis is to develop a rating scale of Chinese EFL college students' argumentative writing ability in the context of classroom assessment, using a theoretically-based data-driven approach to the scale design. Research questions are (1) Which discourse analytical measures are successful in distinguishing between argumentative writing samples from Chinese EFL college students majoring in English at different proficiency levels? (2) Is a new theoretically-based data-driven rating scale usable by Chinese EFL teachers of argumentative writing?

In order to develop the scale, potential discourse measures were identified for 12 constructs or knowledge components that the literature survey showed are assumed to be necessary for argumentative writing ability. These discourse measures were identified using a theoretical model of argumentative writing ability based on three models of language use: the CLA model (Bachman, 1990; Bachman & Palmer, 1996), the taxonomy of academic writing skills, knowledge bases and process (Grabe & Kaplan, 1996) and the model of writing competence (Connor & Mbaye, 2002).

Two hundred and fifty-eight writing scripts at four different proficiency levels were analyzed using potential discourse analytic measures. Descriptive statistics and inferential statistics were conducted. A rating scale was built based on the findings of the data analysis: scale descriptors were formulated using discourse analytical measures that successfully distinguished between different writing levels; the number of levels was determined by amalgamating or extending four proficiency levels based on descriptive statistics of between-level differences; modifications were made to descriptors and number of levels to accommodate limitations that human rating imposes, for example, the between-level difference should be obvious to detect and practical to use by human raters.

In order to investigate the usability of the new rating scale, thirty writing scripts, representing a wide variety of proficiency levels, were selected. Three argumentative writing teachers who were potential users of the new rating scale were selected to rate the thirty writing scripts using the new rating scale. Quantitative analysis was conducted on the rating results to investigate whether the ratings produced by the three raters using the new rating scale were reliable. A follow-up questionnaire was conducted to investigate how the raters' perceived the new rating scale. Content analysis was conducted on their questionnaire feedback. The usability of the new rating scale was then discussed in terms of Bachman and Palmer's (1996) usefulness framework of a test: construct validity, reliability, authenticity, impact, and practicality. A usability argument for the new rating scale was then given.

There follows four main sections. The first is a summary of findings that have been arrived at in this study. The second is focused on the implications. The third is focused on the limitations. The fourth is suggestions for further research.

## **8.2 Findings**

Based on a comprehensive theoretical model of argumentative writing ability and statistical analysis results, five constructs were selected as the basis for the features or traits in the new rating scale. They are: accuracy, fluency, mechanics, coherence, and argument structure. Ten discourse measures that were successful in discriminating between different levels of argumentative writing performance of Chinese EFL college students, and practical to use by raters, were selected for the development of level descriptors of the new rating scale. Five discourse measures were selected from qualitative analysis of the writing scripts. Table 8.1 presents constructs and discourse measures which were chosen as the basis for the development of the new rating scale.

Table 8.1 Discourse analytic measures included in the rating scale

Constructs	Measures
Accuracy	Ratio of error-free T-units
Temporal fluency	Number of word tokens
Mechanics	Number of paragraphs
Coherence	The proportion of parallel progression The proportion of extended parallel progression The proportion of discourse-related topical progression The proportion of extended sequential progression The proportion of unrelated topical progression
Argumentation structure	The proportion of introduction The proportion of level-1 reasons (qualitative analysis) The proportion of level-2 reasons and below The proportion of non-functional elements (qualitative analysis) The proportion of yoursides arguments (qualitative analysis)

The analysis also identified a number of discourse measures which were not discriminating between different levels, not practical to be used by raters in a rating process, or not common in the writing sample. These measures were not included in the rating scale. A number of constructs were also excluded from the design of the rating scale because no discourse measure was found suitable. These constructs are: repair fluency, lexical complexity, grammatical complexity, cohesion, topic-level argumentative structure, register, the soundness of argumentation, and appeals of argumentation. The constructs and measures which were not included in the rating scale are presented in Table 8.2.



Table 8.2 Discourse analytic measures not included in the rating scale

Constructs	Measures	Reasons for exclusion from pilot study or main study
Accuracy	Number of errors per T-unit (E/T), Number of errors per clause(E/C)	Not practical to use; excluded from the pilot study
Repair fluency	Number of revisions	Not discriminating; excluded from the pilot study
Lexical complexity	Sophisticated word types per word types (SWT/WT) Lexical words per words (LW/W)	Not discriminating; excluded from the pilot study
Grammatical complexity	Mean length of clause (MLC) Mean length of sentence (MLS) Clauses per T-unit (C/T) Complex nominals per clause (CN/C) Coordinate phrases per clause (CP/C)	Not discriminating; excluded from the pilot study
Mechanics	Number of capitalization errors	Not common; excluded from the pilot study
	Number of spelling errors	Not discriminating;
	Number of punctuation errors	excluded from the pilot study
Cohesion	Number of references	Not discriminating; excluded from the pilot study
	Number of lexical cohesion	Not practical to use; excluded from the pilot study

		study
	Number of incorrect use of cohesive devices	Not common; excluded from the pilot study
Coherence	Proportion of SP1 progression	Not discriminating;
	Proportion of SP2 progression	Excluded from the main
	Proportion of related progression	study
	Number of metadiscourse markers	Not discriminating; excluded from the pilot study
	Number of errors in metadiscourse markers	Not common; excluded from the pilot study
Topic-level argumentative structure	Number of paragraphs per argumentative move	Not discriminating; excluded from the pilot study
Register	Number of personal pronouns, contractions, formal and informal metadiscourse markers, colloquial word, 'knowledge-base' use of because	Not discriminating; excluded from the pilot study
The soundness of argumentation	A 1-3 scale of acceptability	Not practical to use; excluded from the pilot study
Appeals of argumentation	Number of R8-the appeal of cause/effect-means/end-consequences The development of appeals	Overlapping with the proportion of justification in argumentation structure; excluded from the pilot study

	Number of other types of the appeals	Not common; excluded from the pilot study
Argumentation structure	The proportion of standpoints	Not discriminating;
	The proportion of conclusion elements	excluded from the main study
	The proportion of functional elements	Not common; excluded from the main study

Quantitative findings from Phase 2 showed mixed results on the reliability of the new rating scale. The scales of mechanics and fluency were found to have good reliability. The quantitative findings were consistent with the quantitative findings for these two scales. Raters responded positively to these two scales as these scales provided precise and clear descriptors. The scales of accuracy and coherence were shown to be problematic as they were found to have poor reliability; no explicit negative feedback was found on these two scales, although one rater responded that she would consider using an automatic scoring system to evaluate language accuracy, and in the training session it was also found that there was a difference when raters employed different rating behaviors (e.g., judgements based on counting or judgements based on estimation). It is not clear why the scale of coherence did not work, and no suggestion or criticism was made of the scale. It is possible that raters had difficulty in identifying different types of topical progression or the two types of topical progression (i.e., unrelated and discourse-related topical progression) are contradictory in providing reliable criteria. Compared with the scale of coherence, it was surprising to find that the scale of argumentation was moderately reliable while raters found this scale difficult to use. It could be possible that raters tended to rate globally when they had difficulty and the quality of argumentation is strongly correlated with the overall quality of the writing scripts.

Qualitative findings showed that the new rating scale was not practical in terms of use,

but literature suggests that moderations could be made to make the scale less descriptive in operational situations (e.g., Weir, 1990). It was also found from the qualitative findings that the new rating scale provided benefits to teachers in terms of clear, specific and detailed feedback they could provide to students in their class. Also, the new rating scale could be used to develop detailed training materials for raters, and raters suggested that the time spent on training was necessary to guarantee a proper use of the scale. It was also found from the qualitative findings that the new rating scale was not authentic, as raters did not rate holistically, apart from one rater who experienced difficulty in using the scale of argumentation.

### **8.3 Implications**

The implications of the study are both theoretical and practical.

#### **8.3.1 Theoretical implications**

The first theoretical implication relates to the scale classification. In a summary table (see Table 2.1), four different types of rating scales were established based on previous research, in terms of score reporting, number of trait scales, task generalization, generic or context-specific, and ESL/EFL setting (Weigle, 2002; North, 2003; Fulcher, 2003). According to the table, the multiple trait scale was distinguished from the analytic scale as the former was developed for a type of task, is context-specific, and is not commonly used in ESL/EFL setting. This study suggests that it may be possible to categorize a subtype of the multiple trait scale as a theory-based and data-driven multiple trait scale. This scale is more generalizable than the typical multiple trait scale developed by Hamp-Lyons (1991) because of its use of a comprehensive list of traits of writing at the start of the design process. This is different from Hamp-Lyons's (1991) scale design process. Hamp-Lyons decided on the traits of her scale herself during the analysis (a typical scale based on experts' intuition – see Section 2.2.4). The degree of context-dependence of the new scale is also different from the typical multiple trait scale, as the new scale only takes

into account the discourse type to which the test responses belong. Other context factors like topics and prompt characteristics are not taken into account in the scale. Though more generalizable than the typical multiple trait scale, the new scale is less generalizable than an analytic scale as the score cannot be generalized to other types of writing task, like describing graphs, tables, or diagrams. Therefore, the summary table of scale classification can be expanded in the following manner (see Table 8.3).

Table 8.3 Extended scale classification by including the theory-based and data-driven multiple trait scale

Scales	Primary trait	Multiple trait		Holistic	Analytic
		Typical	<b>Theoretically-based and data-driven</b>		
Score or scale	Single score	Multiple scores	Multiple scores/+single score	Single score	Multiple scores
Task generalization	Specific to a particular task	Specific to a type of tasks	Specific to a type of discourse: argumentative discourse	Generalized to a variety of task types	
Generic or context-specific	Context-specific	Context-specific	Context-specific	Generic	
ESL/EFL setting	Not commonly used			Commonly Used	

The second theoretical implication relates to the construct of a second language/foreign language argumentative writing test. Argumentative writing ability is a fundamental writing ability for Chinese EFL college students enrolled in English programs. It also prepares Chinese EFL learners for learning more complicated academic writing at higher levels. However, most current ESL/EFL writing tests emphasize writing as a language ability while they neglect writing as a cognitive activity, for example, containing reasoning skills. This is reflected in most EFL/ESL rating scales (e.g., International

English Language Testing System (IELTS) writing rating scale, the rating scale of TOEFL writing in different versions, the TEM4 rating scale, and the Michigan English Language Assessment Battery (MELAB) rating scale)). These scales are more focused on generic linguistic, syntactic, organizational features, like accuracy, fluency, flow of ideas, topic development, thesis statement, while the quality of argumentation is disguised under features such as communicative effectiveness, task fulfilment, or content. One might argue that reasoning or the quality of argumentation is not part of language ability, and therefore should not be included in a language test. However, this does not justify not evaluating the quality of argumentation in argumentative writing tests as it is impossible for raters to evaluate quality of writing without taking into account the quality of argumentation. This research sets out to build a construct of argumentative writing ability in the communicative language testing context and includes, as an integral part, the ability to make arguments. The theoretical framework is represented in Table 8.4. The argumentation knowledge highlighted in bold includes the structure, substance and appeals of argumentation. This theoretical framework is expected to bring better understanding of argumentative writing ability for writing test developers and scale developers.

