Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
THE REACTIVITY AND VERIDICALITY
OF CONCURRENT VERBAL REPORTS
IN SECOND LANGUAGE WRITING

CHENGSONG YANG

A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY IN EDUCATION
THE UNIVERSITY OF AUCKLAND, 2016
ABSTRACT

Much concern has been raised over the issue of whether administering concurrent verbal reports (CVRs) alters the very cognitive processes that the reports are supposed to represent and keep intact (the reactivity issue) and whether they reflect the processes completely and accurately (the veridicality issue). This thesis reports on an empirical study designed to address both issues. It examines the reactive effects of concurrent verbal reporting (CVR) or thinking aloud (TA) in a second language (L2) narrative writing task, in particular, the mediation of L2 writing proficiency, working memory capacity (WMC), as well as task type, in such effects. Eight-five Chinese sophomore participants whose awareness of lexical diversity in writing in English had been raised through instruction engaged in two WMC tests, then wrote a baseline outline-given narrative silently, and then completed a main writing task under either a no-think-aloud (NTA) condition or a think-aloud (TA) condition. The TA group were also asked to provide immediate retrospective verbal reports (RVRs) on their revisions, in addition to CVRs, and answer questions which elicited their perceived reactivity and veridicality. Participants’ compositions were analysed, in terms of 20 measures of fluency, complexity, accuracy, organization, and content, to find out group differences and interactional effects. The results showed that TA caused reactivity because it significantly increased dysfluencies, worsened organization, and most noticeably, impaired lexical diversity. Reactivity was further found to be significantly different among L2 learner writers at different baseline levels of Speed Two (i.e., total number of words produced per minute), General Complexity, Correct Verb Use, or Length Two (i.e., total number of words produced) on the corresponding measure. Reactivity was also found to distinguish different WMC Average or WMC One (i.e., the operation span)
learner writers on Organization, and differentiate different WMC Two (i.e., the reading span) learner writers on Length Three (i.e., total number of syllables in the final composition), Length Four (i.e., total number of syllables produced), and Lexical Diversity. The comparison of CVRs and RVRs indicated that participants ignored reporting transitory metacognitive thinking and supposedly automatized behaviours, among others, making their CVRs incomplete. Their perceived reactivity appeared to be more serious in terms of the intensity and range of effects they felt. Their perceptions on incompleteness corroborated the findings from the comparison of CVRs and RVRs, but they stood by the accuracy of their CVRs. Kellogg’s (1996) model of writing processes in relation to Baddeley’s (1986) accounts on working memory (WM), situated within the information-processing framework, on which Ericsson and Simon (1993) based their argument for the validity of CVRs, was used to explain the reactivity and veridicality problems found in this project.
ACKNOWLEDGEMENTS

I am sincerely grateful to Professor Lawrence Jun Zhang, first and foremost, for his consistent, ready academic support and earnest, useful advice as my main supervisor. I would also like to thank him, for the academic platform he has contributed a lot to establishing here, including the academic lectures by world-renowned scholars he has invited, the atmosphere of research he has been developing by taking in a cohort of equally hardworking PhD students and visiting scholars, and the opportunities he gave us to review articles from some of the important international/China domestic journals, and for the strong support he has offered me in co-authoring with me, from all of which I have benefited. It is due to my strenuous studying here over the past three years and a half with his help and under his guidance and thanks to my constant progressing in an environment that he as an important part has been enhancing that I have now grown up to be a confident and independent researcher. It is for the above reasons that I also thank him for his acceptance of me as one of his PhD students, which made my long-cherished dream to pursue in-depth knowledge come true. I will not forget the first day of my arrival on this picturesque and tranquil land when he and I had a small talk at a Subway where he bought a long ham for me. I will also not forget his kindness, patience, and encouragement in supervising me.

I would also like to thank my co-supervisor, Professor Judy Parr, for her very earnest revision of my thesis, her very pertinent advice, and her great care for my study and life. Her academic works have been exemplary for me and inspiring me. I also cherish heartfelt thanks for her past efficacious administrative work. I always feel like a carefree roadside New Zealand kea foraging for inspirations at this Faculty.
I would also like to express my sincere thanks to Associate Professor Hu Guangwei, at the National Institute of Education, Nanyang Technological University, Singapore, for his having led me to the topic of reactivity in L2 writing when I pursued my MA study.

I would also like to thank other scholars at the Faculty of Education, and at the Faculty of Arts, for their illumination, especially, Professor Rod Ellis, for the lecture and seminars he has given us, Dr. Rebecca Jesson and Professor Timothy Teo, for the comments she and he offered on my PhD proposal, and Dr. Shaofeng Li and Dr. Natsuko Shintani, for their enlightening talk with me.

I would also like to thank The University of Auckland for having offered me the UOA Doctoral Scholarship for the completion of my PhD program in education.

I always feel sorry for only being able to accompany my family in China sporadically while I was studying overseas. I shall keep their understanding and support in the depth of my heart. Towards the final stage of my PhD study, the cheery smiles of my daughter, Shiwen, and the scenes of her playing her naughty deeds have always been flashing in my mind, even when we are oceans apart.
**TABLE OF CONTENTS**

CHAPTER 1 INTRODUCTION

1.1 Verbal Reports and Concurrent Verbal Reports 2
1.2 The Wide Use of Concurrent Verbal Reports 5
   1.2.1 The use of concurrent verbal reports in L1 research 7
   1.2.2 The use of concurrent verbal reports in L2 research and L2 writing 8
1.3 The Problems of Validity and the Significance of My Research 10

CHAPTER 2 LITERATURE REVIEW

2.1 Controversies over the Validity Issue 14
   2.1.1 Ericsson and Simon’s predictions based on the information processing theory 14
   2.1.2 Alternative approaches to reactivity 18
   2.1.3 Reactivity and task type 21
   2.1.4 Reactivity and proficiency 24
   2.1.5 Reactivity and working memory capacity 26
   2.1.6 Veridicality of concurrent verbal reports 31
   2.1.7 Controversies over reactivity and veridicality in SLA and writing research 34
2.2 Empirical Studies on Reactivity 39
   2.2.1 Non-SLA studies using non-verbal and verbal tasks 40
   2.2.2 SLA studies 41
   2.2.3 Reactivity studies in L1 writing 49
   2.2.4 Reactivity studies in L2/bilingual writing 56
   2.2.5 Summary 58
LIST OF TABLES

Table 3.1 Time frame and arrangement of the project 84
Table 3.2 Experimental grouping of participants 89
Table 3.3 Intrarater reliability: Pearson Product Moment correlation coefficients 116
Table 4.1 WMC One, WMC Two, and WMC Average raw and percentage scores 121
Table 4.2 Descriptive statistics of WMC Average and the results of t-tests 122
Table 4.3 Descriptive statistics of WMC One and Two and the results of t-tests 122
Table 4.4 Distribution of participants having three levels of WMC 123
Table 4.5 Descriptive statistics for baseline data 124
Table 4.6 Descriptive statistics for main task data 125
Table 4.7 Distribution of high-, mid-, and low-achievers in 20 groupings 126
Table 4.8 Results of regression analyses of WMC Average on main task performance 127
Table 4.9 Results of regression analyses of WMC One on main task performance 128
Table 4.10 Results of regression analyses of WMC Two on main task performance 129
Table 4.11 Results of regression analyses of baseline task on main task performance 130
Table 4.12 Results of t-tests run on baseline data 132
Table 4.13 Descriptive statistics and results of ANCOVAs for 11 variables 138
Table 4.14 Results of ANCOVA for Speed One after normalization 140
Table 4.15 Descriptive statistics for different Speed One × Condition groups 141
Table 4.16 Descriptive statistics and results of ANCOVA for Speed Three after normalization 145
Table 4.17 Descriptive statistics and results of ANCOVA for Length Three after normalization 146
Table 4.18 Descriptive statistics and results of ANCOVA for Length Four 148
Table 4.19 Results of ANCOVA for Speed Two after normalization 150
Table 4.20 Descriptive statistics for different Speed Two × Condition groups 152
Table 4.21 Results of ANCOVA for Speed Four after normalization 155
Table 4.22 Results of ANCOVA for General Complexity with interaction re-examined 156
Table 4.23 Descriptive statistics for different General Complexity × Condition groups 158
Table 4.24 Results of ANCOVA for CVU with interaction re-examined 161
Table 4.25 Descriptive statistics for different CVU × Condition groups 163
Table 4.26 Results of ANCOVA for Organization with interaction re-examined 165
Table 4.27 Descriptive statistics of Organization for different WMC × Condition groups 167
Table 4.28 Finalized descriptive statistics and results of ANCOVAs and pairwise comparisons 169
Table 4.29 Results of analyses with WMC One as a covariate 174
Table 4.30 Results of analyses with WMC Two as a covariate 178
Table 4.31 Effects of TA on L2 writers’ performance in writing a narrative task 184
Table 4.32 Descriptive statistics and the results of ANCOVAs run on Yang et al.’s data 186
Table 4.33 Contrast of the effect sizes of TA in two studies 188
LIST OF FIGURES

Figure 4.1 Baseline Task Speed One × Condition Interaction 140
Figure 4.2 Subgroup Between- and Within-Condition Comparisons on Speed One 142
Figure 4.3 Baseline Task Speed Two × Condition Interaction 151
Figure 4.4 Subgroup Between- and Within-Condition Comparisons on Speed Two 153
Figure 4.5 Baseline General Complexity × Condition Interaction on General Complexity 157
Figure 4.6 Subgroup Between- and Within-Condition Comparisons on General Complexity 159
Figure 4.7 Baseline CVU and Condition Interaction 162
Figure 4.8 Subgroup Between- and Within-Condition Comparisons on CVU 164
Figure 4.9 WMC and Condition Interaction on Organization 166
Figure 4.10 Subgroup Between- and Within-Condition Comparisons on Organization 168
Figure 4.11 Interaction between Task and Condition for General Complexity 190
Figure 4.12 Interaction between Task and Condition for Syntactic Complexity 191
CHAPTER 1 INTRODUCTION

Verbal reports, as a method to collect data on human thinking, require informants to report their thoughts that are of interest to researchers. Concurrent verbal reports (CVRs), which are real-time records of ongoing thoughts generated by informants while they complete a task, afford better access to the human mind because such reports are able to resist memory decay better than retrospective verbal reports (RVRs) (e.g., Ericsson & Simon, 1993). Given their important role in representing thoughts palpably, CVRs have been widely used as a data-eliciting tool to explore what people think and/or how they make decisions in wide-ranging areas of inquiry, covering accounting, ergonomics, software engineering, medicine, and education (for specific studies, see Bowles, 2010a; Ericsson & Simon, 1993). The latest use of CVRs, in the field of second language (L2) writing, is made in Willey and Tanimoto’s (2015) study, where CVRs were taken as a credible representation of teachers’ editing processes. Arguably, CVRs have greatly assisted researchers in prying open the black box of the human mind over the last century.

However, for quite some time, there has been doubt cast upon the validity of CVRs (e.g., Afflerbach & Johnston, 1984; Cohen, 2000, 2013; Ellis, 2001; Ericsson & Simon, 1984, 1993; Jääskeläinen, 2000; Jourdenais, 2001; Lyons, 1986; McGeorge, & Burton, 1989; Nisbett & Wilson, 1977; Olson et al., 1984; Payne, 1994; Payne, Braunstein, & Carroll, 1978; Rowe, 1985; Russo, Johnson, & Stephens, 1989; Smagorinsky, 1989; Wigglesworth, 2005; Wilson, 1994). Concerns have been raised as to whether gathering CVRs alters the very cognitive processes that the reports are supposed to represent and keep intact (i.e., the reactivity issue) and whether CVRs reflect the cognitive processes completely and accurately (i.e., the veridicality issue). With serious reactivity, the actual processes to be measured would be gone or changed upon measurement, and the CVRs elicited could no longer be trusted because they measured something else and therefore
represented a distorted version of the actual processes, even if they could measure accurately. Also, crucially, if CVRs suffered from serious nonveridicality, that is, if they represented a seriously insufficient version of what they are supposed to represent, and/or if they elicited information that is irrelevant to or deviates from actual processes, they would lose their value as a valid instrument because their contents were partial and/or untrue.

As the opening chapter of this thesis, Chapter 1 establishes the significance of my research by demonstrating how widely CVRs have been used in first language (L1) research and L2 research without appropriately and sufficiently addressing their potential problems of validity. As a start, some definitional issues are clarified to ensure that further discussion of the validity issues is based on a proper understanding of what CVRs are and how CVRs differ from other types of verbal reports, especially introspection. The wide use of CVRs in L1 and in L2 research, their application in writing research in particular, is then examined, and questions are raised about their validity in the existent research. The final section of this chapter focuses on the problems of validity and highlights the significance of my research.

1.1 Verbal Reports and Concurrent Verbal Reports

To start discussion on the validity of CVRs, it is fundamental to distinguish verbal reports given that verbal reports have different forms (Ericsson & Simon, 1993; Cohen, 1991; Leow & Morgan-Short, 2004; Olson, Duffy, & Mack, 1984). They can be categorized by the point of time when they are given, into CVRs and RVRs, with the former being generated simultaneously when a task is performed while the latter are given, preferably immediately, after a task is completed (Ericsson & Simon, 1979). For CVRs, Ericsson and Simon (1993; see also Ericsson & Simon, 1978) made a distinction between three levels. Level 1 verbalization is “simply the vocalization of covert
articulatory or oral encodings” where “there are no intermediate processes, and the subject needs expend no special effort to communicate his thoughts”; Level 2 verbalization “involves description, or rather explication of the thought content” which incorporates “information that is held in a compressed internal format or in an encoding that is not isomorphic with language”; Level 3 verbalization “requires the subject to explain his thought processes or thoughts” where verbalization is “not simply a recoding of information already present in STM, but requires linking this information to earlier thoughts and information attended to previously” (p. 79). Level 1 verbalization was termed talk-aloud, and level 2 verbalization think-aloud. What distinguishes them is the latter’s inclusion of intermediate processes that encode information to make it verbalizable, which are not entailed by the former. What distinguishes Level 2 verbalization from Level 3 verbalization is that the latter involves “additional interpretative processes”. Ericsson and Simon suggested that Level 1 and Level 2 verbal protocols represent the actual thought processes more closely than Level 3 protocols. They therefore recommended use of Level 1 and Level 2 as the ideal means to gain information about thinking processes that enters the short-term memory (STM) and appears in focal attention.

Concurrent verbalization during writing, Smagorinsky (1989) assumed, involved intermediate encoding processes and belonged to Level 2 verbalization. This assumption was made despite Cooper and Holzman’s (1983) previous suggestion that writing was not characterized by problem-solving and that verbalization of writing processes did not belong to this level. Based on Ericsson and Simon’s categorization, CVRs in writing may also incorporate talk-aloud elements, when participants simply vocalize the words that come to their minds spontaneously, which are ready linguistic encodings. Similarly supporting a mixture of the two types, Janssen, van Waes, and van den Bergh (1996)
suggested, “in protocols writers both report activities requiring little translation to an oral code (e.g., rereading), and activities requiring more recoding (e.g., explanation of chosen communicative strategy)” (p. 236-237). Additionally, it should be noted that, not limited to writing, CVRs may also contain metacognitive elements, which provide reasons for certain behaviors or decisions, although the verbalizers may not intentionally mean to justify.

In SLA research on the reactivity of CVRs, a categorization of two types of verbalization has been utilized. Concurrent verbalizations of thoughts per se were referred to as *nonmetalinguistic* and verbalizations with explanations and justifications were referred to as *metalinguistic* (Bowles, 2008; Bowles & Leow, 2005; Ericsson & Simon, 1993; Leow & Morgan-Short, 2004). They were later referred to as *nonmetacognitive* and *metacognitive* verbal reports to cover both verbal and non-verbal tasks (Bowles, 2010a).

For concurrent verbalization, an additional distinction, based on the coverage of thoughts to be reported, was made by Hayes and Flower (1983). They differentiated two types of concurrent verbalization: directed reports and CVRs. While in CVRs, participants are required to report all thoughts that appear in their minds, in giving directed reports, they are required only to report specific information that is part of their thinking.

Recently, Cohen (2013) juxtaposed introspection, think-aloud, and retrospection. He explains regarding the first two, “in the case of think-aloud data, the subjects merely voice their thoughts without trying to analysed or explain what they are doing, while introspection implies that they report on *what they think* they are doing to accomplish a task” (p. 1; italics added). This represents an effort to distinguish think-aloud from introspection, as his former categories (i.e., self-report, self-observation and self-
revelation, Cohen, 1991) did not highlight the distinction. In fact, prior to Cohen, Payne (1994) attempted to disentangle TA from introspection by defining the latter as “the trained observation of the contents of consciousness under controlled conditions” (p. 245).

Cohen’s (and others’) endeavours are appropriate considering the historical association of the use of verbal reports with the introspectionists, who usually asked trained participants to observe their thoughts and report them and then used the reports as data on thinking (Crutcher, 1994). In this case, observation and follow-up reporting may unavoidably entail interpretative elements, making the reports deviate from real happenings in reporters’ minds and approximate to what Ericsson and Simon termed Level 3 verbalization. Ericsson and Crutcher (1991) suggested that the problems incurred in introspective reports should not be attributed to verbal reports in general. Crutcher (1994) exemplified the importance of the TA-introspection distinction by pointing out that in fact some educationists or psychologists had emphasized that participants recalled their thoughts as they were, rather than interpreting them. The fact is that, as Payne (1994) pointed out, the introspective method was “virtually abandoned in 20th-century America because of criticisms by Watson and other behaviorists directed at the objectivity of the method as a basis for scientific psychology” (p. 245).

In this study, CVRs in L2 writing are defined as concurrent nonmetacognitive verbal reports that mingle both Level 1 and Level 2 verbalizations. They are not directed reports, but require all thoughts to be reported. They are not introspections and do not require additional justifications or comments. They are equated with think-alouds (TAs).

1.2 The Wide Use of Concurrent Verbal Reports

The use of CVRs has a long history. The first use of concurrent verbal reports, according to Smagorinsky (1989), may date back to the first two or three decades of the
last century when Duncker (1926) and Claparede (1934) started to use the think-aloud (TA) technique to collect data on thinking in problem solving. In Claparede’s instructions, for example, participants were asked to “tell me everything that passes through your head during the work searching for the solution to the problem” (Claparede, 1934, as cited in Smagorinsky, 1989, p. 464), just as is required in research today.

The usefulness of CVRs in identifying psychological processes in problem-solving tasks has long been recognized (e.g., Newell & Simon, 1972). Hayes and Flower (1980), for example, recognized that CVRs represent simultaneous information on the activities a subject engages in temporal sequence. Smagorinsky (1989) emphasized the process-tapping role, stressing that, in a protocol, the focus is “not the solution to a sequence of actions, but “the processes underlying the sequence itself” (p. 465). Because it has been difficult to find a viable, practical alternative, as a real-time means of gaining access to the workings of the mind, CVRs have continued to be employed. Crutcher (1994) further pointed out that CVRs had been used for more undertakings, for example, testing or evaluating hypotheses or models as opposed to just generating hypotheses (see, e.g., earlier criticisms or thoughts by Nisbett & Wilson, 1977 and Olson et al., 1984), revealing major theoretical concerns and providing refined insights into cognitive processes (Compton & Logan, 1991; Siegler, 1987; Siegler, 1989), exploring mental structures mediating learning and memory (Bellezza, 1986), etc. He realized that the post or ad hoc nature of CVRs in constructing theories had been less controversial due to their increasing roles in testing models on intermediate mental states (e.g., Trabasso & Suh, 1993) and that their roles had been formally recognized and documented.

The usefulness of verbal reports has also been recognized in L1 and L2 research (Bowles, 2010a; Cohen, 1991; Gass & Mackey, 2000; Green, 1998), where they are used to explore the not directly observable “process phenomena” (Gass & Mackey, 2000,
They have also proved to be applicable to a wide range of populations (e.g., Cohen, 1986). CVRs have particularly been favoured in studying the writing process from a cognitive perspective.

In what follows next, the various dimensions of the use of CVRs in L1 and L2 research are presented as a backdrop for this study.

1.2.1 The use of concurrent verbal reports in L1 research

CVRs have been used widely to investigate L1 reading (e.g., Folger, 2001; Harmon, 2000; Olson et al., 1984), writing (e.g., Green & Sutton, 2003; Langer, 1986; Selfe, 1984; Swarts, Flower, & Hayes, 1984; Witte, 1987), and language testing (e.g., Wijgh, 1996; Wolfe, Kao, & Ranney, 1998). In L1 writing research, CVRs, instead of direct behavioural measures (e.g., RT, time on process), were used as manifestations of thinking that occurred during the writing process in a number of important studies that modelled L1 writing processes cognitively and psychologically (Flower & Hayes, 1980a, 1980b, 1981a, 1981b, 1981c; Hayes & Flower, 1983, 1986), or distinguished skilled/older writers from unskilled/younger writers (Bereiter & Scardamalia, 1987; Bereiter, Burtis, & Scardamalia, 1988; Flower, & Hayes, 1980b). CVRs were also used as means to investigate writing subprocesses such as revision (e.g., Zellermayer & Cohen, 1996), to differentiate texts in terms of their cognitive demands (e.g., Durst, 1987), and to explore how cognitive activities are distributed over text production and how they are related to text quality (e.g., Breetvelt, Bergh, & Rijlaarsdam, 1994). Breetvelt et al. (1994), for example, identified 11 categories of cognitive activities, based on Flower and Hayes’s (1983) writing process model, from ninth graders’ CVRs in expository writing tasks. The researchers stated explicitly, “The cognitive activities during the writing process are operationalized as the thinking process that is verbalized and registered during the performance of the writing assignment”. One commonality
among studies that used CVRs to explore various aspects of writing is that all recognized CVRs as intact, real, and true equivalents of cognitive processes that happened during text production.

1.2.2 The use of concurrent verbal reports in L2 research and L2 writing

The application of CVRs has extended to a wide range of L2 research areas studying L2 reading processes in comparison to L1 reading processes (e.g., Block, 1992), mental translation processes (Kern, 1994), L2 reading and test-taking strategies (e.g., Cohen, 1986), differences between L1 and L2 reading strategies (e.g., Davis & Bistodeau, 1993), L1 use in L2 reading (e.g., Upton & Lee-Thompson, 2001), translation (e.g., Enriquez Raido, 2014; Ronowicz, Hehir, Kaimi, Kojima, & Lee, 2005), and the relationship between awareness/attention and learning (Leow, 1997, 1998, 2000, 2001a, 2001b; Rosa & Leow, 2004a, 2004b).

In L2 writing research, verbal reports have important use as a means to obtain online data, for example, to elaborate L2 writing subprocesses (e.g., Manchón, Roca de Larios, & Murphy, 2009; Roca de Larios, Marín, & Murphy, 2001; Zimmermann, 2000), to assess writers’ awareness (e.g., Armengol & Cots, 2009; Cumming, 1989; Qi & Lapkin, 2001; Swain & Lapkin, 1995) and types of thinking (e.g., Cumming, 1990a), to probe into L1 and L2 writing strategies and cognitive processing (e.g., Whalen & Ménard, 1995), to investigate editing of scientific texts (Willey & Tanimoto, 2015), to unveil raters’ decision-making processes while scoring ESL compositions (Cumming, 1990b; Cumming et al., 2002; Vaughn, 1991), to uncover learners’ use of dictionaries during writing (Chon, 2008), to study the effects of linguistic experience on writing fluency (e.g., Chenoweth & Hayes, 2001), to compare L1 writing, L2 writing, and translation from L1 to L2 (e.g., Uzawa, 1996), to investigate the use of L1 and L2 in L2 writing
(e.g., Wang & Wen, 2002), and to determine how L2 writers attend to different types of written feedback (e.g., Sachs & Polio, 2007).

An example showing how CVRs were used to explore L2 writing processes is Roca de Larios et al.’s (2001) study on the temporal expenditure and distribution of formulation processes in relation to L2 (English) proficiency and across L1 (Spanish) and L2 writing conditions. L2 writers’ protocols were coded first to identify episodes that were characterized by little ambiguity in boundary (e.g., task reading and interpretation, rereading of the just written, evaluation, and meta-comments). Then, formulation was segmented from planning based on a stipulation that formulation included “both the verbalization of written material and those other utterances that, because of their strict linear nature (lexical units, syntactic structures, etc.), could be considered clear candidates for becoming part of the text”, while “planning covered all those segments that indicated operations involving the retrieval and/or development of ideas, aims, organization, etc. at a prelinear level” (p. 510-511; italics original). Formulation was also distinguished from revision by the “distance from the point where the modification of the text took place to the point of inscription (Lutz, 1987; Matsuhashi, 1987; Severinson Eklundh & Kollberg, 1996)” (p. 512). The formulation processes were further divided into what they termed fluent versus problem-solving formulation processes. The time spent on formulation (sub)processes and the temporal distribution of formulation were then calculated. Clearly, in this study, CVRs were taken as reliable records for categorizing writing processes and identifying their temporality.

While CVRs were well used in L1 and L2 research, an important concern that arises is their validity, that is, their reactivity and veridicality, which might undermine many of the findings thereby derived. Despite its importance, the validity issue was seldom
addressed, if not unattended, in these empirical studies, leaving an impression that the issue would not pose a trouble at all, when evidence is lacking for such an assumption.

1.3 The Problems of Validity and the Significance of My Research

Discussions on the validity of verbal reports, in fact, have almost invariably accompanied their use. Ericsson and Simon (1984, 1993) made predictions on the reactivity of CVRs in relation to three levels of concurrent verbalization and discussed in detail the completeness of CVRs, together with the problem of epiphenomenality (e.g., reporting of irrelevant, parallel activities independent of actual mental processes). Russo et al. (1989) have also outlined the validity issue as having “at least two forms: reactivity and nonveridicality”. For the latter they made a distinction between errors of omission and errors of commission (p. 760). Others also pointed out the issues of reactivity and nonveridicality (e.g., Brinkman, 1993; Cooper & Holzman, 1983; Crutcher, 1994; Stratman & Hamp-Lyons, 1994). Brinkman (1993), for example, remarked that, “when there are several ways to do a task, it may be that verbalization has the effect that the normal way is changed for a way which is more easily described” (p. 1381-1382). He also suggested that nonveridicality may happen when “in a highly practiced task, the speed of the on-going processes may exceed the ability to verbalize them completely” (p. 1382).

In this thesis, reactivity is taken as concerning whether the reporting alters cognitive processes. Veridicality is viewed as a superordinate term concerning both completeness and accuracy of CVRs. By accuracy is meant whether what is reported in CVRs accurately reflects what is going on in the mind. While completeness concerns errors of omission, in Russo et al.’s (1989) term, accuracy relates to what they defined as errors of commission, and to Ericsson and Simon’s (1993) term of epiphenomenality.
Just as any instrument of measurement is supposed not to change the object it measures in any fashion and is meant to measure it accurately, so is the technique of TA intended to elicit thinking processes. If there were serious reactivity, findings obtained and conclusions drawn from CVR data in previous L2 and L1 studies would be greatly undermined. Specifically as regards writing, the following concerns would hold. First, the writing models established based on TA data depicting both L1 and L2 writing processes or subprocesses (e.g., Flower & Hayes, 1980a, 1980b, 1981a, 1981b, 1981c; Green & Sutton, 2003; Hayes & Flower, 1983, 1986; Zellermayer & Cohen, 1996) would be questionable, in which case the modelling might have been based on altered processes, be they sequential or recursive. Then, if alterations caused by TA differed among L1 or L2 writers of different levels of proficiency, the distinctions between expert and novice or skilled and unskilled writers identified in verbal reports, for example, as regards writing patterns, (sub)processes, difficulties, or strategies (e.g., (Bereiter & Scardamalia, 1987; Bereiter, Burtis, & Scardamalia, 1988; Durst, 1987), as Hayes, Flower, Schriver, Stratman, and Carey (1987) speculated, might not be valid because it would remain problematic as to whether the differences represented in CVRs were caused by the reporting or whether they reflected the actual differences characteristic of different writers. Third, in studies that differentiated texts in terms of their cognitive demands by employing CVRs (e.g., Durst, 1987), there might also be a risk that the additional load of TA might have a confounding role if it did in some way interact with different types of writing tasks by inhibiting or triggering certain levels of operations. In this case, it would be problematic to attribute all differences in cognitive demands of different texts solely to their intrinsic differences. Fourth, investigations into how cognitive processes were related to text quality (e.g., Breetvelt et al., 1994) also would be unduly affected, given the possible effects of TA on both cognitive processes
and text quality. Finally, conclusions concerning the distributions of cognitive activities or (sub)processes (e.g., Breetvelt et al., 1994; Roca de Larios et al., 2001) would also be undermined if the extra work of TA were able to change the sequences or recursions of thoughts or impose or generate processes.

Just as any instrument of measurement is supposed to measure an object completely and accurately, so is the technique of TA intended to elicit thinking processes. If there were serious nonveridicality, findings obtained and conclusions drawn from CVR data in previous L2 and L1 studies would need to be reconsidered. For example, in Leow’s (2000, 2001a, 2001b) series of studies on the roles of attention and awareness in leading to subsequent processing of targeted forms, attention and awareness were operationalized as mentions or other signs in participants’ CVRs. But, if the learners failed to report exhaustively what they heeded in the process of task completion, inferences made between attention and awareness and subsequent learning would have been dubious because attention and awareness may not have been fully embodied in the verbal accounts. Regarding writing, the same concern might be raised with respect to a series of studies that relied on CVR information to investigate attention, awareness, and noticing in L2 writing processes (e.g., Armengol & Cots, 2009; Cumming, 1989; Qi & Lapkin, 2001; Swain & Lapkin, 1995). Also, had nonveridicality posed a problem, a number of other studies, which yielded writing models depicting (sub)processes, reported findings about learner difficulties and strategies, made comparisons between writers of different writing capabilities, or drew other conclusions based on TA data (e.g., e.g., Bereiter & Scardamalia, 1987; Bereiter, Burtis, & Scardamalia, 1988; Durst, 1987; Flower & Hayes, 1980a, 1980b, 1981a, 1981b, 1981c; Green & Sutton, 2003; Hayes & Flower, 1983, 1986; Zellermayer & Cohen, 1996), would have to be re-evaluated. Completeness may be a bigger concern for RVRs than for CVRs due both to
the immediacy of reporting in the case of the latter and to the fact that participants are usually expected to be reporting honestly, but empirical evidence should be sought before a conclusion is reached.

To address the validity of CVRs, quite a number of studies on reactivity have been conducted in cognitive psychology (Ericsson & Simon, 1984, 1993; Fox, Ericsson, & Best, 2011). However, just as will be shown in some detail in my forthcoming review of the literature, only a very limited few of them used verbal or language-related tasks, and there are few studies on veridicality. In L2 research, where CVRs have been used not infrequently to gather process data, and where the reactivity and nonveridicality of CVRs might be greater due to L2 learners’ proficiency in dealing with L2 learning tasks (e.g., Ellis, 2001), the interest in researching the validity of CVRs is just rising (Bowles, 2010). So far, there have been only a couple of reactivity studies on L2 writing (Yanguas, & Lado, 2012; Yang, Hu, & Zhang, 2014) and only one study on the veridicality of CVRs during L2 essay rating (Barkaoui, 2011), although there have been a dozen of studies on reactivity that employed L2 tasks other than writing. The reactivity and veridicality of CVRs in L2 writing research is no doubt a meaningful area worth exploration.
CHAPTER 2 LITERATURE REVIEW

This chapter reviews both theoretical and empirical literature concerning the validity of CVRs. It foregrounds Ericsson and Simon’s (1993) predictions on, and defense for, the validity of CVRs based on their information processing framework. Then, it introduces other perspectives for looking at the validity issues which may challenge Ericsson and Simon’s conclusions. In this part of the review, it presents, most prominently, Russo et al.’s (1989) and Stratman and Hamp-Lyons’s (1994) reactivity models, teases out the possible interactions between reactivity and task type, reactivity and proficiency, and reactivity and working memory capacity (WMC), and examines doubts about the veridicality of CVRs, before it goes on to synthesize controversies over the validity of CVRs in the more specific areas of SLA and writing research. The chapter then reviews important empirical studies, both L1 and L2, on the reactivity and veridicality of CVRs, focusing on a critique of the few empirical studies using writing tasks. It finally points out the lack of validity research in L2 writing and establishes a niche for this study.

2.1 Controversies over the Validity Issue

2.1.1 Ericsson and Simon’s predictions based on the information processing theory

Ericsson and Simon’s (1984, 1993) framework for justifying the validity of think-aloud data is based on information processing theory (Newell & Simon, 1972), which draws an analogy between the human mind and a computer processor and therefore between human cognition and the information-processing process. Based on this framework, all human cognition can be viewed as “a sequence of internal states successively transformed by a series of information processes” (Ericsson & Simon, 1993, p. 11). The information-processing theory hypothesizes the existence of two types of memory storage, long-term memory (LTM) and short-term memory (STM). While the
former is where the bulk of procedural and factual knowledge resides and is to be retrieved and transferred to STM before being reported, the latter holds a modicum of information which is in an active state, likely to vanish soon, and is easily available and readily reportable (see also Dechert’s 1987 assumptions regarding the reportability of information that is in focal attention and outside of it). Ericsson and Simon posited that CVR taps STM, and that “the information that is heeded during performance of a task, is the information that is reportable; and the information that is reported is information that is heeded” (p. 167). In other words, the validity of CVRs is considered in reference to the heeded information only.

Ericsson and Simon distinguished among three levels of verbalization and made different predictions regarding their respective reactive effects (See also, e.g., Ericsson & Simon, 1978, 1984). Their taxonomical accounts made the absence or presence (and the number) of mediating processes between the point of attending to information in STM and the point of verbalizing it a pivotal element for determining the validity of CVRs protocols (Payne, 1994). They predicted that Level 1 and Level 2 verbalizations would not affect the sequence of cognitive processes to be reported or the structure of the processing because neither incurred extraneous processes, but that Level 3 verbalization would because it required justification. These predictions would mean that, in the case of Level 1 and Level 2 verbalizations, the heeded information is only there to be reported, whether it is in a readily reportable form in the case of Level 1 verbalization or is to be verbally recoded in the case of Level 2 verbalization. In other words, no new information is introduced to attention, nor is any otherwise present information suppressed from attention, nor is the sequence and structure of information disturbed. The difference between Level 1 and Level 2 verbalizations in reactivity is that, while in the case of the former, no additional time for processing in terms of oral
coding may be needed, in the case of the latter, some processing time for recoding may be needed, hence a possible prolonging of time on task or latency. As Payne, Braunstein, and Carroll (1978) stated, “the verbal protocol procedure slows down the process slightly but does not change it fundamentally” (p. 36) (see also similar views in Brand, 1987; Cooper & Holzman, 1983; Nisbett & Wilson, 1977; North, 1987).

These predictions of reactive effects appear robust. Ericsson and Simon (1993) reviewed 30 empirical studies adopting a between-groups design in psychological and cognitive research and reported that they found no results violating their predictions, especially their predictions on L2 verbalization. Fox et al.’s (2011) meta-analysis of 94 studies also confirmed the nonreactivity of nonmetalinguistic verbal reports on cognitive processes. In these reviews, however, an overall increase in solution time tended to be apparent, despite no presence of interference with the sequences of internal thought processes. It should be noted that none of the studies reviewed by Ericsson and Simon and their colleagues used writing as the main activity.

As regards the veridicality of verbal reports, there had been concerns prior to Ericsson and Simon’s (1993) synthesis that the information provided by CVRs was incomplete (e.g., Cooper & Holzman, 1983; Dechert, 1987; Dobrin, 1986; Nisbett & Wilson, 1977; Waern, 1988). Dechert (1987), for example, assumed that verbal reports would not contain procedural knowledge because “Human information retrieval is partly declarative and, as such, accessible for verbalization, and partly proceduralized and, therefore, not accessible for verbalization” (p. 97). Waern (1988) also noted that automated mental activities such as letter recognition did not appear in proficient native audiences’ immediate awareness and was not reportable.

Ericsson and Simon (1993) defended the completeness and accuracy of Level 1 and Level 2 verbalizations based on the information processing framework they established.
According to their framework, “only information in focal attention can be verbalized” (p. 90; emphasis original). That is, the issue of veridicality should be considered only in reference to what is verbalizable, which is what is heeded, and CVRs should not be expected to cover all thinking processes, as may have been expected (e.g., Dobrin, 1986; Seliger, 1983). In their view, both Level 1 and Level 2 verbalizations, due to the immediacy of reporting and removal of the effects of memory decay and, due to their involving only heeded information without any addition or any suppression, would not suffer from lack of veridicality. Ericsson and Simon examined previous allegations from studies questioning the completeness of verbal reports. They pointed out that what had been claimed to be missing in CVRs (e.g., recognition cues) were actually not heeded information that did not make an appearance in STM, and that other criticisms were unfair to CVRs because they indicated memory failure in the case of RVRs, where inferences rather than memories may be reported. According to them, RVRs may incur incomplete and/or inaccurate reporting of heeded information because of memory decay in recall due to the lapse of time, which may result in missing information and fabrications, while introspections may suffer from subjectivity and inaccuracy in reporting heeded information due to their reliance on informants’ own observations of and verbalizations on their behaviours. They therefore recommended the use of CVRs over RVRs and over introspections.

To ensure the validity of CVRs, Ericsson and Simon (1993) emphasized the importance of a robust data elicitation procedure (see also Bowles, 2010a, for data collection/analysis considerations). Precautions included asking participants to report their ongoing thoughts rather than their introspections, descriptions, or justifications, thereby minimizing/eliminating the influences of experimenters, etc. Ericsson and Simon (1980) also suggested the importance of making sure that what participants said
was consistent with their behavior. Ericsson and Simon believed that if these guidelines were strictly followed and the undue external experimental factors properly controlled, any possible reactivity and nonveridicality problems could be minimized. Regrettably, the empirical studies reviewed in Ericsson and Simon mostly concerned the reactive effects and empirical support is still lacking for the veridicality of CVRs.

2.1.2 Alternative approaches to reactivity

Different from Ericsson and Simon who made taxonomical accounts of reactivity, other scholars tried to clarify the issue of reactivity by disentangling the factors that might cause it. Amongst these were Russo et al. (1989) and Stratman and Hamp-Lyons (1994). Russo et al. (1989) summarized at least four potential causes of reactivity from the empirical literature, which included: 1) the pressure imposed on processing due to the additional demand of verbalization, 2) the effects of voice, 3) its potential benefits on learning, and 4) expectation of exposure of recordings to researchers. Russo et al. (1989) further pointed out that “These causes of reactivity are independent and task-specific in that any or all of them may be present depending on the primary task” (p. 764).

Specifically, Russo et al. elaborated the processing requirements that verbalization entailed, including preparation and operation of motor programs for articulation, the effort to ensure conformity to the TA instruction and an acceptable volume, recoding of particular oral codes to ensure intelligibility (Werner & Kaplan, 1963), possible oral expression of behaviours of partial automaticity, and recoding of information that is in a form not ready for oral reporting. Russo et al. believed that when these demands of verbalization for processing resources were so high as to exceed the “slack resources” available for it, subjects faced a choice between consuming some of the share of processing resources for normal operation of the primary task, which might cause
reactivity, and concentrating on the primary task without reporting their ongoing thoughts, which causes nonveridicality (p. 764). An analogous caution by Brinkman (1993) stated that thinking could advance too quickly to be available for verbalization, which would cause either reactivity or nonveridicality. Payne (1994) added that, besides the use of attentional resources, TA might change heeded information by diverting attention to what is readily verbalizable (e.g., Schooler, Ohlsson, & Brooks, 1993; Wilson & Schooler, 1991). These accounts all point to the increasing intensity of competition for working memory (WM) resources if a primary task and the TA task are to be juggled. As for aural effects, Russo et al. (1989) held that listening to one’s voice might enhance memorization (Penney, 1975) and performance might be boosted in tasks in which holding partial results in memory mattered and rehearsal was possible. The learning effects, they assumed, might occur via the opportunities afforded for reflection, which “may lead to the discovery of new strategies or to the improvement of old ones” (p. 764) and could be present especially in tasks that contained multiple optional strategies or components. Ahlum-Heath and Di Vesta (1986) provided evidence for the facilitative role of TA for beginners learning problem solving, because “verbalizing made salient the features required for sub-goal analysis that might otherwise have been unavailable (or not immediately available)” (p. 284). Similarly, Williams and Davids (1997) proposed that participants might be encouraged to change their strategy of searching for information due to hearing themselves repeatedly report a single strategy of extracting information. The fourth factor, which Russo et al. (1989) called “motivational shift”, referred to the tendency of subjects to make themselves sound good, knowing that their protocols would be listened to, transcribed, and analysed by experimenters, by exercising more care concerning accuracy (p. 765).
Among the four major potential causes of reactivity that Russo et al. (1989) identified, 1) might be detrimental to task performance, 2) might be more facilitative than detrimental, and both 3) and 4) might be facilitative. Following this distinction, then, the reactive effects of TA may be understood as a trade-off between all these factors that are possibly negative or positive to task performance.

Stratman and Hamp-Lyons (1994) later re-synthesized the reactivity-causing factors. In their synthesis, Stratman and Hamp-Lyons foregrounded the undue effects of improper instructions for CVRs, following Ericsson and Simon’s reiterative cautions, which, different from other factors in their categories which may be intrinsic, can be avoided. They also made explicit the effects of the physical or nonphysical presence of the experimenter. It is possible that informants feel the pressure imposed by the presence of (or thought of) the experimenter.

Both Russo et al.’s and Stratman and Hamp-Lyons’s approaches addressed the issue of reactivity by disentangling its causes. This is different from Ericsson and Simon’s taxonomical accounts which featured uniform predictions as regards the same level of concurrent verbalization. In fact, Ericsson and Simon’s rationale for the nonreactivity of nonmetacognitive (Level 1 and Level 2) CVRs was only obvious in what Russo et al. summarized as the extra processing demands. Ericsson and Simon may deem the other factors in Russo et al.’s list to be negligible or as avoidable, controllable, or able to be minimized. Smagorinsky (1989), following Ericsson and Simon, denied any influence from the physical act of verbalization, but he agreed that among others, “[t]he conditions of the protocol situation, including the specific instructions to the writer, the behaviour of the researcher, and the time constraints” (p. 475) should be identified because they may affect what informants may tell.
Overall, due to one’s familiarity with his/her own voice, merely hearing it generated from concurrent verbalization while writing may incur far less phonological interference than is caused by an unattended speech task secondary to a primary composing task (see, e.g., Madigan, Johnson & Linton, 1994 for minor effects of such tasks). Rather, vocalizing may facilitate recall (Penney, 1975) and speaking aloud helps memorization (e.g., Crowder, 1970), both of which may have a role to play in facilitating or accelerating the writing process. In writing, for example, thinking one’s macro-planning aloud and hearing it may help one keep such planning in mind and therefore speed up writing. While external factors (unwanted experiment conditions, including improper instructions or experimenter intrusion) and some consequent informant factors such as their anticipation of exposure of their performance to the experimenter or researcher) may be controlled, how other factors (the demand of verbalization for WM, increased critical attention and awareness, enhanced memorization and recall) might interact is an interesting question.

2.1.3 Reactivity and task type

Although there are scholars who regard the processing and writing of texts as cognitive problem solving tasks (e.g., Breetvelt et al., 1994; Flower & Hayes, 1980b), there are more concerns about the differences between problem-solving tasks and language-related tasks. Smagorinsky (1989), for example, questioned the application of the findings from reactivity studies utilising problem-solving tasks to writing. He remarked, “… we can only speculate at this point how much bearing their findings have on the use of protocols to study writing” (p. 466). Russo et al. (1989) provided evidence that the reactivity of concurrent verbalization might vary with the type of tasks involved, since in his experiment that utilized a verbal task, a numerical task, a pictorial task, and a mental addition task, positive reactivity displayed as increased accuracy was found in
performing the numerical task, negative reactivity in the form of deceased accuracy was reported on the mental addition task, but no reactivity was found for the verbal task (i.e., the crossword puzzle) or the pictorial task (Ravens Progressive matrices). The mixed findings led the researchers to posit that “the impact of protocol generation depends strongly on the task, suggesting that the causes of reactivity are not general but due jointly to the demands of the task and to verbalization” (p. 762-763).

Stratman and Hamp-Lyons (1994) further considered the differences between language-related tasks and problem-solving tasks. They questioned if the conclusions drawn from nonverbal tasks could be readily transferred to such verbal tasks as reading and writing, which are “ill-defined tasks” where “subjects must specify partly or completely their own goals”, and to which subjects “may generate many equally satisfactory ‘solutions’” (p. 92). To the specific nature of writing tasks may be added the sometimes stream-of-consciousness nature of writing (e.g., Cooper & Holzman, 1983) and the creativity that may be involved. These characteristics of both reading and writing tasks suggest more complications in real-time reporting than what is typical of verbalizing thoughts experienced in a problem-solving task. That is, in verbal tasks, thoughts might not follow an easily traceable and therefore reportable pattern, being hard to define due to lack of regularities or conformities in logic, inference, or sequence. These considerations echo previous cautions about the difficulty in thinking aloud (e.g., Ericsson & Simon, 1984). As has been discussed, this difficulty would constitute a great demand for processing resources, risking reactivity, especially considering that informants may need to slow down their thinking processes in order to track them. Or, otherwise, it might pose a threat to veridicality, due to the error of omission, given that subjects might just skip the intricate and transient thoughts to spare effort for achievement of goals to which they attach greater importance, or, due to the error of
commission, if they turn nonmetacognitive think-alouds into metacognitive think-
alouds when they try to catch these thoughts post hoc (e.g., Ericsson & Simon, 1984). There is another possibility, namely, that informants might eschew complicated thoughts or reduce their creativity in anticipation of the difficulty in reporting them (e.g., Brinkman, 1993).

Smagorinsky (1989) related the issue of reactivity to the extent of automatization of operations required in a given task and argued that the reactive effects found in some general problem-solving tasks that involved perceptual motor processes (e.g., Thomas, 1974) did not apply to writing. He admitted the incompatibilities found between verbalization and manipulation in these tasks (e.g., solving a puzzle, or performing a task requiring switch setting), but assumed that because “[t]he act of writing … is more a transmission of thought to paper, a formal expression of thought, than a manifestation of a decision” (p. 467) (that is, because writing is not as much of a manipulation task given that typing on a keyboard or writing with a pen is largely automatized behaviour), TA in writing would not cause interference with performance in terms of executing writing, namely, actualizing thoughts. He suspected that writing could be viewed as solving maths problems with pen and paper, “since both involve thinking that triggers well-rehearsed sensory motor acts” (p. 467). Smagorinsky rightly pointed out TA would only affect, if at all, controlled processes as opposed to automatized ones. This is consistent with Ericsson and Simon’s (1993) hypotheses, given that automatized processes are not, or little, attended to and therefore not reportable or reported, which leads to their exemption from influence. One omission in Smagorinsky’s comments is that he only touched upon the executing processes of writing, though he did ask for more research on the effects of thinking aloud on writing processes. It should also be noted that L2 executing processes may not be as automated as L1 executing processes are
(Ellis & Yuan, 2004), indicating TA may affect L2 executing processes as well (e.g., Yang et al., 2014).

Besides the above concern that language tasks such as those in writing or reading might engender different reactive effects from those in problem-solving tasks, there have been cautions that the reactivity of CVRs might also vary within the same type of language tasks, like writing tasks with different requirements. For example, Witte (1987), when employing CVRs to explore writers’ pre-texts in composing, admitted that objections to the verbal method may relate to its undesirable influences on the frequency with which writers constructed pre-text and on the contents of the pre-text. He further predicted that these effects would “appear to do so inconsistently across writers and across writing tasks” (p. 400). Also, in a study investigating reactivity in L1 writing, Ransdell (1995), when explaining the limited effects of concurrent verbalization she found in her experiment, suggested that production of descriptive narratives about the first days of college, as a writing task, might not be cognitively taxing and tasks that “require a good deal more planning and organization than does the task involved in this study” might exert more reactivity (p. 96). Further research is evidently needed to explore how the reactivity of CVRs might vary with different writing tasks.

2.1.4 Reactivity and proficiency

Researchers have found individuals differ in the TA task, either in terms of their successful adaptation to the verbalization task, or in terms of the contents of their CVRs. For example, Cumming et al. (2002) observed that raters very experienced in scoring L2 essays were able to offer longer and more elaborate CVRs than did less experienced raters. This suggests that the reactive effects of CVRs might interact with individual differences, particularly, the informants’ ability to handle the primary task successfully.
In L1 research on reactivity, there has been some evidence that the reactivity of metacognitive thinking aloud might have differential effects for students of different abilities. For example, in Short, Schatschneider, Cuddy, Evans, Dellick, and Basili’s (1991) study, bright and average children benefited much more from metacognitive thinking aloud when they completed verbal analogies than disabled and developmentally handicapped children, with developmentally handicapped children benefiting the least. This finding might suggest a similar differential effect of nonmetacognitive thinking aloud among adult learners of language at different levels of proficiency.

In writing research, researchers have relied on the think-aloud technique to distinguish skilled from unskilled writers or revisers in an attempt to retrieve from expert task performances those aspects that might be modelled for classroom or programmed instruction (e.g., Bereiter & Scardamalia, 1987; Hayes, Flower, Schriver, Stratman, & Carey, 1987). However, concern arises as to the extent to which findings of novice-expert differences might be contaminated by the think-aloud tool. For example, when engaging both experts and novices to think aloud during a complex revision task, Hayes et al. (1987) admitted that “what may appear to be a difference between experts and novices in underlying writing or revising processes may sometimes partly be an artefact produced by the interaction between the degree of expertise a subject possesses and the constraint of giving a protocol” (p. 184). Stratman and Hamp-Lyons (1994) added that it was unclear if the novice subjects’ “means-ends” behaviors (coded from the protocols) were more negatively impacted by the TA condition than were the experts’ in Hayes et al.’s (1987) results. They further noted that, given Hayes et al.’s acknowledgement, noted above, they failed to analyse specifically the possible novice-expert differences in their ability to respond to TA in relation to the output measures reported. Stratman and
Hamp-Lyons further pointed out a lack of research on the differential effects of TA in previous literature and that even Ericsson and Simon (1984/93) did not address the question of expert-novice differences. Indeed, educational studies directly addressing this question have largely remained a research gap. Investigation of this gap is certainly warranted because serious differential effects may mean either an artificial exaggeration, or an understatment of actual novice-expert differences (e.g., in strategy use), making implications drawn from expert practices or processes problematic.