Table 8.4 A communicative model of argumentative writing ability

Argumentative writing ability	Language competence	<ul style="list-style-type: none"> <li>Grammatical knowledge</li> </ul>	<ol style="list-style-type: none"> <li>Knowledge of the written code               <ol style="list-style-type: none"> <li>Spelling; b. Punctuation; c. Paragraphing; d. Capitalization</li> </ol> </li> <li>Knowledge of morphology (word-part knowledge)</li> <li>Vocabulary</li> <li>Syntactic knowledge</li> </ol>
		<ul style="list-style-type: none"> <li>Textual knowledge</li> </ul>	<ol style="list-style-type: none"> <li>Cohesion</li> <li>Coherence</li> <li>Rhetorical organization</li> </ol>

		<ul style="list-style-type: none"> <li>• Sociolinguistic knowledge</li> </ul>	1. Formal and informal writing style 2. Idiomatic expression, cultural reference, and figures of speech
		<ul style="list-style-type: none"> <li>• Argumentation knowledge</li> </ul>	<b>1. The structure of argumentation</b> <b>2. The substance of argumentation</b> <b>3. The appeals of argumentation</b>

The third theoretical implication is related to theoretical and operational frameworks of argumentation ability. It is expected that the frameworks can benefit the evaluation of the quality of argumentation in computer-based automatic essay scoring. Automated essay scoring (AES) is commonly recognized as cost-effective, objective, consistent and impartial compared with human raters (e.g., Shaw, 2004). Research shows that current automated essay scoring systems can assess linguistic features, syntactic features, content and organization, but are unable to assess quality of argumentation (e.g., Paul, 2013). Paul (2013) argued that this situation may change as technical issues are resolved, like Natural Language Processing (NLP) techniques in automated identification of structure of argumentation advance. Although the current study does not provide automated identification techniques, the theoretical and operational frameworks of the argumentation ability proposed in the current study are helpful by providing a ready-to-use framework for the development of automated identification techniques. The argumentation ability is theorized in the framework as being able to make a single argument, arrange chains of arguments to defend the author's standpoint, convince readers by employing persuasive appeals, and address both sides in an argumentation. The operational framework provides a summary of quantifiable measures for the structure of argumentation and the acceptability of reasons and persuasive appeals; these measures were trialed on students' essays and proved to be useful with, for example, the depth of reasoning, argument types in terms of hierarchical relations between reasons (e.g., convergent arguments, coordinate arguments), and number of argumentation elements.

### **8.3.2 Practical implications**

The practical implications relate to the weighting of trait categories, score reporting and feedback, rater training, and teaching of argumentative writing.

First, the weighting of trait categories in this scale is flexible. The weighting of trait categories was not covered in the study. It is expected that the weighting of trait categories can be decided by the context in which the rating scale is used. It is sensible to give more weight to certain categories than others. For example, in assessments in which argumentation is the focus, it might be necessary to give more weight to the scale of argumentation. In other situations, language features might be given extra weight.

Second, this scale can provide the score report for both formative and summative assessments. Weigle (2009, p. 286) summarizes the way scores are reported in four common test types. In proficiency tests, where general writing ability of students is expected to be assessed, one averaged score is often reported. In placement tests, one averaged score or groups of trait scores are reported, as students are expected to be placed in a specific course according to the averaged score or different trait scores. In achievement tests, the score reporting is dependent on the focus of the course as the test is usually designed to assess students' achievements in accordance with the teaching objectives of the course. In diagnostic tests, trait scores and detailed information on the strengths and weaknesses of students' writing ability are expected. In this study, the new rating scale is aimed at Chinese EFL college classroom assessments. In summative assessments, where an overall score is needed, separate trait scores like fluency, accuracy and coherence can be averaged using different weighting systems. In formative assessments, when the strengths and weaknesses of students' argumentative writing ability need to be identified and detailed feedback needs to be provided to students, the descriptors in the new scale and the definition of measures which are converted into descriptors will provide the basis for the feedback. For example, if a student scores level



2 on the argumentation scale, he or she may read feedback on his or her performance on argumentation as presented in Figure 8.1. Following the format used by Knoch (2009) in the DELNA diagnostic test, the feedback will first summarize an ideal performance on argumentation that students who take the tests are expected to achieve, then describe how the student performed in terms of argumentation, and finally provide concrete suggestions on how the student can improve his or her performance. Marks are needed to label these features on that student's writing script to better understand the feedback, or, for example, in Figure 8.1 line numbers are used to indicate the position where a particular feature occurs.

In argumentative writing, students generally provide strong justifications for their standpoint. They often use reasons which are well-developed, accurately expressed and relevant to the topic. They also acknowledge the existence of potential opposing arguments and successfully rebut them. A mixture of coordination, convergent and coordinate relations are used to inter-connect reasons to provide strong support.

In your essay, although a number of level-1 reasons were provided, few of them were further justified. Two reasons were not accurately expressed (e.g., in Lines 2 and 5). One reason was nonsensical (e.g., in Line 9). Opposing arguments and rebuttals were not found.

You should justify level-1 reasons. You may want to acknowledge obvious opposing arguments when possible and rebut them in your essay. Efforts should be made to the use of words to make sure the meaning of the sentence is well

Figure 8.1 Feedback on argumentation for test takers

Figure 8.1 illustrates how feedback can be developed based on the new rating scale and benefit from the scale development process. Feedback with a similar format can be

drafted for other traits of writing ability. For example, for feedback on accuracy, the number of errors and the type of errors a student commits would be reported in the 'performance' section, with the help of the taxonomy of error types used in this study. For feedback on coherence, different ways of topical progression would be described in the 'ideal performance' section, and topical progression used by the students would be reported in the 'performance' section. Suggestions on how to identify discourse topic and sentence topic, and how topics are related to each other to avoid unrelated topical progression would be stated at the end of the feedback. For mechanics, the ordering of information, and unnecessary short or long paragraphs would be used to describe students who write a small number of paragraphs, together with the total number of paragraphs in 'performance' section. Finally, a profile of performance on different traits can be provided to show the weaknesses and strengths of a student's ability with trait scores and detailed feedback, thus providing positive feedback in the teaching of argumentative writing.

Third, the scale can be beneficial to rater training. Raters are found to differ in overall leniency, in displaying a pattern of leniency in relation to particular test tasks, or a group of test takers (or a bias), in systematically avoiding using extremes of scores (or central tendency), in tending to rate holistically rather than analytically (or halo effect), and in rating inconsistently from each other (e.g., McNamara, 1996; Myford & Wolfe, 2003, 2004). Rater training is a common practice to improve the reliability of rating. Two major steps in rater training are to identify benchmark essays which represent the different points on the scale or typical problem areas, and to discuss in groups the criteria for different features in relation to these benchmark essays (e.g., Weigle, 2002). By using this scale, feedback can be developed for benchmark essays and definitions of measures can be provided. The feedback, definitions of measures and benchmark essays can better describe different features of these essays and provide a common language for the group discussion of criteria. With these materials for rater training, raters are expected to rate more consistently, extreme scores can be better avoided, and rating criteria can be

clarified, thus facilitating rater training.

Fourth, the scale can also have pedagogical implications. Weigle (2002, p. 137) identifies a number of instructional effects of rating scales. Rating criteria allow teachers, students, and other stakeholders to have a frank discussion of, or a consensus about, the instructional goals of writing courses and the expected outcomes for students. Rating criteria allow teachers to gear their instruction towards the aspects of writing that are emphasized in the criteria but are not valued in their teaching. Research shows that both L1 and ESL/EFL undergraduate writers have difficulty in identifying the concept of arguments and developing their position in a debate, and requirements on argumentative writing are not explicit (Wingate, 2012; Hirvela, 2017). The rating scale for the quality of argumentation can improve this situation if teachers can adjust their teaching goals towards the different aspects of argumentative ability. Students may be motivated to work on those aspects of writing if they are aware of how their writing is scored.

#### **8.4 Limitations**

Although the study was carefully designed, a number of shortcomings must be acknowledged.

The first limitation, one of the major limitations of this study, is that the theoretical model of argumentative writing ability, which provided the basis for the selection of features or aspects to be included in the rating scale, was not validated. This may have reduced the validity of the new scale. Two factors may have minimized this potential reduction. Firstly, the theoretical model was as comprehensive as possible, to include all possible aspects of argumentative writing. Secondly, the analysis of writing scripts in the design process also provided the basis for the aspects to be included in the rating scale. It is therefore hoped that this limitation in the scale design only had a limited effect on the validity of the rating scale.

The second limitation is the way in which four different writing levels, used as the basis for the analysis of the writing scripts, were established. This limitation is twofold. One part was the use of an existing rating scale to assign an overall score to each writing script. A way of breaking the scripts into levels was needed. Because no independent measure of writing ability was available, the existing TEM4 rating scale was used. This means that the existing scale had a direct influence on the development of the new scale. This influence may have been reduced because it is reasonable to assume that the TEM4 score can reliably represent the general writing proficiency of writers, and thus the student writers were reliably rank ordered by the TEM4 score in terms of general writing proficiency.

The other part was the way that the levels were formulated in new trait scales. In the main analysis, essays were grouped into four levels by evenly dividing a score range of 1 to 12 using the existing TEM4 rating scale. This may have had an influence on the outcome of the study. However, this influence may have been reduced as the levels included in the new trait scales were finally justified by the data. One criterion for successful measures is being practical to use. Therefore, those measures with small between-level differences, that would be hard for raters to distinguish, were removed, which in turn reduced the number of levels that were used for describing these measures. Moreover, since I believe there would be a proficiency level above that of the population that I was working with, I added one level at the top to allow for this proficiency level.

The third limitation is that no suitable measure was established for a number of constructs in the analysis. Although a number of measures of lexical complexity and grammatical complexity were identified in the literature review and included in the analysis of writing scripts, none of these measures discriminated between different levels. Similarly, no ratio measure of cohesion was found to be discriminating between four different levels,

although the number of conjunctive devices was found to discriminate between three levels in the pilot study. More time might have resulted in potential measures which could be used in this study, for example, error severity, or particular types of metadiscourse markers. Other measures could have been used in the scale design if such measures were developed. Measures of lexical cohesion could have been included in the rating scale if the measure was easy to be used by raters. Finally, the measures of persuasive appeals, the soundness of argumentation, topical-level argumentative structure, and register were also not found to distinguish between levels in the main analysis. For the constructs of persuasive appeals and register, the measures investigated were unable to distinguish between levels because they were infrequent in the sample writing scripts. Measures could be successful if they were applied to writing scripts by more advanced learners of EFL argumentative writing ability than year-2 English majors. For the soundness of argumentation, measures of acceptability and relevance would be included in the new rating scale if they could be based on relatively quantifiable standards rather than largely dependent on raters' subjective judgments.

The fourth limitation is that the usefulness of the new rating scale was only investigated with a small sample. The underlying assumption of this study is that the newly developed rating scale is more suitable for assessing argumentative writing performance because it is based on an argumentative writing ability model (i.e., it includes the features that are not only general to any type of writing but also those specific to argumentative writing), and it is also based on results of statistical analysis, which makes level descriptors more specific and avoids vague, impressionistic terminology. However, given the scale of the study and the limited time, the usability of the new rating scale was conducted on a small sample and no study was conducted to compare the new rating scale against the TEM4 scale. Only three raters were used and these raters only rated thirty scripts each.

The fifth limitation is that how raters employed the new rating scale was not investigated

in more detail. It was pointed out in the findings section that the scales of accuracy and coherence were not reliable and raters seemed to have difficulty in using the scales, although their questionnaire feedback did not provide evidence (triangulation). The investigation of how these scales were employed by different raters through interviews would provide answers to this question. However, it was too late for any interview to be conducted at the latter stage of data analysis when interviews of raters using these scales were found to be necessary.

### **8.5 Suggestions for further research**

Suggestions for further research mainly follow from the shortcomings identified in relation to the scale development and the usability study. .i.e.

First, a number of further studies are necessary to establish new measures for different traits of writing performance. More detailed research might be able to establish if there are other measures of lexical complexity which can successfully distinguish between different writing levels of the target population. A new word frequency list drawn from a corpus of texts which represents the texts exposed to the target population might be built to provide potential measures of lexical sophistication. For grammatical complexity, it might be possible to establish that grammatical complexity is a good measure for the target population if more research can be conducted to investigate other measures of grammatical complexity. Similarly, particular types of metadiscourse markers and error severity should also be further investigated as they seem to be promising in indicating language development. As has been shown, although successful in discriminating between the different proficiency levels, lexical cohesion was excluded from the rating scale because it is impractical to use in rating. It is thus necessary to establish a practical measure for lexical cohesion. Further research is also necessary to establish less subjective measures for the acceptability and relevance of reasons. For example, it is necessary to establish whether the conditions proposed by Govier (2013), under which a

reason in an argument can be regarded as acceptable, are recognized by raters or teachers of argumentative writing. Further research is also necessary to investigate the use of persuasive appeals in more advanced learners (i.e. writers at level 5 in the current study, or year-4 undergraduate students or postgraduates enrolled in English programs) who may be more mature in their argumentative writing ability.