L2 learners with high writing proficiency may be able to handle the additional cognitive demands imposed by the act of thinking aloud better than those with low writing proficiency. Manchón et al. (2009) reported that more able learner writers possessed “a more multidimensional mental model” which enabled them to “make strategic decisions as to the allocation of attentional resources to various composing activities throughout the writing process” and to “tackle ideational, textual and stylistic problems in addition to those that derive from having to compensate for language deficits” (p. 117). Therefore, there might be a reason to assume that the reactive effects of TA for high achieving L2 writers would be lighter than those for low achieving L2 writers.

2.1.5 Reactivity and working memory capacity

Working memory capacity (WMC) has attracted extensive research in cognitive psychology (e.g., Baddeley, 2007; Conway, Jarrold, Kane, Miyake, & Towse, 2007a; Jarrold & Towse, 2006). WM (e.g., Baddeley, 1986, 2003; Baddeley & Hitch, 1974) is “a multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction” (Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005, p. 770). WM involves “the ability to maintain information in an active and readily accessible state, while concurrently and selectively processing new
information” (Conway, Jarrold, Kane, Miyake, & Towse, 2007b, p. 3). Jarrold and Towse (2006) also viewed WM as indicative of “the ability to hold in mind information in the face of potentially interfering distraction in order to guide behavior” (p. 39). In contrast to its antecedent concept, STM, which was generally defined as a static, temporary storage of information, WM deals with active maintenance of information, at the centre of which function is a cognitive control mechanism that orchestrates information flow and processing, alongside its storage. WMC may, therefore, explain differences in implementing higher order cognitive activities (e.g., reading comprehension), and individual differences in WMC may well be reflected in differences in performance on varied complex cognitive tasks such as language comprehension, including reading and sentence processing (see, e.g., Daneman & Merikle, 1996, for a meta-analysis), and general intellectual abilities, including reasoning and general fluid intelligence (see, e.g., Ackerman, Beier, & Boyle, 2005 for a meta-analysis).

In SLA research, WMC has received wide attention as one factor of individual difference, and has even been viewed as one important facet of language aptitude (e.g., Robinson, 2005a; Skehan, 2002). The effects of individual difference in WMC on, and its association with, L2 learning have been widely researched (e.g., Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009; Juffs, 2005; Kormos & Sáfár, 2008; Mackey, Adams, Stafford, & Winke, 2010; Mackey, Philp, Egi, Fujii, & Tatsumi, 2002; Robinson, 2002, 2005b; Sagarra, 2007, 2008; Tokowicz, Michael, & Kroll, 2004; Trofimovich, Ammar, & Gatbonton, 2007; Walter, 2004; see also Williams, 2012, for a review).

Studies concerning the relationship between individual differences in WMC and writing performance or processes (e.g., Bergsleithner, 2010; also for a review, see Swanson & Berninger, 1996), especially those exploring the effects of WMC constraints
(see Olive, 2004 for a review of studies using the dual/triple-task technique) and those of WMC load release on writing (e.g., Ellis & Yuan, 2004; Kellogg, 1988, 2001) have suggested positive correlations between available WMC and writing performance and/or writing processes. Bergsleithner (2010), for example, found that L2 individuals with smaller WMC were more prone to producing less accurate and less complex narratives. Ransdell, Arecco and Levy (2001) also reported decreased fluency measured by words word-processed per minute for bilinguals writing in their L1, loaded concurrently with a secondary 6-digit memory task, though their overall quality of writing remained constant. Interestingly, Kellogg (1988) provided contrary evidence that a reduction of part of the WM load facilitated writing. He observed better-quality documents when writers employed the strategy of mental or written outlining during prewriting, which may have eased attentional overload by allowing writers to focus processing time on translating ideas into text. His finding has been supported by Ellis and Yuan’s (2004) L2 study of the effects of pretask planning. In their study, pretask planners wrote narratives more fluently, and demonstrated increased syntactic variety (SV), which they attributed to lessened pressure on WM during composing due to pretask planning.

WM has also been related to writing processes in cognitive models of writing (Kellogg, 1996; McCutchen, 1996). Kellogg’s (1996) model distinguishes three basic systems involved in text production: formulation, execution and monitoring. Formulation involves planning and translating, execution consists of programming and executing, and monitoring comprises reading and editing. Similar as these subprocesses are to Flower and Hayes’s (1980) categorizations, the model delineates what components of WM that these processes tap, based on Baddeley’s (1986) tripartite distinction of WM into the two slave systems of the phonological loop and the
visuospatial sketchpad, and a central executive system. While the two slave systems store and process auditory and verbal information, and visual and spatial information, respectively, the central executive is a “multipurpose, limited-capacity system” which assists when the slave systems are overwhelmed. It is called on in controlled processing such as that involved in tasks demanding sustained effort; it “plays regulatory roles … whenever competing behaviors are simultaneously elicited by the environment” (Kellogg, 1988, p. 58). Kellogg suggests that planning involves the visuospatial sketchpad, both translating and reading require the phonological loop, but virtually all writing processes place demands on the central executive system, except for the executing subprocess. Kellogg’s assumption of WM allocation to writing processes seems to have been supported by Levy and Ransdell (2001), who overloaded each component of WM with a secondary task and found that the effects of such overloading were largely felt in the same subprocesses of writing as Kellogg’s model predicted. Therefore, Kellogg’s cognitive model of writing renders it theoretically possible to identify what processes of writing might be affected by TA if the TA demand for WM is specified.

Kellogg’s model, following previous assumptions and findings concerning the limited human WMC or information processing system (e.g., Dechert, 1987) or the limited capacities of certain components of WM (Baddeley, 1986), assumes contention for the limited capacity of the central executive among writing processes (also see Kellogg, 2001) or, as Ellis and Yuan (2004) state, “a trade-off of attention directed at the different processes” (p. 63). Kellogg’s (1996) cognitive model of writing suggests that if TA competes for WM resources that are already stretched, for example, during a cognitively demanding argumentative writing task, it might have a crowding-out effect on writing due to its seizure of attention otherwise allocated to or deserved by some
writing processes. Specifically with regard to verbal reporting, based on his contention theory, Kellogg (1996) predicted that “verbal protocols should at a minimum load the phonological loop and disrupt the quality and fluency of translation” (p. 69). The underlying logic was that concurrent verbalization may involve the phonological loop, on which the subprocess of translating is based and, therefore, there might be a conflict of interest between verbalization and translating due to potentially competing and strong mutual claims. Kellogg did not specify on what kind of translating processes thinking aloud might impact. Also, notably, Kellogg’s (1996) prediction did not extend to other processes, but there is a possibility that concurrent verbalization also places demands on the central executive (Baddeley, 1986) or other attentional resources that writing subprocesses other than translating also require, especially considering previous accounts of the multiple demands of TA for processing resources (Russo et al., 1989). In this case, the effects of thinking aloud might spread to a greater range of writing processes than translating. Ellis and Yuan (2004) applied Kellogg’s model to L2 writing research and rightly argued that all L2 writing processes, including execution, draw on the central executive since L2 learners may not have an adult, native-like automaticity in handwriting. Since more demands are placed on the central executive in L2 writing due to L2 learners’ low proficiency, greater reactive effects may be expected of TA given the competition for WM resources among L2 writing processes.

The possible role of WMC in regulating the effects of thinking aloud has already been suggested in previous efforts that identified the factors potentially contributing to reactivity (Stratman & Hamp-Lyons, 1994; see also Russo et al., 1989). Among these factors, the cognitive effort expended on CVR and, correspondingly, the limited WMC for the dual acts of verbalization and task completion stand out as individual difference factors mediating the relationship between TA and primary task implementation.
processes. This holds given that such cognitive constraints can be intrinsic and other external factors may be properly controlled, pre-empted, or shunned, or may have a negligible role if rigorous, tested experimental procedures are adhered to (for such procedures, see Bowles, 2010a; Cohen, 2000; Ericsson & Simon, 1993; Wade, 1990). More recently, Goo (2010) inferred that WMC might have a part to play in interacting with the potential reactive effects of thinking aloud, based on its predictive role in complex cognitive performance and the cognitive abilities it represents. He drew an analogy between a WM span measure and a task that involves thinking aloud, “they both impose cognitive demands on participants in relation to attentional control” and, more specifically, in the think-aloud task, “cognitive control may be critical in meeting the requirement to verbalize their thoughts while processing information derived from a given task (e.g., reading)” (p. 721). Goo further predicted that high-WMC individuals could better handle the additional task of concurrent verbalization “by utilizing their fine-grained executive attention processes (e.g., Kane et al., 2007) or a higher level of inhibitory control (e.g., Hasher et al., 2007)” (p. 721). But his prediction was not confirmed: in his experiment, somewhat unexpectedly, the high-WMC learners, rather than the low-WMC learners, were negatively affected in their rule learning.

2.1.6 Veridicality of concurrent verbal reports

To continue the discussion on veridicality, it is useful to clarify that Ericsson and Simon (1993) did not maintain the “absolute” completeness of CVRs. Instead, they insisted on and argued for CVRs being relatively comprehensive. That is, they are complete in relation to what should be reported, which is what is heeded and reportable. In fact, before their 1993 work, Ericsson and Simon (1979, as cited in Smagorinsky, 1989) had stated that informants did not report the full spectrum of their thought processes but paused frequently for as long as 15 to 60 seconds even when they were
prompted. The two occasions during which informants should have talked but refrained from talking, instead busying themselves with handling the primary task, included, according to Ericsson and Simon (1979, as cited in Smagorinsky, 1989, p. 468), the time when they were faced with high processing pressure (Durkin, 1937) and “when subjects attend to information that leads to direct recognition of appropriate action (Duncker, 1945)”. Ericsson and Simon (1993) further pointed out that in cognitively demanding problem-solving tasks, but for experimenter or other forms of prompting, subjects would often sacrifice the veridicality of CVRs by forgetting to report concurrently for the sake of concentrating on the primary task. Smagorinsky (1989), following Ericsson and Simon (1979), acknowledged the incompleteness, but recognized the value of pauses in CVRs to reveal the kinds of cognitive processes that went on. He quoted Hayes and Flower as saying:

“The psychologist’s task in analyzing a protocol is to take the incomplete record that the protocol provides together with his knowledge of the nature of the task and of human capabilities and to infer from these a model of the underlying psychological processes by which the subject performs the task. (Hayes & Flower, 1980, as cited in Smagorinsky, 1989, p. 469)

Smagorinsky (1989) argued for CVRs for their far fewer gaps than were found in other forms of data, given no absolutely complete data (see also Robinson, 2001). Although he acknowledged that “[t]he utterance … is more distilled, with the subject’s attention more focused on the writing task” and there is a time delay in completing the task, Smagorinsky defended the similarities of the writing processes reflected in CVRs to those “under natural conditions” (p. 474; italics original) and the resemblance of CVR conditions to “exam conditions” (p. 474; italics original) characteristic of in-school writing. Writing under exam conditions, he hypothesized, “include[d] processes similar to those found in writing done over longer periods of time” (p. 474; italics added). Smagorinsky’s (1989) defence for the validity of CVRs may also help dispel some of
the other early doubts concerning the veridicality of verbal reports, for example, Lyons’s (1986) concern with their completeness and Cooper and Holzman’s (1983) doubt regarding the trustworthiness of the inferences made from the pauses and the completeness of CVRs.

Ericsson and Simon’s (1993) validation of the veridicality of the TA method, based on their information processing framework, however, gave rise to some questioning. According to Payne (1994), one of the criticisms was the difficulty in observing and verbalizing higher-order thinking (Nisbett & Wilson, 1977), leading to the suggestion that “verbal data may reflect the norms for behavior in a task rather than providing a veridical report of the underlying processes employed in the task” (Payne, p. 245); that is, verbal reports might not coincide with behaviours.

In a more detailed fashion, Wilson (1994) critiqued the veridicality of verbal reports by pointing out the importance of what Ericsson and Simon deemed to be nonverbalizable processes/behaviours. Wilson admitted, “The concurrent, think-aloud measure is, perhaps, the best available measure of conscious, easily verbalizable thoughts” (p. 249) and “an excellent methodology to study the contents of consciousness” (p. 251). However, he argued that what Ericsson and Simon acknowledged could be missing from verbal reports, namely, nonconscious processing (e.g., automatic processes) and those thoughts that were difficult to verbalize, were not rare, and could be important, especially outside domains where verbal protocols were typically used and problem solving and expert reasoning were involved. Such absent processes or activities, Wilson pointed out, may include implicit learning, intuition, etc., to which may be added automated and some perceptual processes (e.g., Waern, 1988). While Wilson admitted that in problem solving tasks, the number of unconscious activities could be rare, which may have led to Ericsson and Simon’s insufficient consideration,
he doubted the plausibility of such an assumption, because this rarity was not to be assumed because of the absence of reference. He suggested that “if the goal is to study cognition that might involve nonconscious processing (e.g., automaticity, intuition), the completeness of verbal protocols cannot be taken on faith” (p. 251). As regards further research into the issue of completeness, Wilson recognized that the difficulty often lay in lack of an approach to telling how complete verbal protocols were.

Like Wilson, Russo et al. (1989) also believed that veridicality was difficult to measure, because the criterion for reference was not available. Russo et al. held that reactivity studies should take priority over veridicality studies because if thinking aloud changed thinking processes, veridicality studies would make no sense. However, the two facets of the validity issue may be closely related, because if informants only strategically, selectively reported their thoughts, or skipped over complex thoughts, the reactivity of TA would be reduced. For example, in L2 writing research using the TA method, informants might forget the requirements for thinking aloud and fail to report thoughts. They might report what they are going to write down in a way that resembles word-for-word translation to offset the pressure they face to complete TA while writing. As a result, the level of detail of information in CVRs is decreased and therefore the completeness of CVRs is sacrificed. Therefore, to some extent, the issue of completeness is also a precondition for the issue of reactivity.

2.1.7 Controversies over reactivity and veridicality in SLA and writing research

This section relates controversies over the validity of CVRs more specifically, first to SLA and then to writing research, as this issue has become an important concern in both areas. As the contents overlap, discussion on this issue in L2 writing research is dealt with within the subsection on writing research.
2.1.7.1 Controversies over reactivity and veridicality of CVRs in SLA research

So much caution has been expressed as regards the reactivity of CVRs in SLA research (e.g., Bowles, 2010a, 2010b; Cohen, 2000; Cumming et al., 2002; Ellis, 2001; Erdosy, 2004; Jourdenais, 2001; Wigglesworth, 2005) and their veridicality in this field (e.g., Cumming et al., 2002; Erdosy, 2004; Lumley, 2005; Smith, 2000; Weigle, 1994) that few studies used CVRs without mentioning the controversy over their use, especially their reactivity (e.g., Armengol & Cots, 2009; Cohen, 1991, 2000, 2013; Roca de Larios et al., 2001; Sasaki, 2000). Due to the possible reactivity and incompleteness of CVRs, Lumley (2005), contrary to Wolfe’s (1997) recommendation, opposed using CVRs for rater selection, training, and monitoring.

L2 Researchers have tended to emphasize the additional load that TA might incur in causing reactivity. Jourdenais (2001), for example, suggested that “the think-aloud data collection method itself acts as an additional task that must be considered carefully when examining learner performance” (p. 373). Ellis (2001) also mentioned the difficulty of “dual processing” in producing CVRs, especially for L2 learners (p. 37). Polio (2003) showed a different aspect of concern from the perspective of the reporting language. She was concerned that restricting the reporting language to either L1 or L2 exclusively would cause difficulty in reporting thoughts. This can be an important reason for participants usually being given the freedom to report in whatever language (for example, their L1 or L2) they felt comfortable with (e.g., Sachs & Suh, 2007). In the L2 literature, there have also been accounts of the possible facilitative effects of TA (e.g., Cohen, 2000; Jourdenais, 2001). For example, citing Swain’s (1985) Output Hypothesis, Jourdenais (2001) assumed that TA could help L2 learners learn by noticing the gaps in their interlanguage because the verbalization might act as input. The question remains as to how much L2 learners can learn from such input that *they themselves generate.*
The completeness of CVRs is another concern. For example, when using CVRs to explore raters’ decision making processes in scoring ESL compositions, Cumming et al. (2002) admitted that evidently, “a rater’s verbal report represented only a fragment of the aspects of the compositions that he or she had attended to” and the reports they collected contained “just glimpses of the criteria that the raters actually used to guide their scoring” (p. 71). They therefore warned that CVR data should only be used to make inferences about learner cognition rather than treated as a veridical representation.

2.1.7.2 Controversies over reactivity and veridicality of CVRs in writing research

There have been cautions or observations that CVRs might change or alter thinking processes in writing and that CVRs might not be complete. For example, Perl (1980a) admitted that “it is conceivable that asking students to compose aloud changes the process substantially, that composing aloud is not the same as silent composing” (p. 19). However, while Perl (1980b) reiterated the potential problem of alteration, she emphasized that CVRs helped them “detect certain basic patterns” by “capturing some of the flow of composing” (p. 364).

Conceiving of writing as a cognitively demanding task, Faigley and Witte (1981) also remarked,

Verbal protocols require writers to do two things at once—they must write and they must attempt to verbalize what they are thinking as they pause. Perhaps some subjects can be trained to do both tasks with facility, but many writers find that analyzing orally what they are doing as they write interferes with their normal composing processes, interrupting their trains of thought. Many activities in writing occur simultaneously—from unconscious processes such as ordering the words in a noun phrase to conscious processes ranging from spelling to planning and monitoring. A lot is going on and not all of it gets verbalized. In addition, what writers can verbalize about their composing processes probably is influenced by the nature of the writing task. Unskilled writers who seem bound by concerns about conventions when arguing for mass transit may show very different concerns when writing to a personal friend. (p. 412).

In the above quote, Faigley and Witte raised both the reactivity and the nonveridicality issues. As for reactivity, they were concerned with the cognitive
pressure of processing dual tasks with thinking aloud playing a part, for which they provided observational evidence from writers’ experiences. They also suggested individual differences in adaptation to TA. In terms of the completeness of CVRs, they noted that there were different types of thoughts that occurred in writing, including unconscious ones, and CVRs may not reflect all of them. Faigley and Witte also expressed concern about the effects of writing task type on the contents of CVRs. This suggests the interaction of writing task type with reactivity given the relevance of task type with reportable and heeded information.

Later, when again using CVRs to investigate pre-texts in composing, Witte (1987) raised his concern that CVRs could contribute to an increased number of pre-texts and would “encourage writers to produce more complete or better-formed pre-texts than they would under a silent condition” (p. 399). He also predicted that the impact of think-alouds on the magnitude and extensiveness of pre-text might be contingent upon individual writers and writing tasks, therefore relating the issue of reactivity to individuals and task type.

Ransdell (1995) also acknowledged that TA would impose particular constraints on writing processes. She treated writing as “essentially a multitude of complex problem-solving tasks that must be coordinated in space and time”, which, she suggested, would make writing processes more “susceptible to tasks that include additional time constraints” than cognitive processes that involved fewer subtasks (p. 90).

The above concerns over the issue of reactivity of TA are paralleled by arguments for its nonreactivity. Besides writing researchers who believed that writing could be treated as a cognitive problem-solving activity, similar to the cognitive problem-solving tasks used in cognitive psychology studies that Ericsson and Simon (1993) reviewed (e.g., Breetvelt et al., 1994; Flower & Hayes, 1980b), some researchers showed their
confidence about the nonreactivity of CVRs in writing studies. In Flower and Hayes’s (1981a, 1984b) studies on composing processes using CVRs, for example, the underlying assumption was that reporting contents in STM was easy, given their being in an orally compatible form. Whalen & Ménard (1995) also argued that,

[A] wide variety of verbal protocol data (Cumming, 1988; Ericsson & Simon, 1984; Flower & Hayes, 1983[b]) has amply shown that verbalizing thought processes in no way deforms participants’ natural processing. Nor was there any evidence that the thought processes of our 12 participants were altered or distorted by verbalizing during text production. (p. 387).

Writing researchers have also voiced their concerns over the completeness of CVRs. Durst (1987), for example, while using a think-aloud method to obtain online thinking data during L1 writing, stated that “not all composing processes are revealed in CVRs” (p. 355). He administered retrospective interviews asking students questions about their writing behaviours as a complementary measure.

On veridicality research, both Russo et al. (1989) and Wilson (1994) mentioned the absence of a reference benchmark. When talking about the completeness of CVRs, Wilson suggested, “Ideally, experimental manipulations should be included to assess the extent to which people heed information that does not appear in their protocols” (p. 251). As for experimentation, Robinson (2001) has pointed out the relative ease of investigation into the veridicality issue on well-defined tasks (e.g., arithmetic problems of subtraction), because they “have a finite set of correct sequences or a finite number of solution paths for getting the correct response (Austin & Delaney, 1998; Critchfield & Epting, 1998)” and “converging evidence is available to assess verbal report validity” since “specific patterns of accuracy and latency data can be associated with specific strategies because children’s performance on arithmetic tasks has been well documented (Geary, 1994)” (p. 212).
As far as ill-defined tasks such as writing are concerned, where there appears no absolute benchmark (i.e., the real, complete sequence of thoughts) available to explore the extent of completeness of CVRs, comparisons could perhaps be made by referring to data elicited by other means, which are relatively valid, for example, RVRs, including stimulated recall. Though RVRs may suffer from memory decay, if implemented immediately after task completion, they may provide (as others might argue, part of) additional information that can be reliably used as a means for cross-checking. For example, in writing research, informants may be required to report only certain behaviors or processes of interest, for example, revisions, retrospectively, in the sense of offering immediate directed verbal reports, with the least intrusion from experimenters. The directed retrospective reports on behaviors or processes that have required much control or effort may be trusted, though the reports may not be complete as a whole. Especially, the contents of CVRs can be more directly examined by comparing them with writers’ controlled behaviors shown in video tapes. Controlled behaviors need to be completed with cognitive effort, and within conscience (e.g., LaBerge, 1981; Schmidt, 1992); therefore, in CVR that requires participants to report whatever occurred in their mind, these behaviors need to be reported given that they are reportable and heeded. Other methods for comparison of a similar kind can include using videotaping and keystrokes logging.

2.2 Empirical Studies on Reactivity

This section first reviews empirical studies on reactivity and then reviews empirical studies on veridicality. In reviewing the former, it first briefly summarizes studies using non-verbal and L1 verbal tasks (non-SLA studies) in cognitive psychology, then selectively reviews and synthesizes studies using different L2 verbal tasks (SLA studies),
and finally, and most importantly, examines and critiques reactivity studies concerning both L1 writing and L2 writing.

### 2.2.1 Non-SLA studies using non-verbal and verbal tasks

Research on the effects of verbal reports on problem-solving and decision-making tasks has been conducted in the field of cognitive psychology since at least the 1950s (Bowles & Leow, 2005). These studies typically compared the performance of a think-aloud group or sometimes several think-aloud groups with that of a silent group to see if there were any significant differences in performance that could be taken as evidence for reactivity. The studies mostly examined accuracy in performance and utilized non-verbal tasks, and involved either a nonmetacognitive or a metacognitive group, or both nonmetacognitive and metacognitive think-aloud groups, in comparison with a silent group. Overall, in Ericsson and Simon’s (1993) and Fox et al.’s (2011) reviews, no reactivity was found for nonmetacognitive think-alouds but reactivity was found for metacognitive think-alouds, with both types of think-alouds tending to delay task completion and increase overall solution time, as compared with the silent condition, because of the additional time needed for verbalization. The effects of metacognitive verbalization were found to be facilitative in some studies while detrimental in others. One study conducted by Robinson (2001), for example, examined whether Grade 1, Grade 3, and Grade 5 children, who were asked to solve subtraction problems of different types/supposed difficulties and at the same time report their thoughts/strategies concurrently, showed any statistically significant differences in accuracy and latencies of performance in comparison to a silent group of their peers. The reactivity of CVRs was also examined by comparing their concurrently reported strategy use with the strategy use retrospectively reported by another group of their peers. With the condition confound taken into consideration, Robinson found no reactive effects on the accuracy
measure, no latency effects, and no differences in reported strategy use, except that Grade 3 children spent a considerably longer time solving the problems, than did both their silent and retrospectively reporting peers.