Second, it is also possible that further research could be conducted on applying the scale of argumentation ability to computer-based scoring and the generation of feedback, utilizing the design process of the scale for argumentation ability.

Third, it would be interesting to compare the new rating scale and the TEM4 rating scale. The underlying assumption of this study is that the newly developed rating scale is more suitable for assessing argumentative writing performance in the classroom-based assessment context because it involves detailed analysis of writing scripts and can provide descriptors which are more specific and avoid vague and impressionistic terminology. It would be very interesting to see how these two scales differ from each other on a large sample of raters rating a large sample of writing scripts in terms of construct validity, reliability, authenticity, impact, and practicality using a more sophisticated multi-facet Rasch analysis.

Fourth, further studies could be conducted to explore how test takers respond to detailed feedback (e.g., based on the new rating scale), and whether their writing is improved by acting upon that feedback.

Fifth, a study on how raters employ the new rating scale could be conducted using think-aloud protocol analysis. It was pointed out in the findings section that the scales of accuracy and coherence were not reliable and raters had difficulty in using the scale. However, their questionnaire feedback did not provide any evidence. Think-aloud

protocols would provide more details on how raters use different scales, for example, if they attend to particular descriptors in different scales, if they tend to rate harshly or leniently on particular categories, if they are inconsistent in applying the rating criteria on different writing scripts. This could provide an explanation for the inconsistency of results between the quantitative and qualitative analyses.

## **8.6 Conclusion**

This study developed a rating scale for assessing the argumentative writing of Chinese EFL college students majoring in English, by adopting a theoretically-based data-driven approach. Through this approach, the new rating scale was expected to provide reliable, specific and detailed rating criteria for classroom assessment of argumentative writing of Chinese EFL college students majoring in English. The new rating scale, though shown to be not practical and not authentic as a whole, was shown to be relatively adequate to represent the construct of argumentative writing ability, and useful to provide detailed feedback to the teaching and assessment of argumentative writing ability in classroom. Since the latter quality facets of the new rating scale are of primary importance in the Chinese EFL classroom assessment context, the new rating scale was generally regarded as usable, but efforts need to be made to further investigate raters' use of the scale of coherence and improve the training of raters in using the new rating scale.

The new rating system, comprising five trait scales, will offer potential alternatives to the existing scoring criteria for the assessment of argumentative writing in the Chinese EFL classroom. The scale development process will enable a better practice in rater training, scoring, and score reporting in Chinese EFL argumentative writing assessment, and it will also benefit the teaching of argumentative writing in China. The assessment of argumentation ability using the new rating scale will also contribute to new practice for argumentative writing assessment in the wider EFL context.



## Appendices

### Appendix 1: Interview guide: Developing argumentative writing tasks for Chinese EFL learners in writing courses

This is an unfinished taxonomy of the characteristics of argumentative writing tasks for year-2 Chinese EFL college students and parameters. Please finish the taxonomy by deciding on parameters that are suitable for the assessment of the argumentative writing ability of students whom you teach based on your experience.

•Task characteristic	•Parameters
Word limit	•
Time limit	•
Prompt	•
Topic	•
•Number of tasks	•
•Marking criteria	•
Language	•
Font	•

## **Appendix 2: Questionnaire on the difficulty of argumentative writing tasks**

Thank you very much for participating in this study. The purpose of this study is to investigate the difficulty of argumentative writing tasks suitable for testing Chinese EFL college students' argumentative writing proficiency. You will have 7 argumentative writing tasks. Assuming that you write an argumentative essay on each of these writing tasks, please rate the difficulty of these tasks according to your own writing proficiency. A 5-point rating scale is provided for the rating. Please choose the best option that fits your situation and put the answer in brackets. You are required to give reasons for your choice in terms of factors such as the topic, the number of words, the time limit, and vocabulary in testing instruction.

Your personal information will only be used to understand your answers and will not be passed to individuals or groups other than the researcher.

Please send the completed questionnaire to [kz37@le.ac.uk](mailto:kz37@le.ac.uk) as an attachment entitled Questionnaire on the Difficulty of Argumentative Writing Tasks.

Thank you again for your participation!

Personal information:

Name: \_\_\_\_\_

Grade: \_\_\_\_\_

University: \_\_\_\_\_

Major: \_\_\_\_\_

Your most recent writing performance: \_\_\_\_\_ (writing score/full score, the name of the test)

A 5-point rating scale :

A. Very difficult    B. Difficult    C. Fair    D. Easy    E. Very easy

An example:

Writing task 1: (A) means that task 1 is very difficult for you to write a good essay.

1. Task 1: ( )

- Reasons

---

---

---

---

2. Task 2: ( )

- Reasons

---

---

---

---

- 

3. Task 3: ( )

- Reasons

---

---

---

---

- 

4. Task 4: ( )

- Reasons

---

---

---

---

5. Task 5: ( )

- Reasons

---

---

---

---

6. Task 6: ( )

- Reasons

---

---

---

---

7. Task 7: ( )

- Reasons

---

---

---

---

**Appendix 3: A taxonomy of errors in English writing adapted from the CEM  
error taxonomy (2008)**

Code	Type	Sub-code	Subtype	Example of errors
fm	word form	fm1	spelling	Shrined [fm1]
		fm2	capitalization	china [fm2]
		fm3	word building	transforment
vp	verb	vp1	transitive/intransitive verbs	So I disagree [vp1] the words of Joseph Epstein.
		vp2	finite/non-finite verbs	So choose [vp2] your decision is very important.
		vp3	mood	I suggest students in school all saved [vp3] money.
		vp4	modal (misuse, redundancy, and absence)	Base [vp2] on my personal experiences and this report and my knowledge, as for some persons [np4] opinion that "we should shut down all the factories to improve our air quality", I should [vp4] say no to it.
		vp5	auxiliary (misuse, redundancy, and absence)	The environment protect [wd2] is [vp5] also need everyone's effort
		vp6	absence of a	People become more and more

			main verb within a sentence	busy [aj3] and [vp6] under great pressure.
np	noun	np1	pattern	Nowadays, many universities [np1] students tend to look for love on the net.
		np2	countability /number /agreement	And many of these factories play an important roles [np2]
		np3	case	the air pollutants [np3] discharge
		np4	determiners	Quantifiers: many [np4] precious time Demonstrative determiner: The society has taken great changes all those [np4] last few years. Wh-determiners, and numerals
pr	pronoun	pr1	reference	We have enough money to fulfill all his [pr1] desire.
		pr2	wrong use of anticipatory it	If you have accident or surprising things, it [pr2] is no way that you have no storage money.
		pr3	case	So we [pr3] lives formed.
		pr4	agreement	Both [pr4] of the people found this [pr2] hard to get good work.
		pr5	misuse, absence and redundancy of wh-	You can do anything what [pr5] you want to do.

			pronouns, relative pronoun and interrogative pronoun	
		pr6	misuse of indefinite pronoun (, all/both, few/little, some/any, either/ neither)	Admittedly, both [pr6] of the people found this [pr1] hard to get good work [wd].
aj	adjective	aj1	degree	Ambition can make ourselves more stronger [aj1]
		aj2	-ed/-ing confusion	We can use the saving [aj2] money.
		aj3	predictive/ attributive	An alive [aj3] person must have an aim.
		aj4	pattern	Some people are easy [aj1] to get satisfied, so their ambition could be to live in a healthy way.
ad	adverb	ad1	order	We can see simply [ad1] that there are many advantages of using tomorrow's money
		ad2	modification	I also very [ad2] supported the latter opinion
		ad3	degree	It will encourage us to work hard

				[ad3] and earn more money in order to pay back the loan
pp	preposition	pp1	pattern	In the third place, it is ambition that cultivates the spirit advantage of people during [pp1] the way to succeed.
		pp2	absence of a preposition	Economic in the world is [pp2] a very unstable position now (based on the stocks), at least we can't shut down them [wd1] right away.
cj	conjunction	cj1	wrong use of conjunction	Everyone, whether poor and [cj1] rich, old and [cj1] young, can have the ambition
		cj2	absence of conjunction	You must choose, [cj2] you will continue to study, [cj2] you will go to work
wd	word	wd1	order	At least we can't shut down them [wd1] right away
		wd2	set phrase	Noun: ...the fixed-line phone [wd2]... Verb: some years ago, he made his mind up [wd2] to get PhD. Adjective: ...for those who are interested to [wd2] join into be a competitor. Preposition: in the other hand [wd2], Conjunction: the last but no least



				[wd2]
		wd3	part of speech	An ambition [wd3] and determined person can make full use of what little he has.
		wd4	wrong choice of words especially synonyms and words with similar spelling	Although it is apparent that shutting down the factories is a practical and direct method, it is not considerate [wd4] enough
		wd5	redundancy	First, we are students without having [wd5] the ability to earn enough money to support ourselves
		wd6	repetition	Now [wd6] I remembered it till now.
cc	collocation	cc1	noun/noun	The emission of air pollution
		cc2	noun/verb	PM2.5, which reflects China's severe air pollution problem, has conducted [cc2] a heated discussion on whether polluted [aj4] plants should be shutted [fm1] down or not
		cc3	verb/noun	They are [vp6] work hard, [cj3] broaden their eyes[cc3].
		cc4	adj/noun	They don't have free money[cc4] to give the student to buy books
		cc5	linking	The best way to make healthy [cc5]

			verb/adj	is to exercise
		cc6	adv/adj	...if we have ambition, we will strive for it and our ambition will urge us to be largely hardworking [cc6]
sn	sentence	sn1	run-on sentence	Although compared to factories' emission, these pollutants can be omitted, every little [cc] harms so that they can be a [ar2] enormous threat to our environment. [sn1]
		sn2	sentence fragment	For the factories discharging air pollutions are often blamed as a major culprit of this environmental disaster. [sn2]
		sn3	dangling modifier	Agreeing with Epstein, ambition is not only feature that should not be refrained from.... [sn3]
		sn4	illogical comparison/ non-factual statement	First, the sewage of factories is the main source of PM2.5. [sn4]
		sn5	topic prominence	The expensive but useful books, maybe you can't afford it, so the saved money can help you. [sn5]
		sn6	coordination	Once we have the habit of saving money, it will help us in our future career or lives, and cherish our existing belongings. [sn6]

		sn7	subordination	Unless all of the cars dumped, trains forbidden [sn7], the PM 2.5 problem would not get much better.
		sn8	structural deficiency	Firstly, air pollution [vp7] not just because these industries and these companies, It's a comprehensive question which should involves [vp5] the world's efforts not only just China close down the polluted companies. [sn8]
		sn9	voice	...all their cost will offord [fm] by their parents. [sn9]
		sn10	unreadable sentences	What if our daily life needing [vp] that[np8] products, the economic growth, and the unemployed people after shutting down these factories? [s10]
		sn11	translated sentences	As an old saying, the sky can't fall down the pie. [sn12]
tn	tense	tn1	subject-predicate agreement	It make [tn1] you feel so bad.
		tn2	writing choice of tense	In the past, saving money is [tn2] a good habit.
		tn3	tense agreement	One of the ridiculous questions I met during the interviews is [tn3] that a recruiter asked two

				candidates to make comparison on themselves, and let them to decide who is[tn3] better.
		tn4	tense-related verb forms	Nowadays our standard living have improve [tn4].
pc	punctuation	pc	punctuations	First, the sewage of factories is the main source of PM2.5.[sn4] Of course we couldn't shut them down,[pc] if so our economy will be affected absolutely.[sn8]
ar	article	ar1	definite article	Some people support it, [cj3] others take the [ar1] negative point of it.[sn1]
		ar2	indefinite article	They take it as a [ar2] excuse of pursuing material comfotation [fm3].
		ar3	absence of article	To live in [ar3] easy and beautiful life.
		ar4	overuse of article	The [ar4] scientists create us an e-age, a digital a [ar4] colorful world.
cn	connection at sentence level	cn1	missing connection	Some people have no jobs. [cn1]They have received little education
		cn2	wrong connection	If we see the good thing, we want to buy it. Thus [cn2], we have the idea that we should save more money to do other things, we don't buy it any more.