As for non-SLA studies that used verbal tasks, Bowles and Leow (2005) were able to identify six of them, which used such tasks: anagrams (Lass et al., 1991; Russo et al., 1989), artificial grammar activities (Mathews, et al., 1989), sentence assembly and word puzzles (Rhenius & Deffner, 1990), verbal analogies (Short et al., 1991), and revision of text (Stratman & Hamp-Lyons, 1994). No reactive effects on accuracy in performing verbal tasks (in the case of Russo et al., 1989, the anagrams) were found, except that the 94 fifth graders in Short et al.’s (1991) study performed better on verbal analogies while thinking aloud metacognitively. As for time on task, Lass et al. (1991) and Rhenius and Deffner (1990) reported latency, and Russo et al. (1989) found increased time on task, while the rest of the three studies (Mathews et al., 1989, Experiment 2; Short et al., 1991; Stratman & Hamp-Lyons, 1994) did not report time, making the reactive effects on time inconclusive. The limited number of reactivity studies using L1 verbal tasks, which used different populations and perhaps dissimilar tasks (not strictly comparable), make it premature to draw sound conclusions. Considering the interaction between reactivity and task type suggested by researchers (e.g., Deffner, 1989; Russo et al., 1989), there is an evident need for more empirical studies using L1 verbal tasks. There is also a clear paucity of endeavour to establish the relevance of L1 studies to L2 ones.

2.2.2 SLA studies

As far as I am able to identify, there are 13 SLA studies on reactivity to date. Six of them used L2 reading tasks (Bowles & Leow, 2005; Goo, 2010; Leow & Morgan-Short, 2004; Polio & Wang, 2005; Rossomondo, 2007; Yoshida 2008), three employed writing tasks (Sachs & Polio, 2007; Yang et al., 2014; Yanguas & Lado, 2012), and four used
tasks other than reading and writing (Barkaoui, 2011; Bowles, 2008; Sachs & Suh, 2007; Sanz, Lin, Lado, Bowden, & Stafford, 2009). The total number of 13 SLA studies is based on Bowles’s (2010b) pool of nine SLA studies on reactivity, plus four updates (Barkaoui, 2011; Goo, 2010; Yang et al., 2014; Yanguas & Lado, 2012). As can be seen, there actually are not many SLA studies on reactivity. In what follows, the 10 SLA studies using tasks other than writing are selectively reviewed, but the three empirical studies on L2 writing are considered in the section reviewing empirical studies on writing, following a review of studies on L1 writing.

2.2.2.1 Studies using L2 reading tasks

Studies using L2 reading tasks have explored the effects of CVR on reading comprehension, and/or embedded grammar learning, and have touched upon the mediating roles of learner proficiency, task type, and WMC. Leow and Morgan-Short (2004) was the first SLA study that explored the issue of reactivity of CVR in L2 reading. In this study, reactivity was operationalized as the significant differences between a nonmetacognitive thinking aloud group and a silent group in their performance in terms of reading comprehension and concomitant L2 grammar development. In this experiment, 77 beginner English-native learners of Spanish each read, either silently or while verbalizing nonmetacognitively, a text in Spanish that contained 17 occurrences of the targeted form, the impersonal imperative. Participants’ text comprehension, intake, and controlled written production of the targeted form were measured by an 11-item comprehension task, a multiple-choice recognition task, and a fill-in-the-blank task, respectively. Comparisons of the performance of the ± think-aloud groups found no statistically significant differences in their scores in these three aspects, indicating no reactivity for concurrent nonmetacognitive verbalization. Time on task was not measured. One limitation of this study, as the researchers admitted, was the constraints
imposed by the specific morphological form targeted, which may limit the applicability of the findings to other linguistic forms that may possess different degrees of saliency or assume different degrees of meaning. Another limitation to the findings lay in the level of reading proficiency demonstrated by the first-year participants, who differed from higher-level learners in terms of prior knowledge and the strategies used. Leow and Morgan-Short also suggested the roles of length of text and the times of exposure as two potential reactivity-related factors to be empirically tested. Leow and Morgan-Short conducted pretests using the same recognition task and controlled production task as were used in posttests, but they could also have tested the participants’ comprehension prior to the reading task.

Leow and Morgan-Short’s (2004) study was replicated by Polio and Wang (2005), who investigated how CVR might affect advanced learners’ reading comprehension and subsequent learning. In this experiment, 30 advanced-proficiency Chinese L1 learners of English read a passage that contained frequent and infrequent phrasal verbs as target forms under ± think-aloud conditions. The same types of post-tests were administered measuring the participants’ reading comprehension, written production, and recognition. The test scores were analysed using nonparametric statistical tests. Significant difference was found only in the comprehension measure, with silent readers comprehending better than verbalizing readers ($p = .01$), indicating detrimental reactivity. Analyses of participants’ verbal reports showed that the participants in this experiment might be using more complex strategies than the translation strategy that dominated the participants’ reading process in Leow and Morgan-Short’s study. In the former case, verbalization of on-going thoughts could be more demanding than it was in the latter case, hence the detrimental reactivity in this study. The limited number of
participants may affect the power of analyses in this study, and it remains unclear why the reactivity did not extend to the other two measures.

Yoshida (2008) explored the possible effects of concurrent verbalization in completing different types of reading tasks. Sixty-four Japanese English majors each completed one of the three types of reading tasks (i.e., read-only, read-and-answer-questions, and read-and-write-outlines) under a think-aloud or non-think-aloud condition, before they were asked to recall the propositions embedded in the passage twice, once immediately and once a week later. The think-alouds contained metacognitive elements since participants were also asked to explain the answers they gave to the questions attached to the reading passage and the outlines they wrote while reading. The two-way analysis of variance (ANOVA) performed on the participants’ recall scores revealed that neither reading condition nor task type exerted any statistically significant main effects on immediate or delayed recall, neither was there any significant interaction between reading condition and task type. However, another two-way ANOVA run on the time spent indicated that TA did cause a significant slowdown. Yoshida attributed the nonreactivity of TA to its insufficient cognitive demand and the possibility of its potential detrimental effects on memorization due to the additional load of verbalization being cancelled out by its possible memory-enhancing effects derived from “reading aloud difficult parts repeatedly” (p. 206). Yoshida also mentioned the effects of the anxiety-releasing collective think-aloud setting in a laboratory, where “participants may have been more relaxed than when individually conducted and participants did not have to feel much pressure about their performance on the think-aloud task” (p. 207). The merits of this study included its pretesting, its interest in reactivity with task type, and its measuring of recall of
propositions. But still the limited number of participants in each of the six segments of group assignment might affect the validity of its results.

Goo (2010) explored how WMC (WMC) might interact with reactivity. The WMC of 42 English-speaking learners of Spanish was measured with a listening span task and an operation span task. A fill-in-the-blank written production pretest was also administered to ensure that all participants were similar in their lack of acquisition of the targeted form, the Spanish immediate future. In the experiment sessions, they engaged in a reading task featuring the targeted form, under two conditions: TA and non-think-aloud (NTA), and their performance on reading comprehension and written production was then tested. Different from the previous findings of nonreactivity on reading performance reported by Leow and Morgan-Short (2004), Bowles and Leow (2005), and Rossomondo (2007), a significant detrimental effect on overall learner reading comprehension was found in this experiment. In the written production posttest, which gauged rule learning, no overall effect of TA was found but a differential effect stood out—thinking aloud significantly hampered high-WMC learners’ learning of the form while reading the passage but did not affect low-WMC learners. Goo attributed his finding of reactivity on reading comprehension to different target structures, which might “affect the level of difficulty of a given text and complexity of sentences included in the text” (p. 735). He also explained the high-WMC group’s vulnerability to TA in terms of concomitant form learning as the result of a mechanism characteristic of their “fine-grained controlled processes” that “enabled them to concentrate, with great intensity, on retrieving goal-relevant information (e.g., contents of the text) during the comprehension test, simultaneously suppressing goal-irrelevant information (e.g., grammatical information on the target structure and related target items) incorporated in the text used in the reading activity that preceded the comprehension test” (p.
In other words, the secondary task did not benefit from the cognitive advantage that high-WMC learners had, but would be sacrificed, to some extent, in the case of the additional requirement of concurrent verbalization.

2.2.2.2 Studies using L2 tasks other than reading and writing

Four studies used tasks other than reading and writing, including essay rating (Barkaoui, 2011), learning to play a computerized game (Bowles, 2008), story-retelling in a synchronous CMC context (Sachs & Suh, 2007), and taking a computerized lesson (Sanz et al., 2009). An example was Sanz et al.’s (2009) study. The researchers conducted two experiments which were similar in these features: 1) the equal number of participants ($N = 24$) and the equal assignment of groups ($N = 11$ for the verbalization group; $N = 13$ for the silent group); 2) the targeted form (i.e., the assignment of semantic function to noun phrases in Latin sentences; 3) the context (i.e., a computerized lesson); 4) the same assessment materials, and 5) the use of reaction times. The two experiments differed primarily in the removal of a grammar lesson component from Experiment 2.

Analyses of aural interpretation, grammaticality judgment, and production scores from the pretests and the posttests using $2 \times 2$ repeated-measures ANOVAs ($\text{Group} \times \text{Time}$) showed divergent results between the two experiments. In Experiment 1, ANOVA results on the three tests showed neither main effects for Time and Group, nor significant interaction between Time and Group. In Experiment 2, however, significant interaction was found between Time and Group on both the grammaticality judgment and production tests, $F(1, 23) = 4.800, p = .039$, and $F(1, 23) = 5.010, p = .035$, suggesting positive effects. Further analyses of the participants’ responses to the old items that appeared in the treatment and the new items in all the three tests using repeated-measures ANOVA procedures revealed that facilitative effects occurred in production of the old items and in grammaticality judgment of the new items. Reaction times, in
Experiment 1, analysed with a $2 \times 2$ repeated-measures ANOVA, showed a significant interaction between Time and Group, $F(1, 21) = 7.170, p = .014$, only for latency scores from the grammaticality judgment tests, suggesting a retardative effect on subsequent task performance. Nevertheless, in Experiment 2, no effects of TA on the speed of subsequent test taking were found across the three tests.

The researchers explained the temporal effects in Experiment 1 as the result of TA “favour[ing] the development of a strategy to slow down and reflect on the nature of the language” (p. 62), and the significant facilitative effects of TA in Experiment 2 as representing its more salient effects in increasing awareness when interacting with a less explicit instructional training condition than when interacting with an explicit instructional training condition, as was the case in Experiment 1. They supported their finding of positive reactivity by turning to Russo et al.’s (1989) summary of reactivity-causing factors, as regards the roles of increased reflection and auditory input, and also to Swain’s (1985, 1995, 2000) output hypothesis, which suggested the roles of language production (in this case, TA) in promoting learning by providing additional input and developing awareness. Sanz et al. concluded that “reactivity depends not only on the task and on the nature of the assessment tool but also on the nature of the dependent variables (i.e., latency vs. accuracy)” (p. 34). Their suggestion was that TA should be used cautiously. The study has considerable merit, as seen in its employment of a variety of tests and reaction times as measures and its examination of the issue of reactivity in a pedagogical setting, although the small sample sizes are a major issue.

2.2.2.3 Summary

In summary, among the 10 SLA studies that utilized tasks other than writing, five reported that nonmetacognitive thinking aloud generated reactive effects (i.e., Barkaoui, 2011; Goo, 2010; Polio & Wang, 2005; Sachs & Suh, 2007; Sanz et al., 2009), while
the other five reported no reactivity for this type of verbalization (i.e., Bowles, 2008; Bowles & Leow, 2005; Leow & Morgan-Short, 2004; Rossomondo, 2007; Yoshida, 2008). The reactive effects reported included negative effects (i.e., Goo, 2010; Polio & Wang, 2005), positive effects (i.e., Sachs & Suh, 2007; Sanz et al., 2009), a mixture of both positive and negative effects or effects of an unidentified nature (i.e., Barkaoui, 2011). The negative effects were found on reading tasks, in terms of decreased reading comprehension for the TA group (Goo, 2010; Polio & Wang, 2005) and, additionally, hindered rule learning, especially for high-WMC think-aloud learners (Goo, 2010). The positive effects were found on a story-retelling task in a synchronous CMC context, in terms of text completion (i.e., Sachs & Suh, 2007), and in completing a computerized lesson, benefiting accuracy of grammaticality judgment (of new items) and production (of old items) (i.e., Sanz et al., 2009). Changes in thinking processes and/or alterations of performance were also reported in implementing an essay rating task (Barkaoui, 2011). These mixed results seemed to confirm Russo et al.’s (1989) prediction that reactivity might be task- and measure-dependent. Regarding time on task or reaction times, the findings tended to be consistent, suggesting that thinking aloud increases time spent on task (Bowles & Leow, 2005; Yoshida, 2008) or elongates reaction times on a grammaticality judgment posttest (i.e., Sanz et al., 2009, Experiment 1). This is consistent with Ericsson and Simon’s (1993) prediction and conclusions.

There are several concerns with these studies reviewed above. The two studies using reading tasks that reported detrimental effects on the primary measure of reading comprehension (Goo, 2010; Polio & Wang, 2005) involved small sample sizes. The two studies that reported facilitative effects also suffer from the same problem (Sachs & Suh, 2007; Sanz et al., 2009), and the effects were found on a secondary measure of text completion rather than the primary measure of story-retelling (Sachs & Suh, 2007). In
contrast, one study using a reading task that did not find reactivity (i.e., Leow & Morgan-Short, 2004) had a more robust sample size, and another study that reported no reactivity on performance but used a L2 problem solving task also had a large sample size (i.e., Bowles, 2008), though more studies reporting facilitative effects suffered from the problem of limited participants (Bowles & Leow, 2005; Rossomondo, 2007; Yoshida, 2008), despite the insights they provided. Also, among the 10 studies, some were not dedicated studies (i.e., Sachs & Suh, 2007; Rossomondo, 2007; see also Sachs & Polio, 2007, reviewed below). Several studies explicitly included a metacognitive think-aloud group (i.e., Bowles, 2008; Bowles & Leow, 2005), but other studies had a blurred distinction between nonmetacognitive and metacognitive think-alouds (Sanz et al., 2009; Yoshida, 2008), making it hard to draw a specific conclusion. In the following section, I shall review studies on reactivity in the field of writing, to see if there is a different picture on this issue with regard to the writing task. I shall start with reactivity studies in L1 writing first before I present the details of L2 writing.

2.2.3 Reactivity studies in L1 writing

There were few empirical studies on the reactivity of CVR in L1 writing. The earliest might be the case study conducted by Berkenkotter (1983), which was commented on by Smagorinsky (1989), who compared writing under natural conditions and under TA. Her participant was an expert writer, Donald M. Murray, who wrote first at home for two months with a tape recorder on whenever composing, then in a laboratory setting, wrote for an hour on a topic (“Explain the concept of death to 10-12 year old readers of Jack and Jill magazine”) and provided CVRs, with the researcher present and a recorder switched on, and, finally, under the same laboratory conditions, revised an article within two days. While Murray’s writings were successful at the home setting, his writing was obviously not as fruitful and smooth (i.e., he produced only 17 words in an hour) when
he thought aloud under the laboratory setting. He said “I have rarely felt so completely trapped and inadequate. To find equivalent feelings from my past I would have to go back to combat or to public school” (Murray, 1983, p. 169). In interpreting his struggle in writing under the CVR condition, Smagorinsky doubted if his difficulty in writing could have been largely attributable to this specific topic. He proposed that “A more logical approach would have been to give Murray the same task (i.e., write an article, write a personal narrative, state and defend an opinion) to perform under both conditions” (Smagorinsky, 1989, p. 471)

To my knowledge, there are four empirical studies that explored the effects of TA on L1 writing processes or L1 writers’ performance (i.e., Janssen et al., 1996; Levy & Ransdell, 1995; Ransdell, 1995; Stratman & Hamp-Lyons, 1994). Stratman and Hamp-Lyons (1994) conducted a pilot study, in which 12 graduate and undergraduate university students were engaged in two revision tasks, one with think-alouds and the other silently, separated by an interval of several weeks. The two faulty texts employed in the study were similar (e.g., the syntax of each sentence, words of the same syllables and familiarity, the same type and number of mistakes placed at the same locations, etc.) and had the same “comparison-contrast” structure (i.e., the block pattern). The researchers examined differences in the number of revisions made between the two conditions against two broad categories of error detection/removal and content changes. The results were mixed. Stratman and Hamp-Lyons simply compared the number of revisions made between the TA and non-TA condition. There was no statistical analyses of their results, perhaps due to the small number of participants ($N = 12$). The rather mixed results, therefore, should not be taken as strong evidence for or against the reactive effects of CVRs. The salience that was given to revision in an experimenter-designed revision task also reduced the relevance of their findings to on-line revision,
which may not receive as much emphasis in on-line writing tasks.

Among the four empirical studies, three asked students to compose in an L1. Ransdell (1995) examined the effects of thinking aloud on such performance indexes as words per minute, the total number of words, the average length per clause, the total number of clauses, and clauses produced per minute, as well as intrusions, which were defined as “any comment[s] that did not fit the content, genre, or topic of the letter the subject was writing” (p. 93). In her experiment, 38 students each were required to complete a letter on a computer describing the first days of school, under three conditions, TA, retrospective-replay, and silent, with the orders of these conditions counterbalanced. Ransdell found that TA retarded production of words per minute, when compared with the no-protocol condition and the retrospective-replay, and production of clauses per minute, when compared with the retrospective-replay. Significantly more intrusions were also detected for the TA condition \((M = 27)\) than for both the retrospective condition \((M = .06)\) and the silent writing condition \((M = .02)\). She drew the conclusion that TA would not substantially change writing processes and the thought sequences of writing. She explained the more intrusions under the TA condition as participants “simply reporting on what is unusual about a task”, indicating that “CVRs are no more intrusive than other experimental procedures” (p. 96). She supported Ericsson and Simon’s (1993) predictions as regards the nonreactivity of Level 2 think-aloud and believed that CVRs could offer valid bases for establishing writing process models. She pointed out that because the writing task employed in her experiment was easy, the pressure of dual processing of the writing task and verbalization was displayed in the form of a reduction in processing speed, but the impact of hearing one’s voice should not be excluded. She speculated that the reactivity of CVRs might more obviously be seen in complex writing tasks. Some of the limitations of her study included the short
duration of concurrent verbalization (only 12 minutes) and the different writing tasks (i.e., three segments of the letter) under the three conditions. Besides, more dimensions of writing should have been considered, including more indexes of fluency (e.g., dysfluencies), complexity (e.g., general or subclausal complexity) and accuracy.

Another study, conducted by Levy and Ransdell (1995), examined changes in the effort spent on writing subprocesses under a think-aloud condition. This study involved 10 university students writing over a period of 12 weeks. The participants wrote in the Windows Notepad interface, timed digitally. A monitor zone on the screen generated ticks and recorded the reaction time. SimulScan was additionally used to screen-record the course of writing and record think-alouds. The experiment lasted from Week 2 to Week 11. The writing topics were open, to be completed within 40 minutes, for submission to a magazine. At the beginning of each week, participants’ reaction times were tested while they pressed a footswitch, and these times were averaged. During Week 2 and Week 3, participants wrote to familiarize themselves with the experiment environment. During Week 4, they practiced responding to ticks by pressing the footswitch while they wrote. During Week 5, they were trained on thinking aloud as an addition requirement. In the following six weeks, they were asked to write, think aloud, and respond to ticks. The last writing session had only 20 minutes for writing, but they needed to report their writing activities retrospectively. Altogether eight weeks (Week 4-Week 11) of the participants’ writing activities provided data for analyses. The effort expended for a certain subprocess of writing was inferred by calculating interference response times (IRTs) (i.e., the differences between the actual reaction times in writing and the average reaction time. Some research assistants coded typing, revisions, pauses (positions and durations), movements of the cursor, and paragraphing from video recordings while turning off the audio, whereas others watched the video recordings
with the audio on, to code writing activities such as planning, rereading, speaking, writing content and writing the same content. The researchers then decided on the categories of writing subprocesses. Then, the starting and ending points of these subprocesses were determined, and the subprocess-tick correspondences were also worked out. The researchers drew the conclusion that the effects of TA were negligible because, in planning and text generation subprocesses, the differences between the IRTs during Week 4 and those during Week 5 were lower than 50 milliseconds and were not statistically significant, indicating no extra effort exerted on these two subprocesses with the additional demand of verbalization. Besides, there was no difference in the number of words produced during text generation, across the two conditions, within one writing session, between different sessions, or within the first 10 minutes of a session. The researchers observed that, although participants found themselves somewhat uneasy with think-alouds at the beginning, during Week 5, they were able to write as usual. This study was not one dedicated to the issue of reactivity. Conclusions drawn by comparing 10 participants’ IRTs in two weeks and comparing the number of words the 10 participants wrote for different writing tasks may have questionable reliability. The researchers also could have measured more dimensions of writing performance.

Presenting an opposing view, Janssen et al. (1996) concluded that TA disturbed normal writing processes based on an examination of changes in the real-time organization of writing. Their data were pause duration in combination with pause location in the text and in the process. The data were collected with the resident software program Keytrap, which registered all keystrokes and all pauses between keystrokes when the participants wrote on a computer word processor (WordPerfect 5.1). They conducted two experiments, both of which involved two writing tasks and two writing conditions (± think-aloud), administered in a Latin-square design, with task and
condition alternated for the second writing session of each experiment. Experiment 1 involved 20 students, within one hour, writing two business reports each of about two pages’ length, in two writing sessions separated by approximately two weeks. All participants were familiar to the genre, experienced in using WordPerfect, and wrote using written instructions, which detailed the aim, the audience and the context of the business reports. In Experiment 2, 28 different students were asked to write two simple explanatory texts to explain some Dutch events or customs. For this experiment, the time set for writing both texts was about 20 minutes and there was an interval of about one week between the two writing sessions. Different from Experiment 1 which measured pauses within sentences, between sentences and between paragraphs, Experiment 2 replaced the measure of pauses within sentences with the measures of pauses within word groups (e.g., verbal phrases) and between word groups, and set the minimum pause length as 0.5 seconds, in contrast to that of 1.3 seconds set in Experiment 1, to achieve finer measurements. Both experiments showed that in the TA condition, pause time increased significantly on almost all levels of measurement. This finding corroborated Ericsson and Simon’s prediction on a temporal effect of CVRs. Also, by relating these pauses in different locations to different levels of planning or processes of monitoring and planning to revise, the researchers further concluded that thinking aloud disturbed normal writing processes. The researchers also suggested that the reactivity of TA differed with writing tasks, since they discovered larger effects in complex business letter writing than in simpler description of events. One problem with this study is the relevance of pauses to reactivity. Questions that can be asked include whether longer pauses in typing on the keyboard indicate changes in the structure or the sequence of writing processes, during which participants might well engage in recoding of their thoughts and verbalizing them.
The three reactivity studies on L1 writing employed different writing tasks (i.e., writing a letter, writing around open topics for submission to a magazine, writing two business reports, and writing two explanatory texts). Their designs or approaches to investigation also differed. Ransdell (1995) adopted a repeated-measures design operationalizing reactivity as differences in several performance indexes. Levy and Ransdell (1995) integrated reactivity research into the long stretch of continuous writing sessions, focusing on changes in a temporal measure, IRTs, in relation to writing subprocesses, while looking at the number of words produced. Finally, Janssen et al. (1996) measured pause lengths and pause locations pertaining to different levels of writing subprocesses and operationalized reactivity as significant changes in pause duration. The former two studies pointed to little reactivity except for a slower speed of production (i.e., Ransdell, 1995) or slightly more cognitive effort taken (i.e., Levy & Ransdell, 1995). Levy and Ransdell, especially, observed that the reactive efforts could be minimized or eliminated with increased practice. In contrast, Janssen et al. (1996) concluded that concurrent verbalization disturbed writing processes, although their measurement of pauses was still temporally related, and suffered from the indefinite nature of pauses. Regardless of the seemingly mixed findings, there was an overall tendency for TA to affect writing temporally (e.g., Ransdell, 1995; Janssen et al., 1996). Further L1 investigations into this issue are clearly needed given the small number of L1 reactivity studies and some of the drawbacks to these previous studies (e.g., notably, the small number of participants in Ransdell and his colleague’s studies and the ambiguous association of changes in pause features with interference of writing processes in Janssen et al’s study). How TA might occur with L2 learners and how reactivity, if there is any, might differ in L2 from L1 is also worthy of exploration.
2.2.4 Reactivity studies in L2/bilingual writing

Three empirical studies are identified that concerned L2 writing (Sachs & Polio, 2007; Yang et al., 2014; Yanguas & Lado, 2012), all of which operationalized reactivity in terms of performance indexes. The first study on reactivity in L2 writing was conducted by Sachs and Polio (2007), who examined, as an additional interest, how TA affected L2 writers’ use of teacher feedback.

As a dedicated reactivity study, Yanguas and Lado (2012) investigated the effects of TA on writing in a heritage language. The participants were 37 bilinguals of English and Spanish, the latter being their heritage language. All participants were enrolled in two sections of the first of a sequence of three Spanish language programs. The study adopted a quasi-experimental design, with one section of the students forming a TA group ($N = 20$) and the other a no-think-aloud (NTA) group ($N = 17$). The writing task used was a semi-guided writing task, to be completed within 25 minutes, by both the think-aloud group and the silent group. The reactivity of CVR was operationalized as the statistically significant differences between the ± think-aloud group in terms of three categories of performance indicators, namely, fluency (the number of words and the number of words per T unit), accuracy (error-free T units) and lexical complexity (Uber Index, Jarvis, 2002). The scores of these four dependent variables were compared across groups using a series of one-way ANOVAs. The results of the analyses indicated statistically significant differences between the ± think-aloud groups in terms of both accuracy ($p = .005$) and lexical complexity ($p = .052$), the effect sizes being medium ($d > 0.50$), with the TA group surpassing the non-TA group. Both groups, however, performed equally well in terms of fluency. These results suggested facilitative reactivity overall. Some drawbacks of this study included its lack of pretests, especially considering the diversity of students’ language background and possibly their varying
command of Spanish, and its sample size. Moreover, the researchers quoted some content from think-alouds as evidence for participant gains from the act of verbalization. Such evidence is questionable because it remained unknown whether silent L2 writers had similar thoughts.