		cn3	illogical reasoning	At this time, have some people no work. So [cn3] they are not good habit.
cp		cp1	all types of inconsistencies in the use of pronouns at the paragraph or discourse level	Our speed of economic development remains fast and steady, while the air pollution is so serious with plenty of factories that we should pay more attention on [vp2] it. Essentially, recently with the debate of smog, which is one of the air pollution, this topic [cp] is becoming hot.
para	paragraph	para1	wrong paragraphing Parts of texts are put into wrong paragraphs	

#### Appendix 4: An exemplar essay coded for the TSA and a diagram of different topical progression types

An exemplar essay:

“Factories Shouldn't be closed to improve air quality

1 In recent years, *air pollution* has attracted increasing attention./2 *Our environment* is polluted by tremendous industrial emissions./3 *More and more factories* have been built in China for economic development./4 Some people hold that *we* should shut down these factories for improving country's air./5 However, I can't agree with *it*./

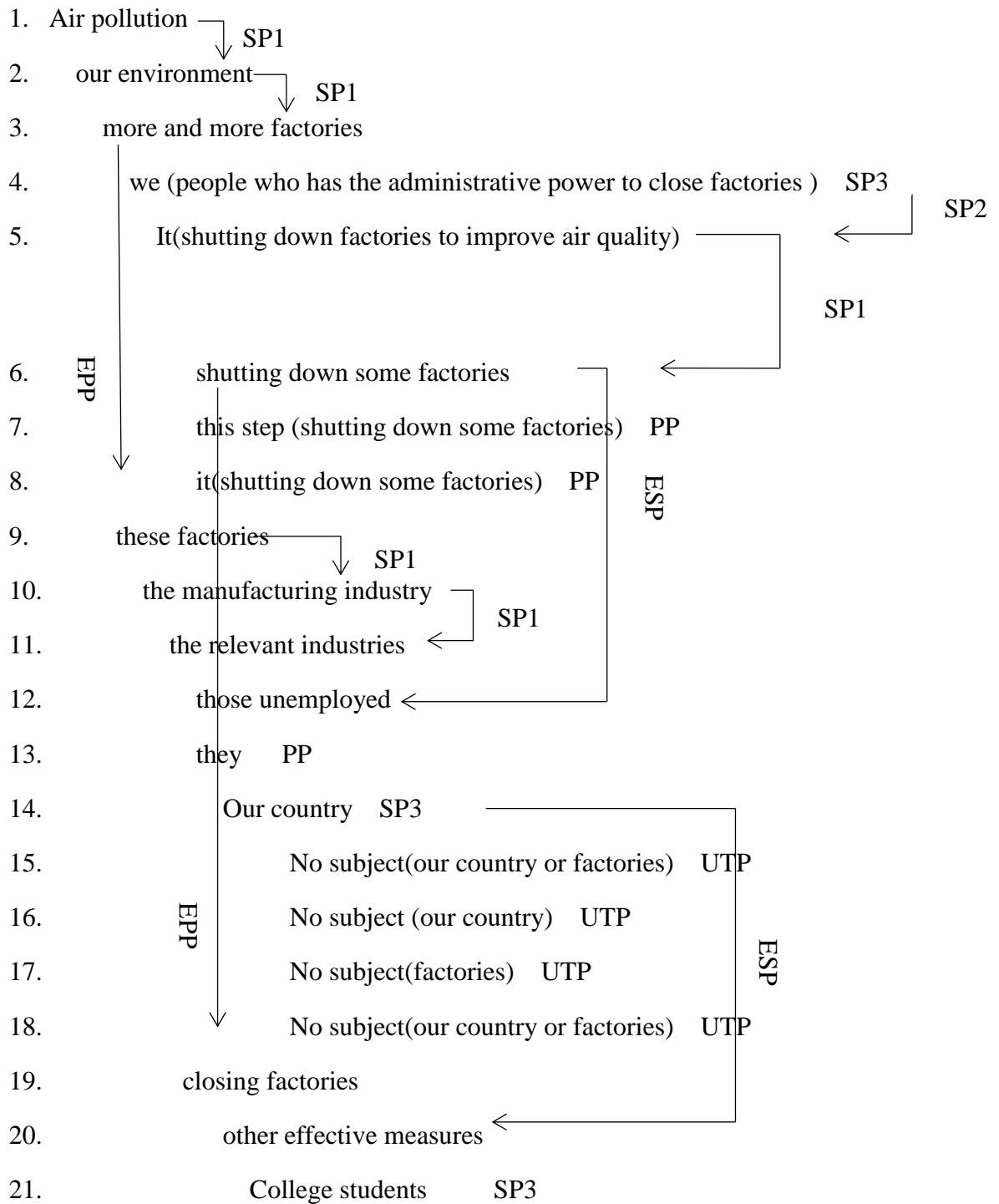
6 For one hand, *shutting down some factories* would hinder economic growth and leave many people unemployed./7 *This step* just can reduce the emission of air pollutants temporarily./8 For another, *it* also has some other terrible consequences./9 *These factories* are important parts of industrial production./10 Without them, *the manufacturing industry* cannot continue to function./11 And the *relevant industries* may also be influenced by this action./12 Besides, these factories are closed, where does *those unemployed* can go?/13 How can *they* survive in the society without jobs?/14 So in a word, *our country* need to take some measures to optimize industrial structure and develop people's awareness to protect environment./15 Try to maintain balance between human beings and those existing ecosystems./16 What's more, urge factories to attach great importance their producing process./17 Try their best to make the manufacturing process more environmentally friendly./18 Unswervingly pursue the scientific outlook on development, and build an environmental-friendly society. /

19 As far as I am concerned, *closing factories* cannot reduce the emission of air pollutants fundamentally. /20 Other *effective measures* should be adopted to solve the problem./21 And as *a college student* we should make contribution to the protection of environment./”

Notes:

Topics are in italics; “/” is used for marking T-unit; T-units are numbered for readability.

Diagram of different topical progression types:



Notes:

PP: parallel progression

EPP: extended parallel progression

SP1: sequential progression 1

SP2: sequential progression 2

SP3: discourse related progression

ESP: extended sequential progression

UTP: unrelated topical progression



**Appendix 5: Argumentation elements coding manual (adapted from a brief version of Chase's argumentative elements, 2011, p. 98–100)**

**1. Introduction (I)**

An introduction is defined as a foreshadow of what is to follow in the writer's presentation of the argument. It may outline the writer's purposes or goals.

Figure 1

For example:

With industrial factories growing up, one of air quality indices-PM 2.5 has drawn world's attention, especially China's air pollution. (I) Those factories have caused tremendous air pollutants. (I) Just like the coin has two sides, shutting down these factories can improve air, and economic growth will be hindered. (I) As far as I am concerned, if the factory can improve its ability to accomplish treatment of the pollution as scheduled, and draw support from blowndown equipments and purifiers (SP)...

**2. Conclusion (C)**

A conclusion is present when the writer gives a closing to what is written. It is located at the end of an essay, sometimes indicated by cohesive devices such as in conclusion, etc.

Figure 2

For example:

Above all, we can see that there are many ways to reduce air pollution. (C) Shutting down the factories is one of these ways. (C) But it is not the best one. (C) We may shut down some of factories, not all. (C) The most important is that we should make join effects to make air fresh and sky blue. (C)

### 3. Standpoint (SP, SN)

A standpoint is the writer's belief or opinion about a controversial topic or a proposition. It has two types: standpoint positive (SP) and standpoint negative (SN). A standpoint positive expresses an affirmative opinion about the proposition, for example, "I think factories should be closed down to improve air quality." (SP) A negative standpoint, on the other hand, expresses a negative opinion about the proposition, for example, "I don't think factories should be closed down to improve air quality." (SN) Sometimes, the writer makes more than one standpoint in an essay. The standpoints are numbered such as SP1, SP2, etc. In actual coding, when it is hard to identify whether the standpoint is positive or negative, it can be coded as ST (standpoint).

Figure 3

For example:

As we all know, recently enormous emission of air pollutants by industries has mostly caused the environmental disaster. (I) So should factories be closed down to improve air quality? (I) It is becoming a hot topic in our society. (I) Different people have different ideas towards it. (I) **In my point of view, factories should not be closed down.** (SN)  
The reasons are as follows. (FM)

### 4. Reasons (R1)

Reasons should answer the questions “why” the writer holds a certain standpoint. They are numbered such as R1, R2, R3, etc.

For example, “For one hand, shutting down some factories would hinder economic growth and leave many people unemployed.” (R1.)

#### 5. Coordinative reasons (R1a, R1b)

Coordinative reasons are multiple reasons that support a standpoint. Each reason depends on another to defend the standpoint. Reasons in coordinative relation are marked with small letters such as R1a, R1b, R1c, etc.

Figure 4

For example:

**First and foremost, the factories is playing an important role in our country.**

**(R1a) It can improve the economic growth. (R1b)** For instance, the factories can produce many little things, which are popular with foreigners. (R1.R2a) We can export them to other countries. (R1.R2b) Foreigners would like to buy them because them ususally cost a little. (R1.R2c) In this way, our country can make some money from other countries so that it can contribute to our economy. (R1.R2d)

#### 6. Subordinative reasons (R1.R1)

Subordinative reasons consist of a series of reasons where one reason represents a standpoint for the following reason. Reasons in subordinative relation are labeled as leveled such as R1 (level-1 reason), R1.R1 (level-2 reason), etc.

Figure 5

For example:

First and foremost, I think real name authentication can create a healthy and harmonious network environment, (R1) it can reduce a lot of harmful and negative informations from social network. **In this case, users' communications will be more meaningful.** (R1.R1) **In this case, users' communications will be more meaningful.** (R1.R1.R1) The second, real name authentication can prevent online fraud and reduce violations, as we know, people take freedom speech in this country, every people can talk, but sometimes words can kill people.

## 7. Convergent reasons (R1, R2)

Convergent reasons consist of more than one reason for the same standpoint. These reasons do not depend upon each other to support the standpoint. These reasons are numbered such as R1 and R2, R1.R1, R1.R2, and R1.R3.

Figure 6

For example

First and foremost, I think real name authentication can create a healthy and harmonious network environment, (R1)...

The second, real name authentication can prevent online fraud and reduce violations. (R2)...

The third, real name authentication can reduce the spread of rumors and false information, buying and selling zombie forms, and use the network to make frauds. (R3) ...

## 8. Alternative Standpoints (AS)

An alternative standpoint is directly opposed to the writer's stated standpoint. See Figure 7.

## 9. Counterarguments (CA)

A counterargument is a criticism or objection that could be used to undermine a person's standpoint. It supports the alternative viewpoint.

Figure 7

For example

In my point of view, factories should not be closed down. The reasons are as follows.

First and foremost, the factories is playing an important role in our country. It can improve the economic growth.... In addition, the factories facilize our daily life. ... Last but not least, running factories usually needs lots of people, so there are many opportunities of work... Also factories help the government a lot...

**However, there are air pollutants that caused the environmental disaster. (CA1) We also should try to deal with it. (CA2) I think the factories can do something to change it. (AS1) Maybe, they can use something safe or healthy for the air, in proccessing. (AS2) Also the government should work together to cope with it. (AS3)**

## 10. Rebuttal (RB)

A rebuttal is a statement that refutes, weakens or undermines an alternative standpoint, or counterarguments.

Figure 8

For example:

Education is essential for citizen in every country. I think no education means no civilization. Whenever and wherever, human beings put emphasis on education, especially tertiary education.

For example, in one country, more people can accept tertiary education then you will find that their country is very strong with being in harmony. At the same time, the standard living level of citizen is higher than that in other countries. Because tertiary education means faster and better in their daily life.

Someone has said that nothing is more important than to receive education. (CA) **But I think this description is not correct enough. Because tertiary education is based on education. (RB) In most of situation, the one accepting tertiary education is different from the others who just accept based education. (RB. R1) It just like the manager and the stuff. (RB.R2)**

#### 11. Reasons for rebuttal (RB.R)

Reasons that support a rebuttal. See Figure 8.