Yang et al. (2014) is the first empirical study exploring the reactivity issue in English-as-an-L2 writing where English was the participants’ foreign language. They followed previous approaches, inferring reactivity based on performance indexes, but included a wider range of measures than were examined in both Ransdell (1995) and Yanguas and Lado (2012). The participants were 95 non-English-major sophomores. Each of them first wrote an argumentative essay. They were then randomly assigned to three groups: TA, metacognitive TA, and silent, and wrote another argumentative essay on a topic similar to the one used for the baseline writing, under their assigned conditions. During the former two conditions, they wrote individually in a room with a research assistant present for training, recording, or giving prompts when necessary. The results of a series of analysis of covariance (ANCOVA) indicated that the effects of the two types of verbal reporting were similar, with the metacognitive type causing larger effects. Both types of reports caused significantly more dysfluencies, and demonstrated a tendency to cause a decrease in the number of different verb forms used. In terms of temporal measures, metacognitive verbal reporting significantly increased time on task and reduced the speed of production, while nonmetacognitive verbal reporting did not have such effects.

Yang et al. (2014) concluded that TA might cause some reactivity in L2 writing, but this effect is far from serious. Notably, additional aspects of L2 writing performance should have been included in the analysis, for example, the number of errors, length, organization, content, etc., to provide a finer and more comprehensive measure of the reactive effects. Participants’ perceived reactivity should also have been analysed.
The two existing studies involving actual writing (Yanguas & Lado, 2012; Yang et al., 2014) used participants of rather divergent backgrounds (i.e., diverse origins with more than 56% having Central American origin versus Chinese) writing in a language of different affinity to them (i.e., Spanish as the heritage language versus English as the foreign language) and yielded quite different results (facilitative versus detrimental) to different dimensions of writing performance. Overall, it can be seen that the number of reactivity studies in L2 writing is still very limited, which perhaps precludes a judgement.

2.2.5 Summary

In her meta-analysis, Bowles (2010b, p. 121) proposed that “the answer to the question of reactivity and think-alouds is not unidimensional but rather is dependent on a host of variables”, including non-individual and individual factors, but she concluded that “compared to participants completing the same tasks silently, participants who think aloud tend to perform only slightly better or slightly worse on posttests”. As for time on task, she identified a rather consistent increase, though the effect sizes could range widely.

The three studies which were not included in Bowles’s (2010b) meta-analysis that adopted a between-groups design and included statistical analyses (i.e., Goo, 2010; Yang et al., 2014; Yanguas & Lado, 2012), to some extent, violated her concluding remarks. Goo’s (2010) study reported negative reactivity on L2 learner’s reading comprehension, with a medium effect size, (TA vs. NTA), \( t(40) = -1.99, p = .054, d = .62 \). Yanguas and Lado (2012) reported facilitative effects of thinking aloud on two measures of heritage language writing, error-free T units, \( F(1, 36) = 8.896, p = 0.005 \), and lexical variety, \( F(1, 36) = 4.057, p = 0.052 \), both with a medium effect size \( (d > 0.50) \). Yang et al. (2014) found a detrimental effect of nonmetacognitive verbal
reporting on a fluency measure, dysfluencies, $p = .050$, with a medium effect size, $d = 0.58$.

To date, only three studies can be identified that have touched upon the relationship between factors of individual difference in expertise and reactivity (Barkaoui, 2011; Goo, 2010; Polio & Wang, 2005). Polio and Wang (2005) explained the negative reactivity they found on reading comprehension of advanced-proficiency Chinese learners of English as the result of their more complex reading strategies, in contrast to the translation strategy predominantly employed by the low-proficiency learners in Leow and Morgan-Short’s study. This may indicate that high proficiency readers are more easily affected by verbalization. Polio and Wang’s finding seems to align with Goo’s (2010) finding that high-WMC learners seemed to be affected more than their low-WMC peers in rule learning, which Goo attributed to the vulnerability of high-WMC learners’ fine-grained cognitive processes to verbalization in dealing with a secondary task. Barkaoui’s (2011) study suggested that the effects of TA on rating processes, rating criteria and/or scores assigned did vary individually, given raters’ rather diverse accounts. He also concluded that experienced raters were more likely to find TA difficult and/or to report that it affected their rating performance negatively than were the novices, since five out of seven raters who reported that TA lowered their confidence and seven out of nine raters who reported that TA drew and/or increased their attention to their rating processes were experienced raters.

As regards the interaction of task type with reactivity, the SLA studies conducted so far do not provide a consistent answer. The six studies using L2 reading tasks (Bowles & Leow, 2005; Goo, 2010; Leow & Morgan-Short, 2004; Polio & Wang, 2005; Rossomondo, 2007; Yoshida, 2008) reported no reactivity of nonmetacognitive verbal reporting on text comprehension (Bowles & Leow, 2005; Leow & Morgan-Short, 2004;
Rossomondo, 2007), recall of propositions in the reading passage (Yoshida, 2008), written production of old and new exemplars of the targeted form (i.e., Bowles & Leow, 2005), recognition and written production of the target forms (i.e., Leow & Morgan-Short, 2004; Polio & Wang, 2005), or written production indicative of rule learning (i.e., Goo, 2010). They also reported negative reactivity on text comprehension (i.e., Goo, 2010; Polio & Wang, 2005), or positive reactivity on targeted form recognition and production (i.e., Rossomondo, 2007), and increased time on task (i.e., Bowles & Leow, 2005; Yoshida, 2008), with four studies not reporting time on task (i.e., Goo, 2010; Leow & Morgan-Short, 2004; Polio & Wang, 2005; Rossomondo, 2007). Although Bowles and Leow (2005) used a text for reading twice the length of the one used in Leow and Morgan-Short’s (2004) study, both studies reported no reactivity on either text comprehension or accompanying learning. Neither did Yoshida’s (2008) study, which employed three types of reading tasks (i.e., read-only, read-and-answer-questions, and read-and-write-outlines), report any interaction effects between task type and reactivity. Together these seem to suggest, based on the variety of reading tasks that have been utilized, that the L2 reading task, as a type of L2 task, might not have statistically significant interactions with reactivity in general. However, it should be noted that Goo (2010), when explaining his different finding of negative reactivity, assumed that one possible reason could be the different target features employed, a finding which definitely needs confirmation.

2.3 Veridicality Studies

There are not many empirical studies on veridicality. Studies outside L2 research usually investigated the issue by comparing the contents of CVRs with a variety of benchmarks that they thought were reliable. Brinkman (1993) compared verbal report data with performance data. The former included strategies coded from CVRs and RVRs
during and after solution of 12 problems in a computer-based fault diagnosis task (Rouse, 1978). The performance data were strategy encodings derived from computer calculation, based on an algorithm. Two layers of comparisons were conducted. The first used kappa to gauge the agreement achieved between verbal report coding and the algorithm. A moderate degree of agreement was found, ranging from 0.53 to 0.61, with that for CVRs yielding a higher rate than that for RVRs. Another layer of comparison was conducted by examining whether verbal report data and performance data pointed to the same patterns of effects. Specifically, two similar sets of multivariate analysis of variance (MANOVA) were run, one on strategy-related protocol measures and the other on corresponding performance measures, taking into consideration verbalization type and task difficulty as two within-groups factors. The results showed inconsistency. Main effects of verbalization type were found on the protocol data but not on the performance data. The inconsistency might be attributed to the lack of veridicality of RVRs. This was because task effects on performance were found in the performance data for both verbalization conditions, but when verbal reports were examined, the task effects were found present with the CVR data, but absent with the RVR data, indicating the lack of consistency of the RVR data. Brinkman concluded that CVRs could be more valid than RVRs. He explained this difference in validity by turning to Ericsson and Simon’s information-processing model and attributed it to the long duration of the task execution, which resulted in the possibility that not all information that passed STM left its trace on LTM to be retrievable.

Williams and Davids (1997) compared the relative veridicality of concurrent verbalization and an eye-movement tracking system in measuring selective attention. Their study consisted of two experiments. In Experiment 1, 10 experienced and 10 less experienced soccer players each watched 26 clips of film that showed 11-versus-11
soccer simulations, under a verbalization and an eye-tracking condition. They were asked to judge, the moment a clip was occluded, the direction the player, who finally controlled the ball, would pass it. In the verbalization condition they needed to report continuously the area of the screen they paid attention to (i.e., the box that highlighted the player who made the final pass, the left of the screen, or the right) as the offensive play went on. In the other condition, they watched the videos without continuous verbalization and made their judgments but each wore a helmet with an eye-movement tracking system mounted to assess their attention. Williams and Davids reported no sizeable differences in the verbal report data and the eye-movement data, since both types of data pointed to similar findings. For example, both sources of data revealed that the more experienced players spent more time attending to areas of display other than the box, such as the positions of other players. One difference, however, was that less frequent instances of alteration of attention were reported in the verbalization condition than in the eye-movement condition. The researchers interpreted the difference as the inability of verbalization to catch attentional shift sufficiently, particularly considering the dynamic situations in the Experiment.

In Experiment 2, another group of 24 experienced and less experienced soccer players watched 3-versus-3 soccer simulations under the two conditions. In the verbalization condition, they were required to report the areas of the screen from which they extracted information, rather than the areas they focused on. According to verbal report data, experienced soccer players attended to areas other than the box area for a greater time period than was revealed by the eye-movement method. Given that experienced soccer players used more peripheral vision, this difference was interpreted as the greater capacity of verbal reports to help identify patterns of attention allocation compared to the eye-movement records because the latter did not track use of peripheral
vision. They concluded that both methods are valid in gauging selective attention in tasks that involved use of central vision, but verbal reports are more veridical in getting information as regards use of peripheral vision. It should be noted that the CVRs utilized in their study may well be an introspective kind, or a directed kind of verbal reports, because they required the participants to verbalize their use of strategies rather than their intact thoughts.

Robinson’s (2001) study also investigated the veridicality of verbal reports by examining if findings derived from the frequency, accuracy, and latency data obtained from verbal reports conformed to usual expectations, within and across maths problem types. Three categories of strategies were examined: counting, retrieval, and special tricks. As expected, children reported more use of the counting strategy for solving harder problems than for easier ones. Also conforming to expectations, Grade 1 children, in the retrospective and concurrent reporting conditions, reported the most counting, in contrast to Grade 3 and Grade 5 children, confirming the tendency for higher grade children to use less counting. The expected pattern for latencies was also met within grade and in both report conditions, with more difficult problems causing longer latencies. The children’s accuracy figures in relation to problem types were also consistent with the expectation that the problems which entailed less counting should be associated with more accuracy, given that counting was error prone.

Relevant expectations were also met as far as the strategy of retrieval was concerned. When dealing with easier subtraction problems, where retrieval was supposed to occur more frequently than when dealing with harder problems, retrieval did occur more frequently. The use of the retrieval strategy also reasonably increased with age, with Grade 5 children exceeding Grade 3 children, who exceeded Grade 1 children in this use. Also as expected, not much difference was found in median latencies for this
strategy, given that it was usually transient and automatic. In terms of accuracy, since retrieval tended to be accurate generally, no big differences across problem type should be expected. This was true with all Grades and across all problem types except that Grade 1 demonstrated more differences in accuracy across problem types, which were in fact much narrower than those found for counting.

Special tricks, as a kind of strategy, surpassed counting in sophistication but were less efficient than retrieval, and “were expected to be reported more frequently on problems with ‘in-between’ difficulty” (p. 219). More use of special tricks was indeed reported on problems with medium difficulty, which also consumed medium solution times, as was expected. Robinson (2001) also pointed out that the fact that children of all Grades reported a consistent pattern of use of special tricks meant that they reported their use of special tricks in a veridical manner. Because special tricks was a strategy supposed to be acquired approximately at the middle stage of children’s development of subtraction skills (Geary, 1994), Grade 3 children should use special tricks the most, followed by Grade 5 children, and then by Grade 1 children. It was found that this expectation was also satisfied when the reported instances of use of the special tricks strategy was checked across grades. These findings led Robinson (2001) to conclude that both retrospective and CVRs provided veridical data.

L2 researchers tended to relate the issue of veridicality to retrospective reports only (e.g., Leow & Morgan-Short, 2004), concerning such points of doubt as memory decay or double-input exposure in the case of stimulated recall (Leow, 2002). The only study on the veridicality of TA in SLA research was conducted by Barkaoui (2011), who examined the transcriptions of rater interviews and their CVRs to identify comments as regards the validity of think-alouds. In his study, veridicality-related comments were identified which suggested that the participants did not report all the thoughts, actions,
reactions, or feelings they experienced, or that they reported thoughts, actions, reactions, or feelings that they did not actually experience. Barkaoui found that more than a third of the participants, mainly experienced raters, admitted that their CVRs were incomplete due to difficulty in verbalizing automatic, intuitive, and covert processes and reactions and due to the complications involved in a rating task that made their reports selective. The CVRs may not be accurate reflections due to concern of projecting a positive self-image and awareness of an audience for the protocols. Barkaoui also argued that experienced raters were more likely to find their reports incomplete perhaps because they possessed more automaticity in rating and they had too much to say.

In summary, the few veridicality studies that have been conducted outside SLA and used non-verbal tasks show positive results suggesting the veridicality of CVRs by comparing the contents of CVRs with other means of reference. In contrast, the only SLA study, Barkaoui (2011), reported nonveridicality of CVRs, including their lack of both completeness and of accuracy, in essay rating, as perceived by some of the experienced and inexperienced raters interviewed. It should be noted that Ericsson and Simon (1993) themselves did not deny the failure of CVRs to record what Barkaoui regarded as evidence for incomplete verbal reports, namely, automatic, intuitive, and nonverbalizable information. Because verbalizers’ comments could be subjectively and impressionistically negative, especially after they engaged in the ordeal of thinking aloud while scoring, Barkaoui would have needed to manage to compare the raters’ actual thoughts during scoring with some reliable references to tell if the reported thoughts were indeed nonveridical.

2.4 Motivation for the Current Study

Empirical quests for answers to the validity of CVRs in writing research are significant because if thinking aloud changes a writer’s cognitive processes seriously
enough and if think-alouds fail to provide veridical information so that verbal protocols provide only a distorted, incomplete, or inaccurate representation of underlying cognitive activities, previous findings about writing models, strategies, L1 use in L2 writing, awareness patterns, and interaction between cognitive processes and text quality, which have been obtained with data elicited through this tool, would have to be re-evaluated.

Given the paramount role of CVRs as a source of data on thinking, and in light of the limited number of empirical studies on their validity in SLA research and particularly on L2 writing, a need is evident for more investigations to address the validity issue and identify the factors of concern in their use. Until there is sufficient and clear evidence concerning their nonreactivity and veridicality, we cannot fully understand their worth and constraints or put them to prudent use.

Specifically, with regard to L2 writing research, studies are evidently needed that explore the effects of CVR on multiple dimensions of writing performance in interaction with L2 writing proficiency, WMC, and task type, thereby inferring the reactivity of CVR on L2 writing processes. Given that Yang et al.’s (2014) study used untreated L2 learners who were recruited from around 20 classes without teaching and teacher biases, a question remains as to whether L2 writers who are active in some way or pay particular attention to a certain aspect of writing performance (e.g., diction) experience the same extent and aspects of reactivity. Answers to this question correspond to an empirically unverified assumption that CVR impairs creativity in writing (e.g., Brinkman, 1993). Studies are also needed that provide empirical evidence for the veridicality of CVRs, namely, the extent to which they are complete and accurate representations of thoughts and thinking processes. In particular, there has been no study that explores the completeness of CVRs by comparing their contents with a benchmark, for example,
RVRs, on a specific aspect of thoughts or thinking processes in writing, for example, those underlying revisions. This study is therefore designed to bridge the research gaps identified above.
CHAPTER 3 METHODOLOGY

Chapter 3 starts with a presentation of the research questions this study aims to address and provides an overview of its design. Then, it introduces the study itself, including the participants recruited, the instruments used, and the procedure followed for data collection. The rest and bulk of this chapter report how data analyses were conducted.

3.1 Research Questions

To address the research gaps identified at the end of Chapter 2, the following research questions are formulated:

1) Does TA affect L2 writing fluency, complexity, accuracy, organization, and content? Especially, does it affect the diversity of words chosen by L2 learner writers who have recently received instruction on using more complex words and therefore have had a boosted level of lexical awareness?

2) How do the effects of TA, if there are any, interact with task type, as compared with the results of previous research, to be specific, the study conducted by Yang et al. (2014)?

3) How do the effects, if there are any, interact with L2 writing proficiency?

4) How do the effects, if there are any, interact with WMC (WMC One, i.e., the operation span, WMC Two, i.e., the reading span, or WMC Average, i.e., the average of the previous two)?

5) How do participants who think aloud perceive the reactivity of TA?

6) Is the information on revisions elicited by TA incomplete contrasted with that elicited by RVRs administered immediately afterwards?

7) How do participants who think aloud perceive the completeness of their CVRs?

8) How do participants who think aloud perceive the accuracy of their CVRs?
3.2 Overall Design

This study adopted a single factor, between-groups design in which the L2 writing performance of a TA group and that of a NTA group were compared and the interaction of proficiency and that of WM with the effects of TA were examined. All participants first received some lexical training to become lexically active, then took two WM tests, then completed a baseline writing task, in which each of them wrote a narrative within an hour silently, and finally completed a main writing task, in which they wrote a narrative similar to the previous one, within the same time constraint, either silently or while thinking aloud. The baseline task was designed to collect benchmark data which would be used to gauge L2 writers’ writing proficiency, to assess between-group homogeneity, and to statistically control prior between-group differences. Immediately after writing the main task narrative, those in the TA group were required to report retrospectively what they thought when they had made revisions. They were then asked to complete a questionnaire which included questions for reflection on the reactivity and veridicality of CVRs. All participants’ writings were analysed in terms of fluency, complexity, accuracy, organization, and content. Their writing scores were then compared, in relation to their writing proficiency, as well as their WMC. The CVRs and RVRs of the think-aloud group were coded and compared to examine the completeness of CVRs. The participants’ answers to the questionnaire were also analysed to examine their perceived reactivity and veridicality of TA. The data of this study and Yang et al.’s (2014) study were analysed together to shed some light on task effects.

3.3 The Study

3.3.1 Participants

Participants were 85 non-English-major young adult university sophomores at one key university in West China. All of them, out of their own accord, completed all the
required WM tests and writing tasks anonymously, and together with those 4 participants who either withdrew or were excluded from final analyses, were remunerated on a task basis (see Appendix 1 for the information sheet and consent form for participants).

They were volunteers recruited from three sophomore classes of English writing taught by the same teacher (Class 217, N = 25, Class 518, N = 29, and Class 1019, N = 31, out of a total number of 95 students for these classes).

They were from eight different schools or institutes and majored in electrics, telecommunications, industrial and commercial management, finance, clinical medicine, Chinese, sociology, philosophy, artistic design, chemistry, mechanical engineering and automation, measurement and control technology and instrumentation, vehicles, energy and power engineering, or economics.

There were 56 males and 29 females. Except for a female student that did not report her age and English learning experience, the participants were 19.49 years of age on average, ranging from a minimum of 17.58 years old to a maximum of 23.83 years old (SD = 1.041), and they had studied English for an average of 10.08 years, the maximum number of years being 16 and the minimum being 6 (SD = 2.484).

All of the participants had taken a uniform English foundation course during the first university year. The English writing course was one of the English course options they had in the first semester of the second year. For this semester, it contained a preparatory session that started from the second half of the week when the university started and consisted of 32 formal sessions which lasted 16 weeks from Week 1 to Week 16 from the second half of September to January, with each week having two sessions. While teacher autonomy was allowed, half of the total 32 sessions would consist of a teacher lecturing based on a textbook, which contained sample essays different in genre,
followed by editors’ comments on the genre and certain linguistic writing tips, and the other half were practice-oriented, during which students would usually write on an assigned topic of the genre that had recently been covered. Some of the compositions they wrote during the practice sessions would be marked and the scores would be recorded as part of their formative assessment results, but other compositions would not be marked and would be included as part of their formative assessment results as long as they completed those compositions.

The study had been approved by The University of Auckland Human Participants Ethics Committee before the recruitment started (See Appendix 2 for the approval letter). At the research site, the dean’s and the teachers’ informed consent was sought and they offered their full support (see Appendix 3 for the information sheet and consent form for the dean and for teachers). The possibility of voluntary participation in the study was introduced in the preparatory session, which was an introductory session when the syllabus, the textbook, the exercise book, and the assessment scheme were introduced, and questions about the course were answered. After this preparatory session, students were still allowed to re-register for another elective or for another class until the second session of Week 1 started. The first session of Week 1 was a lecturing session on narration. During the session break, with the teacher’s permission, a student, who had helped me with advertising and with whom I had communicated sufficiently and appropriately about the study, went into the classroom with handouts and briefed each class on the research project. In the briefing, the class were told that what they needed to do were to take two WM tests during one of three extracurricular time slots convenient to them, to agree that the narratives they would need to complete in the coming practice session could be used for the purpose of this project, and then to write another similar narrative under a silent or a TA condition, extracurricularly, at a time of
their convenience during the next weekend. They were also told that the assignment of
the writing condition for the second writing task would be random, but all of them would
receive a TA training session first. They were then introduced to the payment schemes,
which covered the narratives they were to write in the practice session, as well as the
benefit of getting detailed feedback and their recordings upon request when these were
available. The principle of freedom of participation, non-participation, and withdrawal
without harm, the methods for anonymity, and the exclusive experimental use and safe
storage of data were emphasized. In particular, it was stressed that each student would
be given a digit code and analyses of the photocopies of all their compositions would
only start after the semester ended and their scores for the course were given. They were
told that their unwillingness to participate was respected and could be indicated by not
being present on the day when the WM tests were held, or by indicating in Chinese their
non-participation or withdrawal on their compositions. Most students were willing to
sign up for the experiment, and some asked if their friends could be taken in. A student
who had some hearing and articulation problems also registered her name. She was later
retained in the silent group. I then did not recruit from other teachers’ classes.

On the day when the WM tests were held, a total of 89 students turned up, in three
batches, for the tests. One student took part in the WM tests, was assigned to a group
(WMC Average high, NTA), and completed the first writing task, but was absent from
the second writing task. One student, though completing the WM tests and the writing
tasks apparently, did not properly complete the WM tests, judging from his/her wide
area of blanks on the answer sheets, and the sparse answers. Another student failed to
get 85% of the math calculations right in the operation span task. Still another student,
when taking the reading span test, did not write down the Chinese characters that he/she
was supposed to recall on the proper page. All these four participants’ data were removed from final analyses.

3.3.2 Materials

3.3.2.1 Working memory measures

3.3.2.1.1 Operation span task

An OSPAN task requires a testee to finish a series of mathematical operations while holding in his/her memory the same number of words or letters, each of which follows one of the operations. The OSPAN task utilized in my project was an automated version of Dr. Randall Engle and his colleagues’ work (Unsworth, Heitz, Schrock & Engle, 2005; see also Turner & Engle, 1989; Kane, Bleckley, Conway & Engle, 2001; Kane & Engle, 2003), which took the form of a script that could be operated in a software tool, Inquisit Lab (developed by Millisecond Software, LLC), just by clicks of the mouse. A trial version of the software for 30 days and the script for the OSPAN task, which was programmed by David Nitz and Jerry Grenard, each for a different version, could be downloaded from the company’s website (http://www.millisecond.com).

At the beginning of the OSPAN test, a testee is given the opportunity to familiarize himself/herself with the test, by practicing on the letter memory part and the operation part separately and together. In the test, a testee is required cyclically to compute a math problem, make a decision on the trueness of a following answer, and to keep in mind a letter that subsequently appears on the screen, until a screen with 12 possible letters appears, which marks the end of a set of tests and where the testee needs to click on the letters he/she was supposed to have memorized, in the right order. He/She needs to complete each math operation within an average duration that is calculated from his/her separate math computation practice trials (plus 2.5 SDs); otherwise, the computer screen will jump to the letter part automatically, counting his/her delay in response as a math
error. There are altogether 75 math problems followed by 75 letters, comprising 15 sets, with 3 sets for each set size ranging from 3 to 7 computation-memorization combinations. A testee has to make sure that 85% of his/her answers to the math problems are right to make his/her recall scores valid, with math errors being either speed errors or wrong calculations. At the end of the test, an OSPAN absolute score and an OSPAN total correct score, together with the total number of math errors, are reported on the screen. The OSPAN absolute score indicates the sum of letters in all perfectly recalled sets, which were used in this study.

3.3.2.1.2 Reading span task

The reading span task tests the ability to hold in mind a sequence of words while completing a series of reading activities. The task I used was a Powerpoint version adapted from the series of such tests dating back to Daneman and Carpenter’s (1980) reading-aloud and sentence-final word recall task, and was made to operate automatically on a computer. To complete this task, one reads a four-character mini-sentence in Chinese in a set silently and judges its semantic appropriateness, then keeps in memory a Chinese character which follows, and then advances to another reading, judgment, and memorization sub-task, until the set finishes, when he/she is required to write down all the Chinese characters in the set in the same order as they have been presented. There are altogether 108 sentences to judge and the same number of Chinese characters to remember, comprising 18 sets varying from two to ten sentence-character combinations in length, with two sets at each set size level. Sets of different size are presented randomly. The reading span score is the number of correctly recalled Chinese characters, regardless of their order.

The four-character Chinese sentences were designed to reduce the possible effects of uneven sentence lengths on processing loads in reading. The majority of them concerned
animals and the rest concerned insects, relating to their physical features (e.g., 虾头长角, Prawns have horns; 驴脸很短, A donkey’s face is short; 猫咩咩叫, Cats purr), abilities (e.g., 青蛙善跳, Frogs jump well; 鸡飞千里, Hens fly one thousand miles), living habits (e.g., 飘虫化蝶, Ladybugs transform to butterflies; 猪啃白菜, Pigs eat cabbages; 螃蟹竖行, Crabs walk forward), relations to human beings (e.g., 苍蝇传病, Flies spread diseases), etc. The animals covered were selected after a search using Baidu, a Chinese language-search engine, by entering the key words “常见动物” (commonly seen animals) and then “常见昆虫” (commonly seen insects). Very special animals and those that existed only in mythologies or fairy tales were excluded. The list finally included mosquitoes, flies, horses, fish, crabs, prawns, sheep, snakes, dogs, mice, cats, turtles, butterflies, sparrows, cows, hens, eagles, tigers, lions, frogs, donkeys, pigs, ducks, ladybugs, bees, rabbits, geese, wolves, dinosaurs, deer, earthworms, snails, ants, and swallows. I then searched in Baidu Baike (http://baike.baidu.com), an online collaboratively built encyclopaedia, for each of these animals or insects to source related information. Four-character sentences were then adapted or created based on the encyclopaedia lines. I avoided difficult knowledge and made sure that no judgment of sentence appropriateness would require specialized or expert knowledge. After quite a few rounds of revision of wording and sentence items, a list of 108 sentences was designed. Half of them were made to be inappropriate semantically and the other half were appropriate. All sentences were statements and all of them were affirmative and none contained negation, except just one (鱼不离土, Fish cannot leave soil). A list of these Chinese sentences in relation to their order of appearance and their positions in the 18 sets is given in Appendix 4.
In a small-scale piloting, two female PhD students in education and a female PhD student’s husband who obtained a bachelor’s degree in computer engineering told me after their trial testing that they had no difficulty in judging the semantic appropriateness of the sentences based on what they knew about animals or insects. They did not raise any concern regarding the obscurity of the wording in or the meaning conveyed by the sentence items, nor expressed their unfamiliarity with the topics covered.

The 108 Chinese characters are all most commonly used Chinese characters belonging to the first level of 3,500 Chinese characters in 通用规范汉字表 (General Standardized Chinese Characters List) developed by the Ministry of Education of the People’s Republic of China and China’s National Language And Character Working Committee. The General Standardized Chinese Characters List has three levels of Chinese characters. The first level represents the most commonly used Chinese characters, the second level follows the first in terms of frequency of use, and the third collects characters from specialized fields. According to China’s Basic Education New Chinese Course Standards, the first level of the list corresponds to the primary school literacy range. Also, all of the characters I selected belonged to the most frequently used 380 characters (e.g., Beijing Language Institute Language Teaching Division, 1986). Among them, only those Chinese characters whose numbers of strokes ranged from six to eight were selected, with the exception of one character 相, which has nine strokes, to equalize effort for handwriting. In ordering these Chinese characters, great attention was paid to avoid association in meaning or in phonology between two adjacent characters and among a greater succession of characters in order not to unduly facilitate memorization. For example, in a two-word set, I saw to it that 这, usually a determiner meaning this, was followed by 明, usually an adjective meaning bright or a noun meaning tomorrow, rather than by 些, which is a plural form, because the former
combination made little sense while the latter combination would make up a Chinese word meaning *these*. For the same reason, I did not put together 更, an adverb meaning *comparatively or more*, and 好, an adjective meaning *good*, because the combination 更好 would mean *better*. A list of these Chinese characters in relation to their order of appearance and their positions in the 18 sets is given in Appendix 5.

The Powerpoint version of the reading span task has three parts, namely, the instructions, the practice section, and the bulk of the reading span task. The instructions tell a testee what the task is about, and how he/she should write down his/her answers on an answer sheet (see Appendix 6 for the instructions). The front side of the answer sheet is for the semantic judgment part of the task, where a testee puts either a tick (✔) in a corresponding box if he/she thinks a sentence that has just appeared is a true statement or a cross (✖) if he/she thinks it does not make sense. The back side is for the recall part, where a testee writes down all the Chinese characters that have appeared in a set. Sufficient space is left on the answer sheet for putting the answers down comfortably. The instructions also tell a testee to judge the semantic appropriateness of a sentence according to his/her common sense or general knowledge rather than based on exceptional and imaginative cases. The practice section has three sets, which contain two, four, and six four-character sentences and characters, respectively.

When the practice section ends, a slide that lasts five seconds notifies the beginning of the actual test. A set of the test begins with a slide showing the code of the set (e.g., Set 1, Set 2, etc.), which lasts one second. In the test, all sentence slides last 3 seconds while all the Chinese characters stay on the screen for one second. When a set ends, a slide indicating “请回忆” (*Please recall*) appears, and a testee is required to turn over to the back side of the answer sheet and write down all the characters that have just been presented. A testee has four seconds for noting each character down, so that the recall
reminder slide lasts as long as four seconds multiplied by the number of characters in the set. These time settings were fixed after my trials and based on feedback in piloting. For example, it was found that when the four-character sentences were presented for four seconds, judging their semantic appropriateness and putting down either a tick or a cross became a rather relaxed task and there was still time over, but when the time slot was set to two seconds, completing the series of actions involved in the judgment section incurred much haste. The length of time set for the recall part was also decided taking into consideration normal speed of writing from my own trials and considering the results of the initial piloting effort. When the whole test finishes, a slide appears indicating the end of the test together with a thank-you note.

For further piloting purposes, the reading span test was implemented in an English class of non-English-major sophomores twice, at the same time of day, during a session break on two Monday mornings. These students were at the School of Medicine of the same university to which the participants for the study were affiliated. None of them participated in the TA study. Each of them was given a ballpoint pen as a thank-you gift.

In this piloting, twenty-four students took the test twice. The test-retest reliability for the memorization part was .723 (the first test yielded a means of 86.542, with an SD of 11.206, and the scores ranged from 66 to 105; the second test yielded a means of 95.542, with an SD of 6.192, and the scores ranged from 85 to 104). The internal reliability for the first test was Cronbach’s Alpha = .789; that for the second was .594. When I examined the memorization scores of the 24 participants in the second test, I found three obviously abnormal cases, where three participants wrote down only one character for an eight-character sequence, only three characters for a nine-character sequence, and only two characters for a ten-character line, respectively. These abnormalities could partly explain the low internal reliability in the second test. Another reason might be
their being less serious about the test, given the repetition. A decision was then made that when giving participants the instructions to do the test, I needed to tell them to try to write as many words as they could recall and never give up easily, especially when they met long sequences. The test-retest reliability for the judgment part was .775 (the first test yielded a means of 100.167, with an $SD$ of 3.908, and the scores ranged from 92 to 108; the second test yielded a means of 100.292, with an $SD$ of 4.676, and the scores ranged from 90 to 107). The internal reliability for the first test was Cronbach’s Alpha = .583; the internal reliability for the second test was Cronbach’s Alpha = .697. I examined the four-character sentences and made sure that no sentences were particularly difficult to judge. Given that the judgment part was not included in the calculation of WM scores, it was then decided that when the test was administered later, participants should be instructed to pay equal attention to the judgment part and if any particular case was found, that is, any participant was found to concentrate only on the memorization part, he/she should be excluded from the analysis.

### 3.3.2.2 Writing tasks

The following writing task was used as the baseline writing task:

**Writing Task One**

**Title:** My past, present, and future (300 words or above)

**Instructions:**

1. Describe your life in the past, for example, your junior and senior middle school classrooms, extracurricular activities, holidays, unexpected events, etc. (you may select only some of these examples) (approximately 80 words), and then make a summary of your past, or develop your understanding of it, or reflect on it (approximately 20 words);
2. Describe your life at present, for example, your university classrooms, extracurricular activities, holidays, unexpected events, etc. (you may select only some of these examples) (approximately 80 words), and then make a summary of your present, or develop your understanding of it, or reflect on it (approximately 20 words);

3. Describe what changes you have gone through. For example, you may write about your appearance, state of mind, thoughts, perceptions, pursuits, etc. (you may choose part of these) (approximately 50 words);

4. Describe (Imagine) your life in the future, for example, your work, marriage, family, growing old, etc. (you may choose part of these) (approximately 50 words).

The following writing task was used as the main writing task in the treatment:

Writing Task Two

Title: My father’s/mother’s/uncle’s/aunt’s (choose one) past, present and future (300 words or above)

1. Describe his/her life in the past, for example, his/her birth, growth, schooling, work, education of you, love for you, attitudes towards you, etc. (you may choose part of these) (approximately 80 words), and then make a summary of his/her past, or develop your understanding of it, or show you gratitude for his/her education of you, love for you, and attitudes towards you, or reflect on them (approximately 20 words);

2. Describe his/her life at present, for example, his/her life, work, spare time, state of mind, communication with you, education of you, love for you, attitudes towards you, etc. (you may choose part of these) (approximately 80 words), and then make a summary of his/her present, or develop your understanding of it, or show you gratitude for his/her education of you, love
for you, and attitudes towards you, or reflect on them (approximately 20 words);

3. Describe what changes he/she has gone through. For example, you may write about his/her appearance, state of mind, thoughts, perceptions, hope, requirements of you, attitudes, etc. (you may choose part of these) (approximately 50 words);

4. Describe (Imagine) his/her life in the future, for example, his/her work, part time, interests, attitudes towards life and you, requirements of you, relationship with you, health, growing old, etc. (you may choose part of these) (approximately 50 words).

The two tasks had been piloted in a case study involving Chinese undergraduate and graduate students of different proficiency. Both tasks were given in Chinese, except for the title. Participants were required to write at least 300 words within 60 minutes and to submit their writing as soon as they finished. The time frame was considered quite sufficient because in the national College English Test (Band 4) which all of them had taken at the end of the first year of college and most of them had passed, they were required to write 120 to 180 words within 30 minutes for an argumentative (or, quite rarely, a narrative) piece with only a very sketchy outline or no outline given at all. The detailed outlines for the above two tasks were supposed to ease L2 writers’ burden of planning and therefore facilitate their writing.

As can be seen, both writing tasks were meant mainly to be narrative. Both entailed the use of different verb forms concerning the past, present, present perfect, and future. Both were also similar in terms of contents and macro-organization of composition, given the almost identical layouts of the outlines—except for a change of focus in the subjects to be described. The two writing tasks were similarly designed so as to ensure
that any homogeneity or heterogeneity in scores of performance on Writing Task One between the TA and NTA groups could maximally represent the two groups’ relative abilities to write on Writing Task Two presumably still under silent conditions.

3.3.2.3 Immediate retrospective verbal reports

The retrospection reports for the TA group to complete immediately after the main writing task asked them to look back at the revisions they had just made on-line and report what they thought when they made them. The use of this type of immediately conducted RVRs to acquire useful information, especially when stimuli are provided, has been supported cogently (e.g., Ericsson & Simon, 1993; Gass & Mackey, 2000). The following are the English version of the instructions for the RVRs, which were given in Chinese at the time of experimentation:

**Experiment Requirement Two:** Now please look back at the revisions/alterations (e.g., deletions, additions, replacements, movements) you made on-line. Circle all of them in red with the red pen provided. Then, one by one, in the order of their occurrence, report what you thought at the moment you made them. Please record your reports. Do not revise again!

The reason for asking the participants to report retrospectively on their revisions was that human revision is understood to be a controlled, non-automatized, or conscious behaviour that involved heeded information. Such information should be reportable and should be reported in CVRs, and anything missing from it can then be interpreted as the CVRs being incomplete. It was noted that such information could be metacognitive, which involved justifications. However, the difference between nonmetacognitive and metacognitive verbal reports is not that the former does not contain any metacognitive elements, but lies in the proportion/percentage of these elements (Bowles, 2008), and whether (the retrieval of) metacognitive elements constitute(s) an indispensable part of
task requirement. It was also noted that revisions or alterations in writing could be made rather swiftly, but I inferred that there must be underlying thoughts, which, at the time of revision, occurred to participants and became what was reportable and should be reported.

3.3.2.4 Reflection questions

The reflection questionnaire asked participants in the think-aloud group to answer in written forms the following questions once they finished reporting retrospectively:

1. Do you think what you reported represents what you thought completely? That is, are there any thoughts that you did not report or could not report? Please elaborate.

2. Do you think what you reported represents what you thought accurately? That is, are there any inconsistencies between what you reported and what you thought? For example, are there any cases when, for some reason, you were not willing to report your real thoughts but reported something else instead? Please elaborate.

3. Do you think thinking aloud affected the way you wrote the essays? Please elaborate.

The questions were adapted from Barkaoui’s (2010) interview questions. They were given in Chinese at the time of experimentation and the participants were allowed to answer them in Chinese. A minimum of 50 words were required for an answer to each of the above three questions.

These questions aimed to elicit thinking aloud participants’ self-perceptions as to the completeness, accuracy, and reactivity of thinking aloud. Data of this type complemented or triangulated quantitative experimental data in contributing to an understanding of the veridicality and reactivity of TA in general and for individuals.
3.3.3 Procedure for Data Collection

Table 3.1 presents the time frame for the experiment.

Table 3.1

*Time Frame and Arrangement of the Project*

<table>
<thead>
<tr>
<th>Time slot</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
<th>Sunday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session 1</td>
<td></td>
<td>Baseline writing</td>
<td>TA training &amp;</td>
<td>TA training &amp;</td>
<td></td>
</tr>
<tr>
<td>Morning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session 2</td>
<td></td>
<td></td>
<td>TA training &amp;</td>
<td></td>
<td>Main task 3</td>
</tr>
<tr>
<td>Afternoon</td>
<td></td>
<td>WM tests 1</td>
<td>TA training &amp;</td>
<td>Main task 2</td>
<td></td>
</tr>
<tr>
<td>session 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Afternoon</td>
<td></td>
<td>WM tests 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evening</td>
<td></td>
<td>WM tests 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The experiment started from the second half of Week One and lasted five days. Participants first completed the two WM tests in a time frame that was convenient to them on Wednesday, after they finished their first lecturing session in this week. They then attended their practice session and completed the first narrative writing task on either Tuesday or Friday. They were then randomly assigned to either a TA or a NTA group and wrote the second narrative writing task, in separate classrooms simultaneously, either on Saturday or on Sunday. These arrangements were subject to the availability of time and classrooms.

3.3.3.1 Prior instruction

Before the participants wrote the baseline narrative, they had attended the first writing session, which was a lecture on Unit 1 of the textbook. This unit was one of the two units on narrative writing. It presented a narrative model text entitled “Unexpected Money”, with a focus of instruction on how to develop a narrative. The students were asked to read the model text themselves first, and were then asked questions about its structure, before the teacher presented the genre knowledge. The writing tips section featured choice of “wonderful words”. It contained around 10 English sentences, with
their Chinese translations given, and with the “wonderful words” printed in bold type. In about 10 minutes’ time, the teacher and the students went through the sentences and appreciated the use of the words. In another five minutes, the students were also introduced how to use several online dictionaries, Google, and the synonym function embedded in Microsoft Word to aid their diction. The lexically focused instruction could have raised the participants’ awareness of lexical diversity before they wrote the baseline narrative.

3.3.3.2 Implementation of working memory tests

All the participants took the two WM tests in three batches, at a multimedia classroom where there were around 40 student computers, a central screen, a projector, the teacher control, and access to the Internet. Each of them was first randomly assigned a code ranging from 10 to 100, and was told to use this code for their identification.

To start the tests, I gave each of the participants a softcopy and a hardcopy of the answer sheet that contained all instructions and procedures. The operation span task involved six steps: 1) Log on to http://www.millisecond.com/download/library/OSPAN; 2) Download Inquisit 4; 3) Download AutomatedOSPAN; 4) Install Inquisit 4; 5) Click Inquisit 4, and 6) Open AutomatedOSPAN and run it. I demonstrated the steps and made sure that every student finished Step 4, the installation, and then locked their screens. I opened the program and showed them how to complete the operation span test. The software was in English, so I explained to them in Chinese the instructions. I presented to them and completed the practice sections for the participants. I emphasized that they would need to do the math part as accurately as possible and recall as many letters as possible. I also showed them how to click the CLEAR, BLANK, and EXIT buttons (for correction, for skipping over a forgotten letter, and for proceeding to the next sequence of letters, respectively). Then after making sure that there were no questions, I unlocked
the participants’ screens and asked them to go through the whole process and to record their scores on their answer sheets. The whole process took about half an hour.

After making sure that every participant finished the operation span test, I locked their screens and began to demonstrate the reading span test. I first explained the instructions on the student answer sheet; they were in both English and Chinese. All of them understood the instructions well because of their experience with the operation span test. I then administered the demo version of the test I had made, which contained two sequences of judgment and memorization tasks, to demonstrate how to complete the judgment and the memorization tasks. I emphasized to them that every time when a four-character phrase appeared on the screen, they needed to make a judgment by putting either a tick or a cross in the corresponding box on the answer sheet. I also reiterated that only when they completed the judgment task could they turn over to the back side of the answer sheet to note down the sequence of Chinese characters they were required to hold in memory. After answering the students’ questions, I then opened the slides that contained the practice sequences and the test proper and set the slides to the auto-play mode. While the slides were played on the teacher’s terminal, all participants in the classroom could see the slides on their computers, as well as on the central projector screen. The whole process took about 20 minutes. The student research assistant and the technician helped me with the tests.

3.3.3.3 Baseline writing task data collection

All participants completed the first writing task during their usual practice sessions. The writing task was administered to familiarize them with writing and writing a narrative, and to test their use of different tenses involved in narration. They were required to write at least 300 words about the task on an answer sheet within 60 minutes. They were asked to write to their best, to keep eligibility, not to use dictionaries, not to
use correction liquid or erasers, and to submit their compositions immediately after they finished them. They were also asked to indicate their starting time and finishing time on their compositions. In this round of writing, all participants completed the writing task within 60 minutes. The compositions of all students, including those who were willing to participate in the experiment, were collected, photocopied, and kept for later analyses.

3.3.3.4 Think-aloud training

The think-aloud training was given to all participants, regardless of their assigned conditions, prior to the second writing task. It started with a presentation of an adapted version of the think-aloud instructions used in Yang et al.’s (2014) study, which were given in Chinese at the time of presentation.

The following is an English version of the instructions:

Report your thoughts while your write

Report your thoughts while your write, so that you or your teacher may get to know your thoughts in the writing process.

You can begin to tell aloud everything that occurs to you, from the moment when you see the writing topic to the moment when you finish writing. Report whatever you think, including your thinking activities prior to writing and while writing (at this time you need to report your thoughts while you write), which may involve your planning, organization, diction, sentence construction, etc. Just report whatever occurs to you.

Just write as you usually do. There is nothing special except that you need to report.

You are free to choose your reporting language, Chinese, English, or Chinese mixed with English, whatever you feel comfortable with.
Note again that you will need to report your thoughts continuously upon your seeing the topic, and during the whole process of completing your writing, you should not pause for more than nine seconds.

Reminder: if you fail to report continuously, you might be asked to re-do the task.

No one will communicate with you. You may be given a gentle reminder when you forget to speak.

Three audio clips were played that contained TA episodes. The first one was a nonmetacognitive verbal report of a math problem-solving process where a linear system of two equations with two variables \((1 + x = y; 2x – 3y = 9)\) was solved; the second was a 36-second long clip that concerned a pre-writing planning process where the topic (An unforgettable person) gave rise to some thinking on whom to write about and a decision was finally made to write about a teacher; the third clip was 46-second long and presented a piece of while-writing reporting that involved writing a sentence \((I \ can \ never \ forget \ him \ because \ I ... \ he \ helped \ me \ a \ lot)\).

The participants’ questions were then addressed and a check was made to ensure that all of them understood what they were expected to do in providing think-alouds. Then all participants were given 15 minutes to trial-write at least 50 words on a topic “My Favorite Book” while thinking aloud. I walked around the classroom to listen to them and made sure all of them were able to think aloud readily. The training lasted about 30 minutes. Afterwards, I asked them if they still had any further questions and then, with the help of the research assistant, assigned those under the silent condition to a different classroom.
3.3.3.5 Main writing task data collection

Before the TA training session and the main writing task, a research assistant and I managed to score all participants’ WM scores. Based on their performance on the WM span tasks (the averages of their operation span and reading span test scores converted into z-scores), the participants were assigned to three WMC groups (high, mid, and low): high for those whose z-scores were higher than .5, mid for those whose z-scores were between -.5 and .5, and low for those lower than -.5, following Mackey et al.’s (2002) approach. The participants in each of the three WMC groups were then randomly assigned to either a TA group or a NTA group, using an online randomizer (www.randomizer.org).

Table 3.2
Experimental Grouping of Participants

<table>
<thead>
<tr>
<th>Groups</th>
<th>WMC High (N = 28)</th>
<th>WMC Mid (N = 37)</th>
<th>WMC Low (N = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think-aloud (N = 43)</td>
<td>N = 15</td>
<td>N = 18</td>
<td>N = 10</td>
</tr>
<tr>
<td>Non-think-aloud (N = 42)</td>
<td>N = 13</td>
<td>N = 19</td>
<td>N = 10</td>
</tr>
</tbody>
</table>

The assignment of participants in each WMC group in relation to their condition in the second writing task is given in Table 3.2.

Right before each batch of participants started the experiment, I made sure, with the help of the technician, that all computers could work, had proper software to record well, and had workable headsets. I also prepared several notebooks and headsets for backup.

All participants randomly assigned to the TA group completed the second writing task while thinking aloud in a multimedia classroom while those assigned to the NTA group completed the second writing task silently in a spare classroom, led by the research assistant. All participants were required to sit at least two seats apart, or were separated by the aisle and a seat. Those who were assigned to the TA condition were first asked to check the recording device before they were given a package of materials.
The package was designed similarly to the one used for the first writing task, except that it provided the think-aloud instructions, reminders for recording, instructions for immediate RVRs, and a reflection questionnaire. I emphasized that the participants would need to start to tell their thoughts continuously as soon as they wrote down their starting time and saw the writing topic on the next page, and that they were allowed to leave the classroom quietly as soon as they finished all the requirements set in the package, including saving their recordings.

During the first several minutes when the think-aloud experiment started, I walked quietly around the classroom to make sure that they did report continuously in conformity with the instructions. Soon later, I stood at the front of the classroom and watched them write and report, in case any incident should happen. Overall, I did notice several participants report in a low voice, and reminded them to raise their voices. But I found them lower their voices again afterwards. I stood close behind them for a brief moment and confirmed that they did report according to the instructions. At that time I thought audio play software that had voice amplifying functions might later solve any problems. Another incident was that a student found, only when he finished the writing and think-aloud part, that he had failed to operate the recording software properly although I had noticed him reporting competently. He was told to go on and record the latter parts of the experiment properly.

Students who were assigned to the NTA condition completed the second writing task in a spare classroom under close invigilation just as they had done the first one. They received a package of materials that resembled the one they had for the first writing task, except that the second writing topic was given in the package.
3.3.3.6 Immediate retrospective verbal reports

Right after they completed the think-aloud writing task and had noted down their finishing time, participants in the think-aloud condition were instructed not to take off their headsets but to go on immediately to report in retrospect what they had been thinking when they made on-line revisions. They were asked to circle the revisions that they reported on using a red-coloured pen provided. All participants recalled and reported their thoughts on revisions earnestly, according to the instructions, but it was noted that they did not report on all the revisions they had made.

3.3.3.7 Questionnaire implementation

All participants who thought aloud, after completing the retrospection task and saving their audio files, were instructed to answer three questions concerning the completeness, accuracy, and reactivity of the think aloud activity they had just experienced, in no fewer than 50 words for each question. It was later noted that a small number of the participants did not meet the word limit, and several of them gave an off-topic answer to the accuracy question.

When they completed the questionnaire, they were instructed to leave their packages of documents on the desks and were allowed to leave. When all participants finished the experiment, the research assistant and I collected all the writing packages and saved the audio files to two hard disks.

3.4 Data Analysis

3.4.1 Data cleanup and verification of task conditions

As I have explained in the participants section, of the 89 students who came for the WM tests, four participants’ data were excluded from analyses. Prior to further data analyses, I checked all the 85 participants’ compositions, typed them into Microsoft Word, and checked the resultant files against the original compositions very carefully.
During the process, however, I found that one student failed to record his/her finishing time for his/her second essay, and one student still used an eraser in making revisions for his/her first essay. For the former case, the student’s essay was not analysed in terms of time and speed; for the second, the essay was excluded from analysis of dysfluencies.