#### 12. Non-functional Unit (NF)

Non-functional elements include: repetitions (NFR), other information that does not appear to be relevant to the topic (NFI), and illegible or nonsensical information (NFU).

Figure 9

For example,

...

Also, people always want to make themselves more mysterious by choosing some cool website name and complex foreign language letter. But their real name cannot offer them this superiority. (R2) In daily life, our real name can not change easily but the website name does. (R2. R1) **I prefer website name on the internet rather than my real name.** (NFI)...

### 13. Functional Markers (FM)

A functional marker serves a particular purpose for the writer, and is often used as a transition to introduce reasons, arguments, and standpoints. See Figure 3.

### 14. Rhetorically Functional Repetitions (RFR)

Rhetorically Functional Repetitions occur when the writer restates previously expressed reasons, arguments or standpoints. These repetitions are rhetorically effective but they don't add to the breadth or depth of the argument.

Figure 10

For example,

...

From its origin, the factories is the biggest source of air pollution.(R4a) **We should solve the problem of air pollution from the root.**(RFR) **So close the factories is the most effective and rapid way.**(RFR) Only to cut it from the source, the air pollution can be managed...(R4b)

NB: Although the examples which demonstrate functional and non-functional elements in argumentative discourse are in sentences, in actual coding, a single element could be in the form of more than one sentences.

Notes:

Bold is used to highlight sample argumentative elements when other elements are also coded in the same text.

Example argumentative elements are generally presented in figures below the definition of each element and those elements which are not presented after the definition are clearly indicated by, for example, ‘see Figure 3’.



## Appendix 6: Rating Sheet: The TEM4 rating scale

Dear rater,

Thank you for taking part in the rating for this research. Your time and opinions are appreciated. Please first fill in the requested information below, then fill in your ratings in the table.

Name:

Specialty:

English teaching experience (years):

TEM4 rating experience (years):

Rating exercise	ID	Ideas and arguments	Language use	Mechanics
	1			
	2			
	3			
	4			
	5			
	6			
	7			
	8			
	9			
	10			

## Appendix 7: Rating sheet: The new rating scale

Dear teacher,

Thank you for taking part in the rating for this research. Your time and opinions are appreciated. Please first fill in the requested information below, then fill in your ratings in the table.

Name:

Argumentative writing teaching experience (years):

Argumentative writing rating experience (test names, number of times):

	ID	Mechanics	Fluency	Accuracy	Coherence	Argumentation
Exempl	ID-X	4	2	3	3	4
Trial rating	31 (ID184)					
	32 (ID517)					
Formal	ID	Mechanics	Fluency	Accuracy	Coherence	Argumentation
	1					
	2					
	3					
	4					
	5					
	6					
	7					
	8					
	9					

	10					
	11					
	12					
	13					
	14					
	15					
	16					
	17					
	18					
	19					
	20					
	21					
	22					
	23					
	24					
	25					
	26					
	27					
	28					
	29					
	30					

## **Appendix 8: Training manual**

The training manual has been developed to explain how to use a new rating scale for classroom assessment of argumentative writing of Chinese EFL college students majoring in English, in both trial and formal rating sessions. The training manual serves as a companion to the new rating scale and three coding manuals (see Documents 1, 2 and 3) and to provide details on how each scale can be applied. The new rating scale is intended to provide reliable and detailed scoring criteria for both college students and writing teachers in classroom assessment of argumentative writing. It consists of five subscales: mechanics, fluency, accuracy, coherence, and argumentation. The three coding manuals are for the textual analysis of accuracy, coherence and argumentation.

The manual comprises five sections. Each section first introduces a subscale. Then the instruction of how that scale can be used is provided, and the textual analysis involved in applying the scale is introduced. After that, exercises are provided. You need to read this manual and do the exercises before a trial rating is conducted. The following procedure is suggested for trial rating: first read the manual, the new rating scale, and the three coding manuals carefully, then complete the exercises provided at the end of each section. If you have any queries on this procedure, please do not hesitate to contact me. My contact details are: nnnnnnnnnnn (Phone), and nnnnnnnnnnn (QQ).

After you feel confident in using the new scale, please rate the two writing scripts (ID 517 and ID 184), and write your ratings (i.e., the level points on each scale) on the rating sheet. You are also encouraged to make notes and marks to support your ratings on the writing scripts where necessary. This information will be used in a discussion of differences between ratings and exercise answers in a formal meeting held later.

Please note that the writing material used for the exercises is also one of the writing scripts

used for trial rating.

### 1. Trial scale for mechanics

This scale is composed of five levels. Writing scripts at each level are described as comprising a number of paragraphs. The use of this scale involves counting the number of paragraphs in a writing script. A set of paragraph number ranges is provided, and a level point is awarded based on the range in which that writing script falls. No exercise is provided for this scale as the rating scale is straightforward.

### 2. Trial scale for fluency

This scale is composed of five levels. Writing scripts at each level are described as comprising a number of words. The use of this scale involves counting the number of words in a writing script. A set of word number ranges is provided, and a level point is awarded based on the range in which that writing script falls. No exercise is provided for this scale as the rating scale is straightforward.

### 3. Trial scale for accuracy

This scale is composed of five levels. Writing scripts at each level are characterized by the proportion of error-free sentences within all the sentences in the writing script. The use of this scale mainly involves the counting of the number of error-free sentences. The error taxonomy in Document 1 is provided for you to identify errors. You are not asked to identify error types in this rating for accuracy, however, it is encouraged that you do so in writing feedback to students.

Exercise: Please read the first paragraph of writing script ID 517 and identify and mark errors.

#### 4. Trial scale for coherence

This scale is composed of four levels. Writing scripts at each level vary in their types of topical progression. The use of the scale involves identifying different types of topical progression. Nine concepts are mentioned in the descriptors: T-unit topic, discourse topic, parallel progression (PP), extended parallel progression (EPP), related sequential progression (RSP), sequential progression II (SP2), sequential progression III (SP3), extended sequential progression (ESP), and unrelated topic progression (UTP). Among these concepts, discourse topic, unrelated topical progression, and discourse-related sequential progression are important in distinguishing between different writing proficiency levels. Other concepts are also useful as they help in understanding the analysis of coherence. More details can be found in Document 2 – “Coherence-Topical Progression Coding Manual”.

##### *Discourse topic:*

Lautamatti (1978) thinks the main idea discussed in the discourse is the discourse topic. Here is an example from van Dijk (1977).

Mr. Morgan is a careful researcher and a knowledgeable Semiticist, but his originality leaves something to be desired.

The sentence topic is Mr. Morgan. The discourse topic is Mr. Morgan’s scholarly abilities.

##### *Unrelated topical progression (UTP):*

The unrelated topical progression is a relation in which the topic of a sentence is neither clearly related to the topic nor to the comment of the preceding sentence. *It is also not clearly related to the discourse topic.* The relation does not fall into any of the relations of parallel or sequential progression (see Document 2 for more details).

Examples of unrelated topical progressions are 1) sentences with no topics (exclusive of

imperative sentence); 2) sentence topics with ambiguous references, especially pronouns; 3) illegible sentences caused by serious language errors.

#### *Discourse-related sequential progression (SP3)*

This is a relation in which the topic of a sentence is neither clearly related to the topic nor to the comment of the previous sentence (as defined in PP, SP1, SP2). The topic of a sentence is also neither related to the topic nor to the comment of the sentence that extends back over a number of sentences. However, it is related to the discourse topic.

Examples:

- 1) It is related to the discourse topic;
- 2) Pronouns such as *we*, *you*, etc. (except for unclear reference use). Only discourse-related sequential progressions which involves the use of these pronouns are used as distinguishing features.

Exercise: Please read the writing script ID 517 and identify sentences whose topics belong to discourse-related topics such as ‘we’, ‘you’, ‘our...’, ‘everyone’, etc.

### 5. Trial scale for argumentation

This scale is composed of five levels. Writing scripts at each level are characterized by different proportions of different argumentative elements within all argumentative elements in the writing script. Ten concepts are involved in the analysis for argumentation: convergent reasons, subordinate reasons, coordinate reasons, rebuttals, alternative standpoints, counterargument, irrelevant reasons, inaccurately expressed reasons, acceptable reasons, weak reasons, and illegible or nonsensical reasons. Other argumentative elements, such as introduction and conclusion, are not used as scoring criteria, although these elements are also important in analysing writing texts.

Exercise: please read writing script ID 517 and complete the following exercises:

- 1) Identify one example of coordinate reasons (R1a, R1b), convergent reasons (R1, R2), and subordinate reasons (R1.R1).
- 2) Identify one example of an unsupported level-1 reason.
- 3) Identify one example of an irrelevant reason.



## Appendix 9: Test paper 1



### Writing task 1: Real name authentication on Weibo

*Weibo has become one of the major social networks for millions of Chinese netizens, providing a platform for users to share daily life and comment on heated topics. Recently, a popular website conducted an online survey on whether Weibo websites should initiate a real name identification system. The results have shown that the 57 percent of participants argue that real name authentication (实名认证) can create a healthy and harmonious network environment, while 45 percent of participants contend that real name authentication disregards users' right of privacy and the rest choose not to respond. Should Weibo require real name authentication? Write an argumentative essay outlining your point of view.*

You have 50 minutes to plan, write and revise your essay. Write at least 300 words.

In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Try to convince your readers with relevant arguments, evidence and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference.

Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).



### Answer Sheet

(Please write on the back of the answer sheet if space is limited)

Name:	ID:	University:	Grade:
Class:	Major:	Gender:	

---



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).

## Appendix 10: Test paper 2



### Writing task 2: Higher education

*Is tertiary education (高等教育) worth going to? This has been a controversial issue for many years. Some believe it remains a good investment. Others argue that increasing unemployment rates for college graduates dissuades high school graduates from applying for college. What do you think?*

You have 50 minutes to plan, write and revise your essay. Write at least 300 words.

In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Try to convince your readers with relevant arguments, evidence and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference.

Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).



**University of  
Leicester**

**Answer Sheet**

(Please write on the back of the answer sheet if space is limited)

Name:

ID:

University:

Grade:

Class:

Major:

Gender:

---



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).

## Appendix 11: Test paper 3



### Writing task 3: Air pollution

*PM2.5, one of air quality indices, has drawn world's attention to China's air pollution. Enormous emission of air pollutants by industries has mostly caused this environmental disaster. The factories discharging air pollutants are often blamed as a major culprit of this environmental disaster. Some experts believe shutting down these factories can improve the country's air. Others believe that shutting down them would curb the economic growth and leave people unemployed. What do you think? Should factories be closed down to improve air quality?*

You have 50 minutes to plan, write and revise your essay. Write at least 300 words.

In your essay, take a position on this topic. You may write about either one of the two points of view given, or you may present a different point of view on this question. Convince your readers with reasons and examples from your knowledge and experience.

You should supply an appropriate title for your essay. You are not allowed to use a dictionary, or other relevant materials for reference.

Marks will be awarded for content, organization, grammar, strength of argumentation, and appropriateness.



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).



**University of  
Leicester**

**Answer Sheet**

(Please write on the back of the answer sheet if space is limited)

Name:

ID:

University:

Grade:

Class:

Major:

Gender:

---



Special thanks to China Scholarship Council (CSC), and Educational Testing Service (ETS).

## **Appendix 12: Student participant information sheet**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

### **Invitation**

You are being asked to take part in a research study on the development and validation of a rating scale for argumentative writing of Chinese EFL college students majoring in English. The research is a PhD project conducted by Keke Zhang under the supervision of Professor Glenn Fulcher and Dr. Jim King in the University of Leicester.

### **What will happen**

In this study, you will be given an information sheet detailing procedures, noting that your participation is voluntary, and explaining the means by which you may revoke your consent for your data to be handled. Having agreed to volunteer to participate in the study by signing the Consent Form below, you will be asked to sit an argumentative writing test, following test instructions. You are expected to write a short essay in at least 300 words and finish the writing as best you can, however no penalty will occur if you cannot finish. Your answers to the writing task will be collected, rated and coded for discoursal measures by teacher raters. Based on this data, a rating scale will be designed.