In terms of the students’ recordings, five were not eligible for analyses. One student failed to run the recording software properly, and four students’ voice levels were too low to be audible, even when the transcriber, who was a postgraduate student in English language and literature, tried to amplify their voices using technical methods. The reasons could be technical, or their mumbling ways of reporting. These students’ WM test scores, their compositions and questionnaires were still included in the analyses. The eligibility of these samples for the reporting condition was judged from the participants’ performance at the time of experimentation, which I watched, from the discernible continuity of their effort to verbalize in some of the recordings, and from their replies to the reflection questions. Therefore, altogether, 38 think-alouds and RVRs were compared and analysed on veridicality.

3.4.2 Independent variables and scoring

The automated operation span task yielded scores automatically as soon as participants completed the task. For the reading span task, the memorization part was scored very carefully, with each correctly recalled Chinese character given one point, regardless of its relative order in a sequence. In fact, the participants mostly followed the original orders of character presentation in their recall and it was rare that points were given to correctly written Chinese characters in places other than their original positions in a sequence. For the memorization part of the reading span task, the internal reliability for all 85 participants’ scores reached .792.
One MA student majoring in accounting helped me to mark the participants’ judgment task performance, that is, to score the ticks and crosses in the participants’ answers. My random inspection of eight participants’ scores he marked showed a 100% rate of agreement. The internal reliability for the judgment part of the test reached .656. When I examined the scores for the judgment part, I found two obvious cases where one participant scored only two points for an eight-character sequence and the other scored only four points for a nine-character sequence. This abnormality may partly explain the lower internal reliability for the judgment part.

3.4.3 Dependent variables and data coding

3.4.3.1 Measures

3.4.3.1.1 Fluency measures

The fluency measures included the following:

1. Time on task (Time). It was calculated based on the timings reported by the participants on their essays, measured in minutes.

2. Speed One, words per minute—the number of words in the final composition divided by the total number of minutes it took a participant to write a composition. I followed Polio’s (1997) instructions on word counting. That is, a contraction was counted as one word (e.g., I’d). Proper nouns, especially those in Chinese pinyinized forms, were also counted. So were numbers. In a number cluster (no blank in between), e.g., 2010.5.7., C106, CET4, or 7:00, one word was counted, but when a number was detached from a word or an analogue of a word (with a blank in between), it was counted separately (e.g., two words were counted for CET 4, or No 4). Hyphenated words were counted separately (e.g., university-entering exam, 2-grade student, or fast-paced society was counted as having three words). Slash-divided words (e.g., his/her) were counted separately.
as well. Titles were not counted. Neither were Chinese characters. Word number counts were not related to spelling correctness. In word counting, I relied on Microsoft Word word counting, either adding or subtracting the particular cases in the participants’ writings which I had highlighted.

3. Speed Two, the total number of words produced (including those deleted or crossed out or smeared) per minute. A letter or other partial forms of a word were still counted as a word.

4. Speed Three, syllables per minute. This is a finer measure of the speed of writing than words per minute (Ellis & Yuan, 2004). A scheme for coding syllables was developed based on Yang et al.’s (2014) scheme (see Appendix 7). The reason for following the British English pronunciations was that I was more familiar with the system, given my early English education.

5. Speed Four, the total number of syllables produced (including the number of syllables in words deleted or crossed out or smeared) per minute. An independent letter/letter cluster was counted as a syllable.

6. The percentage of nondysfluencies (ND), that is, one minus the percentage of dysfluencies. The percentage of dysfluencies is the number of words reformulated (i.e., crossed out or smeared and changed) divided by the number of words in the final form of a composition. A partial reformulation was still counted as a case of dysfluency, even if it concerned only a letter. Words that were deleted and words that were inserted into an original line were counted. When one participant crossed out several words and made reformulations above the deleted, only the words crossed out were counted.

3.4.3.1.2 Complexity measures

The Complexity measures included the following:
1. Length One, the number of words in the final form of a composition; Length Two, the total number of words, including words crossed out and reformulated; Length Three, the number of syllables in the final form of a composition; Length Four, the total number of syllables, including syllables crossed out and reformulated.

2. Syntactic complexity, which was gauged by the number of both finite and nonfinite clauses divided by the number of t-units in a participant’s composition. T-units and finite clauses were coded based on Polio’s (1997) scheme. For example, in counting t-units, a sentence fragment with no verb or copula was counted as a t-unit, but a noun phrase or a subordinate clause standing alone was attached to the neighbouring t-unit if appropriate. Such wrong patterns involving subordination as Although/Even though/Even if ..., but/notwithstanding ... or Now that/Because ..., so ... were coded as having one t-unit, considering that they contained the element of subordination, despite the coordination that followed, taking the influence of the Chinese language into consideration. The same leniency was also given to cases where even was wrongly used instead of even though or even if or even when. Unsuccessful as they were, they were taken as tangible attempts to achieve complexity. Therefore, Even they come back home, they just stay to have a dinner was coded as having just one t-unit. Some complexity in coding coordination was found as well. For example, But as I am stepping into the adult world day by day, I realized there are many things I have to learn, and the truth is usually beneath the surface was coded as having one t-unit, but in When I compare my present with past, I feel that I am richer in knowledge that the past, I have more friends and a more aboard field, two t-units were identified because I was not sure if the last clause deserved to be a run-on version of a continued object clause (perhaps more convincingly so if there were
an and or a that in its front). In coding t-units, reference was also made to Foster, Tonkyn, and Wigglesworth’s (2000) concept of “An independent sub-clausal unit” for including ellipted sentences or similar (successful or unsuccessful) attempts, which consists of “either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation” (e.g., A: | How long you stay here | B: | three months |), “or a minor utterance” (e.g., | Oh poor woman |, | Thank you very much |, | Yes |) (p. 366). A scheme for coding T-units was developed and modified during my coding. A final version is given in Appendix 8.

I also followed Polio’s (1997) scheme for coding finite clauses. Notably, incorrect forms of subordinate clauses that contained no finite verbs were still coded as finite clauses (one erroneous clause and one verb form error were also counted). Therefore, sentences such as We have to think of what the causes of this and what solutions you can suggest and So, my father is a people who strong, stubborn and self-confident were coded as having three and two finite clauses, respectively. Independent constructions without (proper) finite verb forms were coded as finite clauses as well. For example, In a word, my past time fall of unforgettable memories was coded as having one finite clause (fall seemingly as full misused or misspelt). Several cohesive ties having somewhat established status such as what’s more, what’s worse, and what’s important were not counted as finite clauses though each of them contained a finite verb (As far as I’m concern was counted, however). There were also several cases where the subject in a there be construction was wrongly followed by another finite verb. I coded two finite clauses because I thought of no better way to code such a construction and because I suspected that corrections could be made by making the second
finite part a subordinate clause by adding a relative pronoun or by turning the finite verb into a nonfinite one. I thought that in either case, no difference would be made in terms of the measure of syntactic complexity I adopted. Two additional reasons for such a coding method were that I thought it actually already took the form of a finite verb and that a coordination might not be appropriate as well. Therefore, But when the spring festival approach, there are always some of his subordonates come to visit my father was coded as having three finite clauses. Similarly, the following two sentences were coded as having two finite clauses: Today, I have many thoughts want to talk to you, about my past, present, and future, and Going back to his hometown and living with his brothers and sisters, may be the best thing can do for him. Finally, when finite and nonfinite verb forms were used in wrong mixture, they were coded contextually. For example, but thought about it carefully. I believed it worthwhile was coded as having one finite clause and one nonfinite clause, taking thought about it carefully as a peripheral modifier; I having been dreaming of a wonderful life with a harmony family and successful job was coded as having one finite clause, taking the seemingly nonfinite form as an unsuccessful attempt for making the present perfect continuous tense. Independent sub-clausal units (Foster et al., 2000) were also counted as finite clauses. A scheme for coding finite clauses was developed and modified during my coding. A final version was given in Appendix 9.

A non-finite clause consists of a non-finite form (such as an infinitive, participle, or gerund), together with its subject, object, complement, and adverbial, which may be implicit. Foster et al. (2000) pointed out the ambiguity existing between some noun forms and minimal nonfinite clauses (e.g., reading in I like reading, p. 366) and they adopted a conservative method that required
that a non-finite clause comprise one nonfinite verb plus at least one clause element. For example, according to their method, in *they could just make ends meet*, or *I also have all kinds of class to attend*, or *it is time to marry*, or *I don’t like to be beat*, no nonfinite clauses was counted. Their method had been followed by Yang et al. (2014). In this coding, I included all nonfinite forms, even when there were no tangible elements attached to them. I distinguished noun -*ing* forms from participles based on Longman Dictionary of Contemporary English (http://www.ldoceonline.com/), Iciba (http://www.iciba.com), or other resources. For example, independently used *smoking*, *drinking*, *dancing*, *painting*, *cooking*, *teaching*, *reading*, *coming*, *gathering*, *ranking*, *job switching*, *living*, and *speaking*, which were codified in these dictionaries as noun entries, were not coded as nonfinite clauses when used as nouns, but those *-ing* forms that could not be found in the dictionaries as noun entries and/or used as actions would usually be coded as nonfinite clauses, for example, *sleeping*, *imagining*, *eating*, *talking*, *smiling*, *photographing*, *studying*, *struggling*, *sharing*, *calling*, *playing*, etc. Therefore, while in *I love reading*, no nonfinite clause was coded, in *Every common person's life is a story which is worth reading*, a nonfinite clause was coded for *reading*. Also, all *doings in go doing* constructions were coded as contributing to nonfinite clauses because they indicated actions. I did also code nonfinite clauses using Foster et al.’s (2000) method as well. The correlation between the numbers of nonfinite clauses coded following the two methods for all articles reached .966.

Verbless clauses such as *when very young*, *while full of confidence*, *though different from what we get before*, and *busy with countless his work* were counted. A non-finite verb followed by another non-finite verb as its clause element was
counted as having two non-finite clauses (e.g., *There is still a long way to go to solve the problem, going hiking with my friends*). I also coded phrases containing nonfinite verbs such as *to be frank, to be honest, frankly speaking, Needless to say, or like it or not* as nonfinite clauses, thinking that such usages could be encouraged. A nonfinite verb, together with *how* (e.g., *how to love, how to be loved*), a subject (e.g., *For the year to come*), in a verb phrase (e.g., *to bring up, to grow up, to take place, giving up*), or having a clause as its object or adverbial was also counted as a nonfinite clause (e.g., *to work when he was very young, Here, I just want to say, “My mother, I love you!”*). Patterns such as *(be) supposed to, would like to, (be) able to, (be) likely to, (be) expected to, (have) no choice but to, (be) willing to, and (have) no way but to* were counted as contributing to nonfinite clauses, while they were also counted in terms of different verb forms of modality. A scheme for coding nonfinite clauses was developed and modified during my coding.

3. General complexity, namely, the number of words (Length One) per t-unit (Norris & Ortega, 2009). Being a length-based metric, general complexity may not be able to reveal fully the finer formulations of complexification (Wolfe-Quintero, Inagaki, & Kim, 1998). However, it is still “legitimate” as a measure of complexity, which presents “global, unspecified, or general interpretations” and differs from other complexity metrics just by the degree of fineness, and is valuable because “such a broader measure might be able to capture large-scale or long-term variation that would be missed by finer-grained, more specific metrics” (Norris & Ortega, 2009, p. 568).

4. Subclausal complexity, or complexity by subclausal or phrasal elaboration, namely, the number of words per finite clause. Norris and Ortega (2009) argued
that three aspects of complexity should be measured at least, namely, general or overall complexity, complexity by subordination, and complexity via phrasal elaboration, considering that the construct of complexity is multidimensional.

5. Syntactic variety (SV), namely, the total number of different grammatical verb forms in a composition. Grammatical verb forms differ with tense (e.g., simple present, past perfect, or future continuous), modality (e.g., can, may, must, should, have to, etc.), and voice (e.g., passive voice in the past, active voice at present, etc.), which are supposed to indicate low-level complexity (Ellis & Barkhuizen, 2005). Different nonfinite verb forms (e.g., to do, to be done, to be doing, doing, having done) were also coded, so were modality forms that have not acquired official status (e.g., be likely to, be able to, have no choice but to, be willing to). Wrong verb forms that did not match any official tense, modality, or voice forms (e.g., must be do, have be done, and will left) were attributed to categories that bore the closest resemblance to the strange forms (e.g., must be done, have been done, and will leave, correspondingly). A scheme for coding different verb forms was developed and modified during my coding.

6. Lexical diversity (LD) was measured by $D$ (Malvern & Richards, 2002; see also McKee, Malvern & Richards, 2000; Richards & Malvern, 1997), as it was in Yang et al.’s (2014) study, to facilitate comparison. $D$ is calculated via a mathematical equation that relates TTR (Type-Token Ratio) to token size (N):

\[
TTR = \frac{D}{N} \left(1 + \frac{2N}{D} \right) - 1.
\]

Compared with MSTTR, $D$ has the merits of independence of sample size, random sampling from the whole of a transcript to take both long-distance and short-distance repetition into account, and being more informative because “it is representative of the whole of the TTR vs. token curve rather than just a single point on it” (Malvern & Richards, 2002, p. 91). $D$
values were calculated by running the *Vocd* command in the CLAN (Computerized Language Analysis) programs (MacWhinney, 2000a, 2000b), which was written by Gerard Mckee and is freely available to other researchers as part of the CLAN programs.

### 3.4.3.1.3 Accuracy measures

The accuracy measures included the following:

1. Error-free words (EFWs), the number of error-free words per 100 words (i.e., \((1 - \text{the number of errors/Length One}) \times 100\)). All errors were counted, including errors concerning articles, proper nouns, Chinese characters, blanks (left for want of words), punctuations, spellings, and capitalizations, in order not to miss any subtle effects of CVR. Errors were counted separately, even though they concerned a single word. For example, for *literatures*, one spelling error was counted, together with a plural form error. Errors related to collocations were counted once. For example, *lay more attention on* was only counted as having one error considering *pay more attention to* was the most appropriate substitute, despite the fact that two words were involved in the revision. A language point related to one error that was correct in reference to the error was exempted. An example was *He will become more wisdom, more happy, and healthy*.

(W)*Wisdom* was used wrongly as an adjective, but *more* should not be counted as an additional error because *wisdom* has two syllables, indicating a correct use of the form for the comparative degree, even though the correct expression should have been *wiser*. Similarly, in *Another different area between college and high schools is the course you study which is more difficult*, the second *is* should not be counted as an error even if *course* should be in its plural form. Sometimes, the number of errors was decided by comparing the original with a revision that
had the closest possible resemblance to the original but would make the number of errors fewer. Extraneous commas were considered errors. Missing commas were coded case-by-case. The word however used as a conjunctive adverb or the expression For example put at the beginning of a sentence without a comma following it was coded as an error. So was What’s more as a conjunction because of the possible confusion with the following noun as a subject. The missing comma before a coordinating conjunction joining two independent clauses or between a sentence-initial subordinate clause and a main clause was not counted as an error unless confusion was caused. Both commas and missing commas after sentence-initial prepositional phrases, adverbs such as at present, surprisingly, thus, therefore, moreover, so, sometimes, now, then, today, evidently, of course, occasionally, soon, on the other hand, instead, and furthermore, and the conjunction but were accepted, unless confusion was caused. Decisions on verb form use concerning tense, aspect, voice, etc. were made based on the preceding and following discourse. For example, the seemingly inconsistent present tense usages in between or at the end of a series of past events were not coded as errors when they were used to indicate what were felt at present about one’s past. Both American and British English were accepted. Some China English forms were also tolerated if they conveyed meanings understandable to me while not violating grammar rules (e.g., I feel like a fly without a head, meaning I feel aimless). Oral or informal English, including words, abbreviations, and idioms (e.g., ’cause, wanna, or gonna) was also accepted. Any point of doubt that arose was resolved by reference to Google, the online version of Longman Dictionary of Contemporary English (http://www.ldoceonline.com), or/and British National Corpus
A detailed error coding scheme was developed, in reference to Polio’s (1997) and Polio and Shea’s (2014) schemes. It was continuously revised, complemented, and enriched during my coding processes. The final scheme was given in Appendix 10.

2. Error-free clauses (EFCs), the percentage of both finite and nonfinite clauses that did not contain any errors. For run-on sentences, only the last sentence was coded as correct. Attention was paid to attaching an error to the appropriate clause. For example, in *He will be concerned with me as before and give me a hand when I get in trouble with no reward*, the last error, which concerned *with no reward* (a correction would require putting a comma before it, or a position change, or a perhaps more pertinent revision to *without asking for any reward*), should go with the first finite clause so that the subordinate clause *when I get in trouble* is error-free.

3. Correct verb use (CVU), the ratio of correctly used finite verb forms to the total number of finite verb forms. Finite verb forms included those that were not counted as contributing to finite and nonfinite clauses—those in a compound predicate, in tag questions, etc. (e.g., *learned to sing in He went to a music school part time at eight and learned to sing*). Finite verb forms also included null verb forms in constructions that were coded as finite clauses with errors (e.g., *when I old, Maybe it totally due to my study, etc.*). Several ellipted constructions that did not entail finite verb forms were not included in the analysis (e.g., *How about me?*).

Verb form errors included errors in the use of tense and aspect, voice, modality, mood, and form (e.g., *choosed*) and errors due to absence of verb forms (e.g., *As my father older, he didn’t try to do something that was very*
difficult for him). For example, in *Now, she always runs in the morning and do exercises to keep healthy*, one verb form error was counted for *do*. Attention was paid to distinguishing verb form errors from other types of errors such as lexical errors and misspelling. For example, in *I am grateful because these habits are benefit for my whole life*, I did not count any verb form error because I assumed that the student writer may have mistaken *benefit* for its adjective *beneficial*. In *Few days ago, I worked in a stree with tall trees, that our hometown also like this*, I counted *worked* as a lexical error rather than a verb form error, taking it as a wrong execution of *walked*. Similarly, in *It has so much fun and I think I will strive for that life*, I took *has* as a verb use error (for *is*), but did not count any verb form (tense) error. When a verb form error and another type of error concurred, I counted all types of errors cumulatively. For example, I counted three errors for the form *acquainted* in *I have acquainted lots of people, many of whom now have become my life-long friends*, one due to its misspelling, another related to tense, and still another concerning diction or verb use. My judgment on the correctness of verb forms was based on discourse rather than superficial inconsistency. For example, in *They gave me support when I was in trouble, they shared their happiness when I was sad, and I really show great gratitude for that*, I did not count a verb form error for *show*, because I took the use of the simple present as indicating the author’s present reflection. Not using finite verb forms were also coded as verb form errors. For example, in *So I treasure my present life and excited to have more adventures in the future*, one error was coded for not having the link verb *am*. However, in some cases where words of other parts of speech were wrongly taken as finite verbs, no verb form errors (for absence of finite verbs) were counted. For example, in *After my
graduation, I will still my hurried life, no verb form error was coded. Also worth noting was that finite verb forms wrongly used where nonfinite verb forms should have been used were not counted as verb form errors because these finite verb forms were treated as constituting nonfinite clauses.

3.4.3.1.4 Content

The content of the composition was scored based on the relevant scheme developed for assessing content in Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey’s (1981) ESL Composition Profile. The scheme has four levels of content quality: 1) EXCELLENT TO VERY GOOD, scored between 30 and 27 points; 2) GOOD TO AVERAGE, scored between 26 and 22 points; 3) FAIR TO POOR, scored between 21 and 17 points, and 4) VERY POOR, scored between 16 and 13 points. The levels concern different degrees of task fulfilment in terms of knowledge, range and thoroughness in topic development, relevance, and length.

3.4.3.1.5 Organization

The organization of the composition was measured following the relevant scheme for assessing organization also in Jacobs et al.’s (1981) ESL Composition Profile. The scheme has four levels of organization quality: 1) EXCELLENT TO VERY GOOD, scored between 20 and 18 points; 2) GOOD TO AVERAGE, scored between 17 and 14 points; 3) FAIR TO POOR, scored between 13 and 10 points, and 4) VERY POOR, scored between 9 and 7 points. Such features as fluency in expression, clear presentation or support, succinctness, logic, and cohesion are examined.

3.4.3.2 Coding and reliability of coding

3.4.3.2.1 Coding fluency measures and coding reliability

I did all the coding except for syllables, which was done by a female soon-to-be graduate student majoring in English who was paid for her work. I first started with the
measures of fluency. To ensure no mistakes happened, I did the time calculation twice. I also counted the number of dysfluencies from the beginning of a composition to the end once, and then from the end to the beginning for a second time. If inconsistencies occurred, I would count for a third or even a fourth or fifth time. In word number counting, I paid special attention to excluding forms or symbols that did not deserve counting (e.g., Chinese characters) and was careful not to neglect digits, hyphenated or slash-divided words by searching in the Word Navigation panel for digits from zero to nine, hyphens, and slashes twice, before I finally obtained the word count statistics from the Microsoft Word Word Count function. I also made sure that occasional special punctuations (e.g., the Chinese dash) did not increase the number of words improperly.

The above mentioned research assistant counted the number of syllables for all the compositions. I first tested her and then sent her the scheme I had developed. She then coded one composition and sent it back to me. I checked her coding and gave her my feedback. We later had an online discussion, in which I highlighted the problem with diphthongs she seemed to have. She began to code another composition I then sent to her. This time, there was not any problem in her counting except for one word, *maturer* (\([\text{m}a\text{r}\text{ər}]\)), which she thought had four syllables but highlighted it in blue to show her uncertainty. We again discussed the diphthongs. I then told her whenever she had any doubt, she needed to consult the online resources I had listed for her, especially http://www.iciba.com/, where British pronunciations are given alongside American ones. After the initial training, she began to count all the remaining compositions. In her counting, she coded the first 20 compositions twice to ensure the quality of coding. She also referred to the dictionaries frequently for more accurate counting. After she finished the counting, she marked 10 misspellings which she thought could be problematic but for which she had made decisions (*chioce, encourage, sociaty, ohe, othe, howeve, aquair, ...*)
communicating, exercise, and poem). I told her my counts (2, 3, 3, 1, 1, 2, 2, 5, 2, and 2) and let her recode the related eight essays without referring to her previous counting. The intracoder reliability reached .999. I also paid another English major senior year university student to code a randomly selected pool of 12 compositions, which were evenly drawn across conditions (TA and NTA) and tasks (baseline and main), which I also coded. The interrater reliability reached .999 between the two research assistants, and 1.000 between either of the two research assistants and me.

3.4.3.2.2 Scoring content and organization and the reliability of scoring

I scored the contents of the participants’ essays and rated their organization, in separate softcopies. In content scoring, based on the requirements of the writing instructions, in each composition, I first identified the relevant content blocks relating to the past, present, change, and future, and highlighted them in different colors. In doing so, I bore in mind that a content block may occur in different paragraphs and different content blocks may be blended in the same paragraph. Then, I examined the lengths of different content blocks, against the word number requirements given in the instructions. I then further weighed the extent of knowledge, the level of detail, and the vividness displayed by a composition carefully and reiteratively, in reference to the rating scales at hand, before I finally reached a score for its content.

In scoring the organization, I examined coherence and cohesion carefully. I examined intra-sentential and inter-sentential cohesive ties, both grammatical and lexical (Halliday & Hasan, 1976), including, but not limited to, pronominal reference, lexical cohesion and conjunction. I also weighed the coherence of meaning thoughtfully. In the rating process, I marked good or bad/poor/lack of (or sometimes mediocre) organization on sentence boundaries and/or on organization-critical words, phrases, sentences, or lines in different colors, so that I could reach a more reliable decision by looking back
at the organization markers when I finished reading a composition. The scoring decision was finally made by weighing a participant's overall performance in organization, in close reference to the scoring schemes at hand.

Two female PhD students from the Faculty of Education, the University of Auckland had undertaken part of the content and organization scoring work in order to ensure the quality of my scoring. The first colleague directly entered the PhD program after she obtained a master's degree in applied linguistics in China. The second colleague was paid for her work and had been a university English teacher for 12 years, teaching English-major students courses such as intensive reading, writing, listening, stylistics, English for science and technology stylistics, translation, and college English for most of her past career. Both of them completed the content co-scoring work separately with me, but the first colleague did not go on with the organization co-scoring work.

For content scoring, I first asked both the inter-coders to familiarize themselves with the scoring scheme in Jacobs et al.'s (1981) ESL Composition Profile. Either of the two inter-coders and I discussed the scale in detail and in depth and made sure we reached agreement on how the wordings in the criteria could be understood and how the criteria could be related to Chinese university students’ general proficiency. We also agreed upon a working procedure for content scoring; that is, we would adopt a from-the-global-to-the-specific approach, which involved an initial evaluation of the overall length of a composition and the length of each required content block, followed by a thorough scrutiny of the finer sides of the content that were covered by Jacobs et al.’s rating scales. We then studied the writing instructions for both writing tasks carefully and I made sure either inter-coder understood what was required of the participants in terms of content. Next, we studied six sample essays which I had randomly selected from the pool of compositions and had given scores, before we had a discussion on the
contents of these writings and their scoring. Then, the scorer and I scored six essays selected from the pool of essays, independently. We then discussed our discrepancies and different understandings and resolved those issues of an uncertain nature (e.g., vividness of direct quotations, the extent of extra weight given to longer texts, etc.) and some of the peculiarities that arose (e.g., present and change described jointly in a single paragraph, only one aspect of one’s life described, etc.). Both of us then coded 32 compositions, randomly and equally drawn from each of the four task-condition combinations. The Pearson Product Moment correlation coefficient for the scores of the first coder and me reached .787, but that between the second coder and me reached .685 and that between my scores for the two ratings reached .877.