### **Time commitment**

The writing test typically takes 50 minutes.

### **Participants' rights**

You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be withdrawn. You have the right to have your questions about the procedures answered (unless answering

these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

#### Benefits and risks

Your participation in this study is voluntary. Your contribution to the study will lead to a more reliable and valid rating scale, which will benefit your self-assessment of argumentative writing. There are no known risks for you in this study.

#### Confidentiality/anonymity

The data to collect from you will include your exam composition, name, grade, gender, class, university and major. Your name, grade, gender, class, and university will be collected for identifying or indexing your composition for the convenience of data analysis, and will be stored in a separate location from the writing samples. In the presentation and publication of the research where your writing sample is utilized, every precaution will be taken to protect your anonymity. This includes using pseudonyms; real names of both individuals and universities will not be disclosed. All possible use of the data you provide will be only available to the researcher and her supervisory team under the above-mentioned conditions. Under all foreseeable conditions, all use of the data will abide by the Data Protection Act 1998.

#### For further information

Keke Zhang/Professor Glenn Fulcher will be glad to answer your questions about this study and the final results of this study at any time. You may contact her at the email address [kz37@le.ac.uk](mailto:kz37@le.ac.uk) and mobile number 07419 211 523 and him at the email address [gf39@le.ac.uk](mailto:gf39@le.ac.uk) and office number 0116 229 7508.



### **Appendix 13: Student informed consent form**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

#### **Project summary**

The project aims to develop and validate a rating scale, providing more reliable and valid standardized evaluative criteria for classroom assessment of argumentative writing of Chinese EFL college students majoring in English for the benefits of Chinese EFL university teachers as well as students. The project adopts an empirical method, coding writing samples for discoursal measures and developing a rating scale largely based on statistical analysis of coding data. The project is significant in not only providing more reliable and valid evaluative criteria, but also implementing an empirical approach to the development of rating scales in a Chinese EFL context.

By signing below, you are agreeing that: (1) you have read and understood the Participant Information Sheet, (2) questions about your participation in this study have been answered satisfactorily, and (3) you are taking part in this research study voluntarily (without coercion).

\_\_\_\_\_  
Participant's Name (Printed)\*

\_\_\_\_\_  
Participant's signature\*

\_\_\_\_\_  
Date

\_\_\_\_\_  
Name of person obtaining consent (Printed)

\_\_\_\_\_  
Signature of person obtaining consent

## **Appendix 14: Rater participant information sheet**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

### **Invitation**

You are being asked to take part in a research study on the development and validation of a rating scale for argumentative writing of Chinese EFL college students majoring in English. The research is a PhD project conducted by Keke Zhang under the supervision of Professor Glenn Fulcher and Dr. Jim King in University of Leicester.

### **What will happen**

In this study, you will be given an information sheet detailing procedures, noting that your participation is voluntary, and explaining the means by which you may revoke your consent for your data to be handled. Having agreed to volunteer to participate in the study by signing the Consent Form below, you will be asked to rate student participants' test writing samples after being duly trained. You will also be asked to responding to questionnaires following the rating sessions, answering questionnaire questions related to rating.

### **Time Commitment**

The study typically takes four hours in total across 3 sessions.

### **Participants' rights**

You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be withdrawn. You will still receive the gift card detailed below, without penalty.

You have the right to omit or refuse to answer or respond to any question that is asked of you without any penalty.

You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

#### Benefits and risks

You will get some insights into approaches towards the evaluation of students' essay writing in your teaching. And you will also be invited to training sessions on statistical analysis using SPSS involved in this study. There are no known risks for you in this study.

#### Cost, reimbursement, and compensation

Your participation in this study is voluntary. You will receive a gift card worth 50 pounds at the end of the data collection session, to compensate you for the time taken and for any expenditure (e.g., transport costs, etc.) on your part.

#### Confidentiality/anonymity

The data to collect from you will include the results of rating, response to questionnaires, name, email address, years of EFL argumentative writing teaching experience, and years of EFL argumentative writing rating experience. Your name, years of EFL argumentative writing teaching experience, and years of EFL argumentative writing rating experience will be collected for identifying or indexing the rating results and questionnaire responses for the convenience of data analysis. Your email will be only used for the convenience of contact for the purposes of the research. In the presentation and publication of the research where your rating and questionnaire data are utilized, every precaution will be taken to protect your anonymity. This includes using pseudonyms; real names of individuals and

universities will not be disclosed. All possible use of the data you provide will be only available to the researcher and her supervisory team under the above-mentioned conditions. Under all foreseeable conditions, all use of the data will abide by the Data Protection Act 1998.

For further information

Keke Zhang/Professor Glenn Fulcher will be glad to answer your questions about this study and the final results of this study at any time. You may contact her at email: [kz37@le.ac.uk](mailto:kz37@le.ac.uk)/mobile: 07419 211 523 and him at email: [gf39@le.ac.uk](mailto:gf39@le.ac.uk)/office: 0116 229 7508.

## **Appendix 15: Rater informed consent form**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

### **Project summary**

The project aims to develop and validate a rating scale, providing more reliable and valid standardized evaluative criteria for classroom assessment of argumentative writing of Chinese EFL college students majoring in English, for the benefit of Chinese EFL university teachers as well as students. The project adopts an empirical method, coding writing samples for discoursal measures and developing a rating scale largely based on statistical analysis of coding data. The project is significant in not only providing more reliable and valid evaluative criteria, but also implementing empirical approaches to the development of rating scales in a Chinese EFL context.

By signing below, you are agreeing that: (1) you have read and understood the Participant Information Sheet, (2) questions about your participation in this study have been answered satisfactorily, and (3) you are taking part in this research study voluntarily (without coercion).

\_\_\_\_\_  
Participant's Name (Printed)\*

\_\_\_\_\_  
Participant's signature\*

\_\_\_\_\_  
Date

\_\_\_\_\_  
Name of person obtaining consent (Printed)

\_\_\_\_\_  
Signature of person obtaining consent

## **Appendix 16: Teacher participant information sheet**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

### **Invitation**

You are being asked to take part in a research study on the development and validation of a rating scale for classroom assessment of argumentative writing of Chinese EFL college students majoring in English. The research is a PhD project conducted by Keke Zhang under the supervision of Professor Glenn Fulcher and Dr. Jim King in University of Leicester.

### **What will happen**

In this study, you will be given an information sheet detailing procedures, noting that your participation is voluntary, and explaining the means by which you may revoke your consent for your data to be handled. Having agreed to volunteer to participate in the study by signing the Consent Form below, you will be asked to respond to a questionnaire on appropriate parameters of argumentative writing tasks for Chinese EFL college students majoring in English. You will also be asked to double code a small sample of writing scripts for discourse measures anonymously after being duly trained.

### **Time commitment**

The study typically takes three hours across 2 sessions.

### **Participants' rights**

You may decide to stop being a part of the research study at any time without explanation. You have the right to ask that any data you have supplied to that point be withdrawn. You will still receive the gift card detailed below, without penalty.

You have the right to omit or refuse to answer or respond to any question that is asked of you without any penalty.

You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

#### Benefits and risks

You will get some insights into approaches towards the evaluation of students' essay writing in your teaching. And you will also be invited to training sessions on statistical analysis using SPSS involved in this study. There are no known risks for you in this study.

#### Cost, reimbursement and compensation

Your participation in this study is voluntary. You will receive a gift card worth 40 pounds at the end of the data collection session, to compensate you for the time taken and for any expenditure (e.g., transport costs, etc.) on your part.

#### Confidentiality/anonymity

The data to collect from you will include response to questionnaires, coding of student participants' writing scripts,, name, gender, email address and years of argumentative writing teaching experience, and years of EFL argumentative writing rating experience. Your name, gender, and years of argumentative writing teaching experience, and years of EFL argumentative writing rating experience will be collected for identifying or indexing the questionnaire response and coding results for the convenience of data analysis. Your email will be only used for the convenience of contact for the purposes of the research. In the presentation and publication of the research where your questionnaire and coding

data are utilized, every precaution will be taken to protect your anonymity. This includes using pseudonyms; real names of individuals and universities will not be disclosed. All possible use of the data you provide will be only available to the researcher and her supervisory team under the above-mentioned conditions. Under all foreseeable conditions, all use of the data will abide by the Data Protection Act 1998.

For further information

Keke Zhang/Professor Glenn Fulcher will be glad to answer your questions about this study and the final results of this study at any time. You may contact her at email:

[kz37@le.ac.uk](mailto:kz37@le.ac.uk)/mobile: 07419 211 523 and him at email: [gf39@le.ac.uk](mailto:gf39@le.ac.uk)/office: 0116 229 7508.



## **Appendix 17: Teacher informed consent form**

Project Title: Developing a Rating Scale for Classroom Assessment of the Argumentative Writing of Chinese EFL College Students Majoring in English

### **Project summary**

The project aims to develop and validate a rating scale, providing more reliable and valid standardized evaluative criteria for classroom assessment of argumentative writing of Chinese EFL college students majoring in English, for the benefit of Chinese EFL university teachers as well as students. The project adopts an empirical method, coding writing samples for discoursal measures and developing a rating scale largely based on statistical analysis of coding data. The project is significant in not only providing more reliable and valid evaluative criteria, but also implementing empirical approaches to the development of rating scales in a Chinese EFL context.

By signing below, you are agreeing that: (1) you have read and understood the Participant Information Sheet, (2) questions about your participation in this study have been answered satisfactorily, and (3) you are taking part in this research study voluntarily (without coercion).

\_\_\_\_\_  
Participant's Name (Printed)\*

\_\_\_\_\_  
Participant's signature\*

\_\_\_\_\_  
Date

\_\_\_\_\_  
Name of person obtaining consent (Printed)

\_\_\_\_\_  
Signature of person obtaining consent

## Appendix 18: Codes and calculations

Code	Calculation
CN/C	complex nominals per clause
CN/T	complex nominals per T-unit
CP/C	coordinate phrases per clause
CP/T	coordinate phrases per T-unit
C/T	clauses per T-unit
E/C	errors per clause
EFC/C	error-free clauses per clause
EFT/T	error-free T-units per T-unit
E/T	errors per T-unit
LW/W	lexical words per word
MLS	mean length of sentence
MLT	mean length of T-unit
MLC	mean length of clause
SWT/WT	sophisticated word types per word type
WT/W	word types per word

## Bibliography

- Alderson, C. (1991). Bands and scores. In C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp.71-86). London: Modern English Publications/British Council/Macmillan.
- Alderson, C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Anthony, M. & Gladkov, K. (2007). Rhetorical appeals in fundraising. In D. Biber, U. Connor, & T. A. Upton (Eds), *Discourse on the Move: Using corpus analysis to describe discourse structure* (pp. 121–51). John Benjamins, Amsterdam.
- Aristotle. (1932). *The Rhetoric of Aristotle* (L. D. Cooper, Trans.). New York NY: Appleton and Company.
- Aristotle. (1984). Rhetoric. In J. Barnes (Ed.), *The Complete Works of Aristotle* (rev. Oxford ed., Vol. 2, pp. 2152–269). Princeton NJ: Princeton University Press.
- Bachman, L. F. & Palmer, A. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16(4), 449–64.
- Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal*, 70(4) 380-90.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal*, 70(4), 380-90.
- Banerjee, J., & Franceshina, F. (2006). *Documenting features of written language production typical at different IELTS band score levels*. Paper presented at the ESF SCSS Exploratory Workshop 'Bridging the gap between research on second language

- acquisition and research on language testing' Amsterdam, The Netherlands. February, 2006.
- Basturkmen, H., & Randow, J. von. (2014). Guiding the reader (or not) to re-create coherence: Observations on postgraduate student writing in an academic argumentative writing task. *Journal of English for Academic Purposes*, 16, 14–22.
- Beauvais, P. J. (1989). A speech act theory of metadiscourse. *Written communication*, 6(1), 11–30.
- Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28–41.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education ESL. .
- Brindley, G. (1991). Defining language ability: the criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing*, Singapore, Anthology Series 25, SEAMEO/RELC.
- Brown, J. D. (2005). *Testing in language programs: a comprehensive guide to English language assessment*. McGraw-Hill College.
- Burneikaite, N., & Zabaliute, J. (2003). Information structuring in learner texts: A possible relationship between the topical structure and the holistic evaluation of learner essays. *Studies about Language*, 4, 1–11.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1), 1–47.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. In J.C. Richards & R. Schmidt (Eds.), *Language and communication* (pp.2–27). London, UK: Longman.
- Carr, N. T. (2011). *Designing and analyzing language tests: Oxford handbooks for language teachers*. Oxford, UK: Oxford University Press.
- Cerbin, B. (1988). *The Nature and Development of Informal Reasoning Skills in College*

- Students*. Paper presented at the National Institute on Issues in Teaching and Learning Chicago, IL, April, 1988. Retrieved from <https://eric.ed.gov/?id=ED298805>
- Champney, H. (1941). The measurement of parent behavior. In *Child Development*, 12(2), 131–66.
- Chase, B. J. (2011). *An Analysis of the Argumentative Writing Skills of Academically Underprepared College Students*. PhD thesis, Columbia University, USA. Retrieved 4 May 2014, from [http://academiccommons.columbia.edu/download/fedora\\_content/download/ac:131454/CONTENT/Chase\\_columbia\\_0054D\\_10083.pdf](http://academiccommons.columbia.edu/download/fedora_content/download/ac:131454/CONTENT/Chase_columbia_0054D_10083.pdf)
- Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher and F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 35-47). Routledge.
- Cheng, X., & Stephenson, M. S. (1996). Metadiscourse: A technique for improving student writing. *Research in the Teaching of English*, 30(2), 149–81.
- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in Writing: Generating Text in L1 and L2. *Written Communication*, 18(1), 80–98.
- Chinese University Foreign Language Corpus Development and Research-the Corpus for English Majors (CEM) (2008). Shanghai: Shanghai Foreign Language Education Press.
- Cobb, T. (2002). Web Vocabprofile. Retrieved from <http://www.lex tutor.ca/vp/>
- Coffin, C. (2004). Arguing about how the world is or how the world should be: the role of argument in IELTS Test. *Journal of English for Academic Purposes*, 3(3), 229–46.
- Coffin, C., Hewings, A., & North, S. (2012). Arguing as an academic purpose: The role of asynchronous conferencing in supporting argumentative dialogue in school and university. *Journal of English for Academic Purposes*, 11(1), 38–51.
- Cohen, A. (1994). *Assessing language ability in the classroom*. 2<sup>nd</sup> edition. Rowley Mass. Newbury House/Heinle and Heinle.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Newbury Park, CA: Sage.
- Connor, U. (1990). Linguistic/Rhetorical Measures for International Persuasive Student

- Writing. *Research in the Teaching of English*, 24(1), 67–87.
- Connor, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second language writing*. Cambridge: Cambridge University Press.
- Connor, U., & Farmer, M. (1990). The teaching of topical structure analysis as a revision strategy for ESL writers. *Second language writing: Research insights for the classroom*, 126-139.
- Connor, U. & Lauer, J. (1985). Understanding persuasive essay writing: Linguistic/Rhetorical approach. *Text*, 5(4), 309–26.
- Connor, U. & Lauer, J. (1988). Cross-cultural variation in persuasive student writing. In A.C. Purves (Ed.) *Writing Across Languages and Cultures* (pp. 138-159). Newbury Park, CA: Sage.
- Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics*, 22, 263–78.
- Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, 78(1), 98-104.
- Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, 34(2), 213–38.
- Crammond, J. G. (1997). An analysis of argument structure in expert and student persuasive writing. Unpublished doctoral dissertation, McGill University. Retrieved 4 June 2014, from [http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object\\_id=37709&local\\_base=GEN01-MCG02](http://digitool.library.mcgill.ca/R/?func=dbin-jump-full&object_id=37709&local_base=GEN01-MCG02)
- Crammond, J. G. (1998). The uses and complexity of argument structures in expert and student persuasive writing. *Written Communication*, 15(2), 230–68.
- Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written communication*, 10(1), 39–71.
- Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework*. TOEFL Monograph Series 22. Princeton, New

- Jersey: Educational Testing Service.
- Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, 86(1), 67–96.
- Dafouz-Milne, E. (2008). The pragmatic role of textual and interpersonal markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse. *Journal of Pragmatics*, 40, 95–113.
- Davies, A. (1990). *Principles of language testing*. Oxford, Blackwell.
- Davidson, F. (1992). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Eds). *Assessing second language writing in academic contexts*. Norwood N. J., Ablex.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press. Retrieved from: [http://works.bepress.com/annie\\_brown/11/](http://works.bepress.com/annie_brown/11/)
- Ellis, R., & Barkhuizen, G. (2005). *Analyzing learner language*. Oxford: Oxford University Press.
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4, 139–56.
- Ferretti, R. P., Lewis, W. E. & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' argumentative writing strategies? *Journal of Educational Psychology*, 101(3), 577–89.
- Ferretti, R. P., MacArthur, C. A. & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology*, 92(4), 694–702.
- Foreign Language Teaching Advisory Committee of Higher Education English Group (2000). *English Syllabus for English Majors of Higher Education*, Foreign Language Teaching and Research Press, Beijing; Shanghai Foreign Language Education Press, Shanghai.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299–323.

- Fulcher, G. (1987). Tests of oral performance: the need for data-based criteria. *English Language Teaching Journal*, 41(4): 287–91.
- Fulcher, G. (1989). *Lexis and Reality in Oral Testing*. Washington DC: ERIC Clearinghouse for Languages and Linguistics, ERIC\_NO: ED298759.
- Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. PhD thesis, University of Lancaster, UK. Retrieved from [info/articles/store/FulcherPhD.pdf](#)
- Fulcher, G. (1995). Variable competence in second language acquisition: A problem for research methodology? *System*, 23(1): 25–33.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208–38.
- Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman/Pearson.
- Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, 28(1), 5–29.
- Fulcher, G. (2017). Criteria for Evaluating Language Quality. *Language Testing & Assessment*, 179-192.
- Gaies, S. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly*, 14, 53–60.
- Govier, T. (2013). *A practical study of argument*. Cengage Learning.
- Grabe, W., & Kaplan, R. B. (1996). *Theory & practice of writing: An Applied Linguistic Perspective*. London and New York: Longman.
- Green, R. (2013). *Statistical Analysis for Language Testers*. Palgrave Macmillan.
- Gui, Sh. Ch. & Yang, H. Zh. (2003). *Chinese Learner English Corpus*. Shanghai: Shanghai Foreign Language Education Press.
- Gyllstad, H. (2007). *Testing English Collocations – Developing Receptive Tests for Use with Advanced Swedish Learners*. Lund: Lund University, Media-Tryck.
- Halliday, M. A. K., McIntosh, A. & Stevens, P. (1964). *The Linguistic Sciences and Language Teaching*. Bloomington, Indiana: Indiana University Press.



- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M., Matthiessen, C. M., & Matthiessen, C. (1994/2014). *An Introduction to Functional Grammar*. Routledge.
- Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In Kroll, B. (Ed.) *Second language writing* (pp. 69–87). New York, Cambridge University Press.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts* (pp. 241–76). Norwood N.J., Ablex.
- Hasan, R. (1984). *Coherence and cohesive harmony*. In J. Flood (Ed.), *Understanding reading comprehension* (pp.139–54). Newark, DE: International Reading Association.
- Herriman, J. (2011). Themes and theme progression in Swedish advanced learners' writing in English. *Nordic Journal of English Studies*, 10(1), 1–28.
- Hieke, A. E. (1985). A componential approach to oral fluency evaluation. *Modern Language Journal*, 69(2), 135-42.
- Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan*, 2, 21–30.
- Hirvela, A. (2017). Argumentation & second language writing: Are we missing the boat? *Journal of Second Language Writing*, 36, 69–74.
- Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press.
- Hoey, M. (1994). Signalling in discourse: a functional analysis of a common discourse pattern in written and spoken English. In M. Coulthard (Ed.), *Advances in written text analysis* (pp. 26–45). London: Routledge.
- Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly*, 14, 35–43.
- Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics*, 19(1), 24–44.
- Hunt, K.W. (1970). Syntactic maturity in school children and adults. *Monographs of the*

- Society for Research in Child Development*, 35(1), 1-67.
- Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of second language writing*, 6(2), 183-205.
- Hyland, K. (1998). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of pragmatics*, 30(4), 437-55.
- Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. John Wiley & Sons, Inc.
- Hymes, D. H. (1967). Models of the interaction of language and social setting. *Journal of Social Issues*, 23(2), 8-38.
- Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.): *Sociolinguistics* (pp. 269-93). Harmondsworth: Penguin.
- Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics*, 25, 205-27.
- Hughes, W. & Lavery, J. (2008). *Critical thinking: An introduction to the basic skills (5th edition)*, Peterborough, Ontario: Broadview Press.
- Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. *NCTE Research Report* No. 3. Retrieved from <https://eric.ed.gov/?id=ED113735>
- IBM Corp. Released 2013. *IBM SPSS Statistics for Windows*, Version 22.0. Armonk, NY: IBM Corp.
- Intaraprawat, P., & Steffensen. M. S. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, 4(3), 253-72.
- Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. and Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- Jafapur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System*, 19(4), 459-65.
- Jin, Y., & Fan, J. (2011). Test for English majors (TEM) in China. *Language Testing*, 28(4), 589-596.
- Jones, R. L. (1981). Scoring procedures in oral language proficiency tests. In Read, J.A.S. (Ed.), *Directions in Language Testing (Anthology Series 9)* (pp. 100-7). RELC:

Singapore University Press.

Kennedy, C., & Thorp, D. (2002). *A corpus-based investigation of linguistic responses to an IELTS academic writing task*: University of Birmingham.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Knoch, U. (2007). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, 12(2), 108–28.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Frankfurt: Peter Lang.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155-163. Koponen, M. & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency*, Ann Arbor, Michigan: University of Michigan Press Robinson.

Kuhn, D. (1991). *The skills of argument*. Cambridge, UK: Cambridge University Press.

Kummer, J. L. (1972) Aspects of a theory of argumentation. In E. Gulich & W. Raible (Eds.). *Textsorten* (pp.25–49). Frankfurt am Main: Athaneum.

Lado, R. (1961). *Language testing*. New York: McGraw-Hill.

Lantolf, J. & Frawley, W. (1985). Oral proficiency testing: a critical analysis. *Modern Language Journal*, 69, 337–45.

Larsen-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, 25, 21–33.

Lautamatti, L. (1978). Some Observations on Cohesion and Coherence in Simplified Texts. Eric. Retrieved from <https://eric.ed.gov/?id=ED275191>

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

- Leech, G., & Svartvik, J. (2013). *A communicative grammar of English*. Routledge.
- Lee, Y., Gentile, C. & Kantor, R. (2008). Analytic scoring of TOEFL CBT Essays: Scores from humans and E-rater (TOEFL Research Report RR-08-01). Princeton NJ: ETS.
- Lennon, P. (1991). Error: Some problems of definition, identification and distinction. *Applied Linguistics*, 12, 180–196.
- Liang, M. & Xiong, W. (2008). Wenben fenxi gongju PatCount zai waiyu jiaoxue yu yanjiu zhong de yingong [Patcount and its application in foreign languages teaching and research]. *Computer-assisted Foreign Language Education*, 5, 71-76.
- Li, Q. H. (2010). *The development & validation of the rating scale for TEM 4 writing section*. Unpublished postdoctoral report, Shanghai International Studies University, Shanghai, China.
- Li, Q. H. (2014). TEM4 Xiezuofenxiangshi pingfenbiaozhun yu zhengtishi pingfenbiaozhun duibi yanjiu [A comparative study of analytic scoring and holistic scoring for TEM4 writing subtest]. *Foreign Language Testing and Teaching*, 3, 11-20.
- Linacre, J. M. (1989). *Multi-faceted measurement*. Chicago: MESA Press.
- Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Sweden: Gleerup.
- Liu, M., & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System*, 33(4), 623–36.
- Liu, F., & Stapleton, P. (2014). Counterargumentation and the cultivation of critical thinking in argumentative writing: Investigating washback from a high-stakes test. *System*, 45, 117–28.
- Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper & L. Odell (Eds.), *Evaluating Writing* (p.33-69). Urbana, IL: National Council of Teachers of English.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–96.
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers's language development. *TESOL Quarterly*, 45(1), 36–62.

- Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96(2), 190–208.
- Lyons, J. (1977). *Semantics (Vols I & II)*. Cambridge CUP.
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave MacMillan.
- Marttunen, M. (1994). Assessing argumentation skills among Finnish university students. *Learning and Instruction*, 4(2), 175–91.
- Matthews, M. (1990). The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *English Language Teaching Journal*, 44(2), 117–21.
- McCann, T. M. (1989). Student Argumentative Writing Knowledge and Ability at Three Grade Levels. *Research in the Teaching of English*, 23(1), 62–76.
- McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.
- Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, 14, 139–78.
- Miller, K. S. (2000). Academic Writers On-Line: Investigating Pausing in the Production of Text. *Language Teaching Research*, 4(2), 123–48.
- Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters.
- McGavigan, P. (2009). The acquisition of fixed idioms in Greek learners of English as a foreign language. Unpublished doctoral dissertation, Swansea University, UK.
- Mislevy, R. J. (1995). Foundations of a new test theory. In N. Frederiksen R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–40). Hillsdale N. J.: Lawrence Erlbaum Associates.
- Montee, M. & Malone, M. E. (2014). Writing scoring criteria and score reports. In A. Kunnan (Ed.), *The companion to language assessment, Vol. 2* (pp. 847–59). Oxford: Wiley-

Blackwell.

Mugharbil, H. (1999). *Second language learners' punctuation: Acquisition and awareness*.

Unpublished PhD dissertation, University of Southern California.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, 4(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement*, 5(2), 189–227.

Nation's Discipline Evaluation University Ranking of Foreign Languages and Literatures (2012). Retrieved from <http://edu.people.com.cn/n/2013/0129/c1053-20361429.html>

Neuner, J. L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English*, 21(1), 92–105.

Newman, R. (1977). *An attempt to define through error analysis the intermediate ESL level at UCLA*. Unpublished thesis, University of California at Los Angeles.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, 23(4), 445–65.

North, B. (1996/2000). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. PhD thesis, Thames Valley University/New York: Peter Lang.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, Peter Lang.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph 24. Princeton: Educational Testing Service.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, 15(2), 217–63.

Nussbaum, E. M., Kardash, C. M., & Graham, S. E. (2005). The Effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, 97(2), 157.

- O'Keefe, D. J. (1999). How to handle opposing arguments in persuasive messages: a meta-analytic review of the effects of one-sided and two-sided messages. In M. E. Roloff (Ed.), *Communication yearbook* (Vol. 22, pp. 209–49). Thousand Oaks, CA: Sage.
- Oller Jr, J. W. (1983). A consensus for the eighties. *Issues in language testing research*, 351–6.
- 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.
- Osgood, C.E., and Tannenbaum, P. (1957). *The measurement of measuring*. University of Illinois Press.
- Paterson, D. G. (1922). The Scott Company graphic rating scale. *Journal of Personnel Research*, 1, 361-376.
- Paul, D. (2013). Covering the construct: An approach to Automated Essay Scoring motivated by a Socio-cognitive framework for defining literacy skills. *Handbook of automated essay evaluation: Current applications and new directions* (pp. 298–312). Routledge. New York.
- Pennington, M., & So, S. (1993). Comparing writing process and product across two languages: a study of 6 Singaporean University student writers. *Journal of Second Language Writing*, 2, 41–63.
- Perelman, C. (1982). *The Realm of Rhetoric* (W. Kluback, Trans.). Notre Dame IN: University of Notre Dame Press.
- Perkins, D. N. (1985). Post primary education has little impact on informal reasoning. *Journal of Educational Psychology*, 77, 562–71.
- Pigaiwang. (2019). <https://www.pigai.org/>
- Pollitt, A. (1991). Giving students a sporting chance. In J.C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 46-59). Modern English Publications/British Council, London, Macmillan.
- Polio. C.G. (1997). Measures of linguistic accuracy in second language writing research. *Language learning*, 47(1), 101–43.
- Polio, C. (2001). Research methodology in second language writing research: The case of text-

- based studies. In T. Silva & P. K. Matsuda (Eds.), *On second language writing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26, 10–27.
- Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an Acquisition-Based Procedure for Second Language Assessment. *Studies in Second Language Acquisition*, 10(2), 217-243. doi:10.1017/S0272263100007324
- Qian Y., Andrés Ramírez, J. & Harman R. (2007). EFL Chinese students and high stakes expository writing: A Theme analysis. *Colombian Applied Linguistics Journal*, 9, 99–125.
- Qin, J.J. & Karabacak, E. (2010). The analysis of Toulmin elements in Chinese EFL university argumentative writing. *System*, 38(3), 444–56.
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
- Roid, G. H. (1994) Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education*, 7(2), 159–70
- Ricci, V. (2005). Fitting distributions with R. R project web site Retrieved from <http://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf>
- Revision Group of Syllabus for Test for English Majors-Band 4 of Higher Education (2004). *Syllabus for English Majors-Band 4 (2004 version)*. Shanghai: Shanghai Foreign Language Education Press.
- Rusfandi. (2015). Argument-Counterargument Structure in Indonesian EFL Learners' English Argumentative Essays: A Dialogic Concept of Writing. *RELC Journal*, 46(2), 181–97.
- Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19<sup>th</sup> Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press.
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future



- directions. *Science Education*, 92(3), 447–72.
- Schlepppegrell, M. J. (1996). Conjunction in spoken English and ESL writing. *Applied Linguistics*, 17(3), 271–85.
- Schmidt, R. (1992). Psychological Mechanisms Underlying Second Language Fluency. *Studies in Second Language Acquisition*, 14(4), 357–85.
- Schneider, M. & Connor, U. (1990). Analyzing topical structure in ESL essays: Not all topics are equal. *Studies in Second Language Acquisition*, 12, 411–27.
- Schultz, R. A. (1986). From achievement through proficiency through classroom instruction: some caveats. *Modern Language Journal*, 70(4), 373–9.
- Schwarz, B. B., Neuman, Y., Gil, J., & Ilya, M. (2003). Construction of collective and individual knowledge in argumentative activity. *The Journal of the Learning Sciences*, 12(2), 219–56.
- Seliger, H. W., & Shohamy, E. (1989). *Second Language research methods*. Oxford: Oxford University Press.
- Shaw, P., & Liu, T.-K. (1998). What Develops in the Development of Second-language Writing? *Applied Linguistics*, 19(2), 225–54.
- Shaw, S.D. (2004). Automated writing assessment: A review of four conceptual models. *Cambridge research notes*, 17, 13-18.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420-428.
- Simon, S. (2008). Using Toulmin’s argument pattern in the evaluation of argumentation in school science. *International Journal of Research & Method in Education*, 31(3), 277–89.
- Simpson, J.M. (2000). Topical structure analysis of academic paragraphs in English and Spanish. *Journal of Second Language Writing*, 9(3), 293–309.
- Skehan, P. (1998) *A cognitive approach to language learning*. Oxford: Oxford University Press.
- Skehan, P. (2003). Task-based instruction. *Language Teaching*, 36, 1–14.

- Smith, P. C., & Kendall, J. M. (1963). Retranslation of expectations: an approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149–54.
- Soleymanzadeh, L., & Gholami, J. (2014). Scoring Argumentative Essays based on Thematic Progression Patterns and IELTS Analytic Scoring Criteria. *Procedia - Social and Behavioral Sciences*, 98 (Supplement C), 1811–819.
- Stapleton, P., & Wu, Y. A. (2015). Assessing the quality of arguments in students' persuasive writing: A case study analyzing the relationship between surface structure and substance. *Journal of English for Academic Purposes*, 17, 12–23.
- Swales, J. (1990). *Genre Analysis: English for Academic and Research Settings*. Cambridge: CUP.
- Tan, H., & Eng, W. B. (2014). Metadiscourse Use in the Persuasive Writing of Malaysian Undergraduate Students. *English Language Teaching*, 7(7), 26–39.
- Nation's Discipline Evaluation University Ranking of Foreign Languages and Literatures (2012) Retrieved from <http://edu.people.com.cn/n/2013/0129/c1053-20361429.html>
- Thorndike, E. L. (1904/1912). *An introduction to the theory of mental and social measurements*. New York, Teachers College Columbia University.
- Thordardottir, E., & Ellis Weismer, S. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language and Communication Disorders*, 36, 221–44.
- Toulmin, S. E. (1958/2003). *The uses of arguments* (updated Ed.). Cambridge: Cambridge University Press.
- Toulmin, S. E., Rieke, R. D., & Janik, A. (1979). *Introduction to reasoning*. New York: Macmillan.
- Tsang, W. K. (1996). Comparing the effects of reading and writing on writing performance. *Applied Linguistics*, 17, 210–33.
- Turner, C. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language*

- Review*, 56(4), 555–84.
- Turner, C. E. (2012). Classroom assessment. In G. Fulcher and F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 79-92). Routledge.
- Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, 36(1), 49–70.
- Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*, 49(1), 3–12.
- Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College composition and communication*, 36, 82–93.
- van Dijk, T. A. (1982). Episodes as units of discourse analysis. In D. Tannen (Ed.), *Analyzing discourse: Text and talk*, 177–95. Washington, D. C.: Georgetown University Press.
- van Eemeren, F.H., Grootendorst, R., & Henkemans, F.S. (2002). *Argumentation: Analysis, evaluation, and presentation*. Mahwah, NJ: Erlbaum.
- Wang, H. (2016). Implications of “Statement of the adjustments of items of Test for English Majors-Band 4 (TEM4)” on Teaching of English Majors, *English on Campus*, 8, 30-31.
- Watson Todd, R., Khongput, S., & Darasawang, P. (2007). Coherence, cohesion and comments on students’ academic essays. *Assessing Writing*, 12(1), 10–25.
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
- White, E. (1993). Holistic scoring: Past triumphs, future challenges. In: Williamson M. M., & Huot, B. A. (Eds). *Validating holistic scoring for writing assessment* (pp. 79-108). Cresskill. N.J., Hampton Press.
- Witte, S. P. & Faigley, L. (1981). Coherence, cohesion and writing quality. *College Composition and Communication*, 22(1), 189–204.
- Witte, S. P. (1983). Topical structure and revision: An exploratory study. *College composition and communication*, 34 (3), 313–41.
- Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. *Thinking & Reasoning*, 14(1), 1–27.

- Wolfe, C. R. (2012). Individual differences in the “myside bias” in reasoning and written argumentation. *Written Communication*, 29, 477–501.
- Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication*, 25, 183–209.
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (No. 17). University of Hawaii Press.
- Xue, G., & Nation, I. S. P. (1984). A university word list. *Language learning and communication*, 3(2), 215–29.
- Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, 23(1), 31–48.
- Young, R. (1995). Conversational styles in language proficiency interviews. *Language learning*, 45(1): 3–42.
- Zhang, M. S. (2000). Cohesive Features in the Expository Writing of Undergraduates in Two Chinese Universities. Retrieved from <http://journals.sagepub.com/doi/abs/10.1177/003368820003100104>
- Zhang, L. J. (2010). A dynamic metacognitive systems account of Chinese university students’ knowledge about EFL reading. *TESOL Quarterly*, 44(2), 320–53.
- Zou, Sh. (1997). *The Test for English Majors (TEM) Validation Study*. Shanghai, Shanghai Foreign Language Education Press.
- Zou, Sh. (2011). *Construction and research on Corpus for English Majors*. Fudan, China: Fudan Press.
- Zou, Sh. (2017). *Language Assessment* (2<sup>nd</sup> Edition). Shanghai, Shanghai Foreign Language Education Press.