When I compared my scores (\(M = 21.625, SD = 2.860\), ranging from 16 to 27) with the second coder’s scores (\(M = 25.938, SD = 2.078\), ranging from 19 to 29), I found she tended to give higher and more centralized scores. I then compared our scores and found out that for 14 of the compositions our differences were greater than five points; two of these compositions were seven points apart, one was eight points apart, and one was different by nine points. I then read through some of the compositions that generated the greatest differences and discussed them with her. She admitted that she had not paid as much attention as I had done to the word requirements set in the writing instructions for different temporal descriptions, for she had scored the compositions in printouts while I had completed the scoring on their softcopies, in which case I could easily work out the lengths for the different parts. Another discrepancy concerned how essays that met the length requirements could be differentiated. We agreed that detailed, fine, and multifaceted descriptions should be awarded additional marks. Another agreement we reached was the maximum score we could give. Based on our exchanged understanding of Chinese non-English-major sophomore students’ English writing proficiency, we
decided that a score of 27 should not be given lightly and a score of 28 should be rare. The face-to-face discussion lasted about three hours. Then, the co-coder decided to code the compositions for a second time on the softcopies I sent her. This time, her scores ($M = 23.313$, $SD = 2.923$, ranging from 19 to 27) and mine reached an interrater reliability rate of .806. Then, at a time convenient for both of us, we had an online discussion on our scores, especially the 11 essays which showed differences greater than 2 points. Both of us modified our scores. I raised the scores for two compositions each by 1 point ($M = 21.688$, $SD = 2.867$, ranging from 16 to 27). She modified her scores for 11 essays by different degrees ($M = 22.406$, $SD = 3.057$, ranging from 17 to 27). The Pearson Product Moment correlation coefficient showed a relevance of .950 then. The discussion lasted about two hours. I then began to score the rest of 132 essays in their softcopies.

Our mutual scoring for determining the reliability of the organization coding followed a similar procedure. The work was started with an approximately five-day interval after the content co-scoring work. The second co-coder and I first familiarized ourselves with the organization scoring scheme in Jacobs et al.’s (1981) ESL Composition Profile. We then discussed the scales carefully, making sure that both of us understood the benchmarks properly and shared a similar understanding of them. We also agreed that when we scored the compositions in terms of their organization, we should read them carefully, sentence by sentence, and mark every place of interest (e.g., sentence boundaries, pronouns, discourse markers, etc.) by putting either a tick or a cross nearby (in the printouts), indicating good or poor organization, or by highlighting it in different colors indicating different qualities of organization (in the softcopies). In this way, we would make a reliable final decision based on a review of organization performance in the interesting places.
At the beginning of coding, I gave the inter-rater four essays for trial scoring, which I had scored. She gave very close scores. We then discussed these essays. After this trial scoring, we went on to score three more essays independently. Our scores turned out to be similar and close. Subsequently, we held a discussion and exchanged notes. We then scored the same 32 manuscripts as we had used for content co-scoring. When my scores ($M = 15.375$, $SD = 2.366$, ranging from 9 to 18) were compared with hers ($M = 14.250$, $SD = 1.437$, ranging from 11 to 17), the Pearson Product Moment correlation coefficient reached .456. We then spent a whole afternoon discussing the scores, those bearing a difference of more than 2 points in particular. There were eight such divergent compositions, with the greatest difference being 6 points. In our discussion, we agreed that the length problem should be taken mainly as a content issue. I also briefed the inter-coder on what connectives should have been given a bonus, for this group of non-English-major students (e.g., *even though*, *surprisingly*, etc.), and what means of organization should be taken as indicative of an average level (e.g., *firstly*, *secondly*). Our long discussion on the 32 compositions we had rated involved debates and reflections. Finally, altogether, I made a minor adjustment of four scores each by one point among my scores ($M = 15.250$, $SD = 2.171$, ranging from 10 to 18), while the inter-coder made eight adjustments in her scores ($M = 14.719$, $SD = 1.800$, ranging from 11 to 18). Especially, she accepted my score that bore a 6-point difference from her original score, and rescored one composition so that our score difference for that particular composition was narrowed down from five points to just one point. After this rescoreing, the Pearson Product Moment correlation coefficient reached .844. I then began to score the remaining 131 essays in their softcopies.

An intra-rater reliability index was also calculated for my content and organization scoring work after I later finished coding the complexity measures. I randomly drew 17
compositions (approximately 10% of the total number of compositions, evenly distributed across condition and task except for one composition) from the total number of compositions that were not included in the inter-coding and rated their contents and organization in blind copies. Pearson Product Moment correlation coefficients were then calculated as indexes for intra-rater reliability. The reliability for content scoring reached .765 and that for organization scoring reached .889.

I then checked the scores and resolved the differences between the two scores that were greater than two. Especially, in Composition 29, there was a whole paragraph where the participant wrote about how a father’s love differed from a mother’s love. In my previous scoring, I had thought it was seriously off topic, so I gave 16 points and rated it as a Very Poor composition, other parts considered. In this re-scoring, I thought it was somewhat related to the instructions, since it indicated a reflection. I modified the score to 19. I then began to examine all the compositions that contained parts that I had thought to be off topic, which I had not highlighted but marked in red. I identified 20 such compositions that contained off topic elements and re-scored them. The Pearson Product Moment correlation coefficient was calculated as the intra-rater reliability for coding the 20 essays, which reached .875. I then carefully examined the discrepancies between the new scores and the corresponding previous scores of these compositions and made final decisions on the appropriate scores.

3.4.3.2.3 Scoring complexity and accuracy measures and the reliability of scoring

I coded complexity and accuracy, which included such constructs as t-units, finite clauses, nonfinite clauses, the lexical measure D, different verb forms, errors, erroneous clauses, and wrong verb forms. To facilitate calculating statistics and modifications, I used a digit code to represent each construct or measure that was involved in coding complexity and accuracy.
The series of coding was completed in the same documents, one construct/measure after another, so that any preceding coding could be checked repetitively when the subsequent coding was being carried out. During my coding, I developed separate coding schemes based on previous work (e.g., Polio, 1997; Polio & Shea, 2014; Yang et al., 2014) and kept complementing or modifying them as my coding progressed. I paid special attention to cases that had not been covered by previous literature.

In coding different verb forms, I identified all finite verbs, which included verbs that did constitute finite clauses and those that did not, and nonfinite ones, which included all different forms of nonfinite verbs, and examined them case by case. Whenever a different verb form occurred, I recorded it in a chart, with great care.

I completed the first round of error coding painstakingly, by referring to dictionary resources and other online resources frequently. I coded erroneous clauses and erroneous finite verb forms along with individual errors. During the process, I developed an initial coding scheme based on Polio (1997) and Polio and Shea (2014).

To calculate D values for lexical diversity, I first downloaded the CLAN programs (MacWhinney, 2000a, 2000b) from the CHILDES web site (http://childes.psy.cmu.edu). I then examined all the softcopies of the participants’ compositions, converted all the words into lemmas (e.g., went changed to go), and corrected misspelt words, before I transformed and embedded all the processed compositions into the standard CHAT format (Codes for the Human Analysis of Transcripts) of the CHILDES project (Child Language Data Exchange System) (MacWhinney & Snow, 1990). I made sure that the scripts could be recognized and processed by the software. I then ran the Vocd command and yielded D values.

For calculating error coding interrater reliability, two PhD students from Faculty of Education, the University of Auckland counted errors for a portion of the participants’
compositions first. One of them was paid by the hour; the other received my promise to code reciprocally. Both coded a total of four essays following a similar procedure: 1) the inter-coder studied with me a brief guide to error coding, together with examples; 2) the inter-coder was introduced to the several resources or dictionaries I had used in my coding, which included British National Corpus (http://www.natcorp.ox.ac.uk), to which I had made a donation, Google, Longman English Dictionary Online (www.ldoceonline.com), and www.iciba.com, an online English-Chinese dictionary; 3) the inter-coder was given two compositions randomly chosen from the pool of compositions to code as a trail; 4) the inter-coder and I compared our coding and discussed our error counts one by one; 5) the inter-coder and I studied the detailed coding scheme I had developed during my first round of coding the 170 essays; 6) the inter-coder went on to code another two randomly selected essays; 7) I compared our results and then the inter-coder and I held a discussion to resolve the discrepancies. The female PhD student was retained for the remaining work of inter-coding. Coming from China as I did, she completed her bachelor’s degree in English at one of the top five universities in China and then stayed at the university as an English teacher. Five years later, she obtained her master’s degree in linguistics and applied linguistics from the same university she worked in. And at the time of my study she is a second-year full-time PhD student researching second language writing while retaining her affiliation to her university. She had taught reading and writing for five years before she taught listening and speaking for seven years, mainly to non-English-major freshmen. She has many years’ experience in rating College English Test essays.

I first emailed her two essays for trial coding, together with a brief introduction to how to code. When she finished her coding, we discussed our work face to face. During our discussion, we talked about the errors we each coded one by one. Especially, we
addressed the discrepancies we had, which included lack of double/multiple counting, difficult/obscure sentences, or some other illusive points.

After the discussion, I gave her the error coding scheme I had developed after I finished the first round of error coding. She and I studied the coding schemes very carefully, item by item, and example by example. I then sent her two essays for continued coding. When she finished, we held a long discussion on the discrepancies we had and reached an agreement for each case of discrepancy. When all discrepancies were resolved, I sent her 24 randomly selected compositions to code, which covered the two conditions and tasks evenly. It took her two weeks to finish coding them. The Pearson Product Moment correlation coefficient for the number of errors we coded for the 24 essays reached .960. I then examined our coding of errors, and after that wrote the results of my double check to her, including notes on all our discrepancies. She offered feedback, and I made corresponding decisions based on our exchanges.

I then began to re-code the remaining 132 essays, which took me about 40 days.

In this second round coding of errors, changes were also made as regards $t$-units, finite clause, non-finite clauses, and other measurement-related units. Some of the changes were due to my negligence (e.g., so let’s stop here), some due to re-coding of finite clauses as nonfinite ones, some due to re-counting of infinitives in be able to, would like to or similar constructions, and others due to exclusion of independently appearing -ing forms that have become established nouns (smoking, drinking, etc.).

I then calculated Pearson Product Moment correlation coefficients for indexes of intra-rater reliability for my coding of finite clauses, nonfinite clauses, $t$-units, EFCs, correct verb forms, and errors. I compared the finalized figures for these constructs in the 170 essays with the corresponding statistics I had obtained before I coded errors for the second time.
Table 3.3 shows the Pearson Product Moment correlation coefficients calculated for all the comparisons.

Table 3.3

Intrarater Reliability: Pearson Product Moment Correlation Coefficients

<table>
<thead>
<tr>
<th>Item</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of finite clauses</td>
<td>.998</td>
</tr>
<tr>
<td>Number of nonfinite clauses</td>
<td>.978</td>
</tr>
<tr>
<td>Number of T-units</td>
<td>.993</td>
</tr>
<tr>
<td>Number of EFCs</td>
<td>.951</td>
</tr>
<tr>
<td>Number of wrong verb forms</td>
<td>.942</td>
</tr>
<tr>
<td>Number of errors</td>
<td>.974</td>
</tr>
</tbody>
</table>

3.4.4 Analyzing/Coding answers to reflection questions and verbal reports

The participants’ answers to the three reflection questions on the reactivity, completeness, and accuracy of their CVRs were analysed following a content analysis approach. Based on this approach, I abstracted their yes or no answers, their perceived extent of and reasons for the (non)reactivity, (in)completeness or (in)accuracy of CVRs, the aspects of writing processes or performance where they felt reactivity had occurred, and the aspects of their CVRs where they felt incompleteness or inaccuracy may have happened. The participants’ misplaced responses were also analysed and allocated to the right categories.

The research assistant who had done syllable counting was paid to transcribe the participants’ retrospective and CVRs regarding the places of revision. She was required to make a detailed chart in the Excel format that presented a clear comparison of the contents of the two types of reports relating to all points of revision, one by one, and made a note if there was any particularity.

I then contrasted the juxtaposed contents of both types of reports, studied them carefully, and made initial decisions on the relative completeness of CVRs based on the information gaps revealed. After this initial coding, I started a second round of coding by listening to the participants’ CVRs carefully and repeatedly, in reference to the
participants’ RVRs and their written texts, modifying or complementing the research assistant’s transcripts whenever necessary. I inferred what I thought should be in the student writers’ minds and what they should report, taking into consideration of the contents of both types of reports in reference to the revisions they made and the linguistic contexts where the revisions occurred. During this re-coding process, I developed three categorizations of incomplete concurrent verbal reports (i.e., lack of information on what one had intended to write, lack of information on what one actually wrote, and lack of reporting on metacognitive thinking) and identified cases of complete reporting. When listening to the participants’ RVRs and reading the transcripts of the RVRs, I was aware that the participants’ RVRs might not be reliable, given the possible memory decay due to the elapse of time, and given their sometimes descriptive and explanatory nature, so I treated the RVRs with discretion. When the second round of work finished, I checked my judgments throughout the transcripts again, and would listen to the recordings repeatedly whenever doubt arose. It should be noted that not all points of revision were reported on, especially those that were viewed as minor ones or escaped from the participants’ notice, given that the retrospective verbal reporting was completely done autonomously by the participants.
CHAPTER 4 REACTIVITY: RESULTS FROM QUANTITATIVE ANALYSES OF PERFORMANCE DATA

This chapter presents the results of my quantitative analyses of data that were obtained by comparing the performance of the TA group and the NTA group. It starts with an overview of the methods and procedures I used for my data analyses. Then it presents the results of my preparatory analyses that addressed data normality, prior group differences in WM and (baseline) writing performance, and correlations (by linear regression) of WM or baseline writing performance scores with main writing performance scores. Next, it reports the results of my main analyses that examined group differences in main task performance scores and the interactions of WMC or baseline writing performance scores with the reactive effects. The chapter ends with a summary of the results from the quantitative analyses.

4.1 Procedures and Methods for Data Analyses

My data analyses involved the following major procedures and methods:

1. WMC scores, which included WMC One scores (the operation span scores in percentage), WMC Two scores (the reading span scores in percentage), and WMC Average scores (the average of the operation and reading span scores in percentage) of the TA group and the NTA group were compared using a series of t-tests, to examine if there was any statistically significant group difference between the two groups in terms of their WMC;

2. Participants were grouped into high-, mid- and low-WMC groups based on the converted z-scores of their WMC Average, WMC One, or WMC Two, and the groupings were checked by conducting Brown-Forsythe and Dunnett’s T3 post hoc pairwise comparisons;
3. Participants were then grouped based on their baseline writing scores on the 20
dependent variables examined, and the 20 groupings were checked likewise;

4. A series of regression analyses of WMC Average, WMC One, or WMC Two on
main task writing scores, and a series of regression analyses of baseline writing
scores on main task writing scores were conducted;

5. The differences between the TA and NTA groups in baseline performance in all
the 20 aspects were checked, by running a series of t-tests, after the baseline data
normality was checked, based on the Skewness and the Kurtosis of the data and
the results of both the Kolmogorov-Smirnov and Shapiro-Wilk tests;

6. The normality of all main task writing scores were checked likewise;

7. A series of initial one-way ANCOVAs was run on the main task data, controlling
for the baseline data and WMC Average, for the baseline data and WMC One,
and for the baseline data and WMC Two, followed by pairwise comparisons
through the LSD procedure, using a custom model including Condition (TA or
NTA), the two covariates (the baseline data and one of the three WMC scores),
and the two covariate × Condition interaction terms;

8. Where there was no covariate × Condition interaction found in 7), a one-way
ANCOVA was run with only one covariate × Condition interaction term kept,
before another one-way ANCOVA was run with the other covariate × Condition
interaction term, for confirmation;

9. If still no significant interaction was found between either of the two covariates
and Condition, the data were then re-analysed by running an ANCOVA in the
full factorial mode, without any interaction term, to obtain final results;

10. Where there was statistically significant interaction found between either of the
two covariates and Condition in the initial ANCOVA operations in 7), a
succession of two one-way ANCOVAs were run, keeping one covariate $\times$ Condition interaction term while excluding the other, to confirm the interaction, or lack thereof, found in the initial ANCOVAs, and the results of between-groups comparisons were reported as were found in the ANCOVA analysis showing the interaction;

11. A scatterplot was then drawn, in the case of an interaction, to illustrate the differential effects of Condition on participants of different traits where the interaction was found;

12. A further analysis of the interactional effect was conducted, by running an ANCOVA to compare three subgroups (high, mid, and low) grouped based on the measure or type of WMC which interacted with Condition, with the main task data for the measure concerned as the dependent variable, Condition and grouping as the independent variables, the corresponding baseline data and WMC Average, or WMC One, or WMC Two as two covariates, keeping the Condition $\times$ Grouping interaction term only, to examine if there was significant Condition $\times$ Grouping interaction, and post-hoc between- and within-condition comparisons of subgroups were also carried out to examine subgroup differences, especially those across conditions;

13. A further but finer analysis of the interaction effect was then conducted to compare the two conditions at seven different levels (i.e., the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles) of the baseline measure or of WMC (WMC Average, WMC One, or WMC Two), which interacted with Condition;

14. Where homogeneity of variance was not met in the above ANCOVA analyses, data would be re-examined and outliers in relevant data would be removed, if applicable, to ensure normal distribution and satisfaction of the assumption of
homogeneity of variance before further analyses were conducted again for confirmation and/or modification, restarting from 7);

15. To confirm the main effects of TA found in the above series of ANCOVA analyses, a series of ANCOVAs was run, with the baseline scores as the only covariate where WMC (Average, One, or Two) did not show a significant correlation with main task performance, following the same procedures described between 7) and 14), wherever applicable, and,

16. A series of t-tests was also run, to compare the main task data across the two conditions.

The alpha for achieving statistical significance was set at .05. In all ANCOVA operations, Type III sum of squares were used. Partial eta squared ($\eta^2$) was calculated. The effect sizes of pairwise comparisons were also calculated by dividing the adjusted mean difference by the square root of $MS_{\text{error}}$. In terms of the magnitude of effects, following Cohen’s (1988) benchmarks, $\eta^2 = .01$ was defined as small, $\eta^2 = .06$ as medium, and $\eta^2 = .14$ as large, and $d = 0.2$ was defined as small, $d = 0.5$ as medium, and $d = 0.8$ as large.

### 4.2 Independent Variable: Working Memory Scores

Table 4.1 presents the statistics of WMC One and WMC Two, both in raw scores and in percentages, and WMC Average percentage scores.

<table>
<thead>
<tr>
<th>WMC</th>
<th>N</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMC 1</td>
<td>85</td>
<td>92.059</td>
<td>10.723</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.852</td>
<td>.099</td>
</tr>
<tr>
<td>WMC 2</td>
<td>85</td>
<td>53.412</td>
<td>16.344</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.712</td>
<td>.218</td>
</tr>
<tr>
<td>WMC AVG</td>
<td>85</td>
<td>.782</td>
<td>.137</td>
</tr>
</tbody>
</table>
The correlation between WMC One and WMC Two scores was Pearson’s $r = .413$ ($p = .001$).

Table 4.2 presents the WMC Average scores of both the NTA group and the TA group and the results of an independent samples $t$-test. The scores of both the groups met normal distribution (NTA: Skewness = -.721, Kurtosis = .222; NMTA: Skewness = -.714, Kurtosis = -.387).

As Table 4.2 shows, there was no statistically significant difference between the NTA group ($M = .780$, $SD = .129$) and the TA group ($M = .784$, $SD = .146$) in terms of their WMC Average, $t (83) = .142$, $p = .887$. This indicates that both groups began on an equal footing in terms of WMC Average.

Table 4.2

<table>
<thead>
<tr>
<th>Groups</th>
<th>WMC AVG (M, SD)</th>
<th>t-test</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA (N = 43)</td>
<td>.784, .146</td>
<td>0.142</td>
<td>.887</td>
</tr>
<tr>
<td>NTA (N = 42)</td>
<td>.780, .129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The WMC One and WMC Two scores of the TA group and the NTA group were also compared, separately.

Table 4.3

<table>
<thead>
<tr>
<th>Groups</th>
<th>WMC 1 (Operation Span)</th>
<th>M, SD</th>
<th>t-test</th>
<th>p</th>
<th>WMC 2 (Reading Span)</th>
<th>M, SD</th>
<th>t-test</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA (N = 43)</td>
<td>.717, .229</td>
<td>.202</td>
<td>.841</td>
<td>.960</td>
<td>.852, .107</td>
<td>-.051</td>
<td>.960</td>
<td></td>
</tr>
<tr>
<td>NTA (N = 42)</td>
<td>.707, .209</td>
<td></td>
<td>.853, .093</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 lists the descriptive statistics of these two WMC scores and the results of $t$-tests. As Table 4.3 shows, there was no statistically significant difference between the TA group and the NTA group in either WMC One or WMC Two.
Participants were grouped into three levels of WMC groups, based on their WMC Average, WMC One, and WMC Two z-scores, following the -.5 and .5 benchmarks. Table 4.4 presents the distribution of participants of different levels of WMC Average, WMC One and WMC Two in either experimental condition.

Table 4.4

Distribution of Participants Having Three Levels of WMC

<table>
<thead>
<tr>
<th>Groups</th>
<th>WMC AVG High</th>
<th>WMC AVG Mid</th>
<th>WMC AVG Low</th>
<th>WMC 1 High</th>
<th>WMC 1 Mid</th>
<th>WMC 1 Low</th>
<th>WMC 2 High</th>
<th>WMC 2 Mid</th>
<th>WMC 2 Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>N = 15</td>
<td>N = 10</td>
<td>N = 17</td>
<td>N = 15</td>
<td>N = 11</td>
<td>N = 16</td>
<td>N = 14</td>
<td>N = 13</td>
<td></td>
</tr>
<tr>
<td>NTA</td>
<td>N = 13</td>
<td>N = 19</td>
<td>N = 10</td>
<td>N = 17</td>
<td>N = 13</td>
<td>N = 12</td>
<td>N = 16</td>
<td>N = 14</td>
<td>N = 12</td>
</tr>
</tbody>
</table>

Note. TA: N = 43; NTA: N = 42.

To confirm the representativeness of these groupings, I conducted Brown-Forsythe and Dunnett’s T3 post hoc pairwise comparisons due to violation of homogeneity of variance. Statistical significance was found in all tests: for WMC Average, Brown-Forsythe $F (2, 36.580) = 229.440$, for WMC One, Brown-Forsythe $F (2, 39.878) = 206.120$, and for WMC Two, Brown-Forsythe $F (2, 48.658) = 210.646$, all $p = .000$, and all pairwise comparisons using Dunnett’s T3 were significant ($p = .000$ for all pairwise comparisons), indicating that all the groupings were valid in distinguishing the participants in terms of their WMC.

4.3 Quantitative Analyses of Dependent Variables

This section reports the results of main quantitative analyses. It first presents the descriptive statistics concerning all the 20 measures as regards fluency, complexity, accuracy, contents, and organization. Then, it provides three groupings of the participants (High, Mid, and Low) in each of these 20 aspects of their baseline writing performance. Next, it presents the results of my regression analyses, which examined the correlations between the possible covariates (WMC Average/WMC One/WMC Two and baseline writing scores) and main task performance scores. After that, it reports the
results of a series of $t$-tests which examined the prior differences between the TA and NTA groups in baseline writing performance, to check group homogeneity, and to establish a basis for use of covariates, given that either significant correlation or prior between-group difference is a prerequisite for including a covariate into an ANCOVA analysis. Finally but most importantly, the section presents the results of my ANCOVA analyses, which compared the main task performance scores of the two groups in each of the 20 aspects, with each aspect of baseline writing scores as a covariate and WMC Average, or WMC One, or WMC Two as the other covariate, while examining the possible interactions between either of the covariates and Condition.

### 4.3.1 Descriptive statistics of baseline and main task writing performance

Table 4.5 presents the descriptive statistics (i.e., means, standard deviations, and minimum and maximum values) for the baseline data.

**Table 4.5**

*Descriptive Statistics for Baseline Data*

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>NTA</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$</td>
<td>$SD$</td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>41.357</td>
<td>9.432</td>
</tr>
<tr>
<td>Speed 1</td>
<td>8.765</td>
<td>2.642</td>
</tr>
<tr>
<td>ND</td>
<td>.911</td>
<td>.051</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length 1</td>
<td>342.810</td>
<td>52.403</td>
</tr>
<tr>
<td>Length 2</td>
<td>365.683</td>
<td>57.105</td>
</tr>
<tr>
<td>Length 3</td>
<td>462.786</td>
<td>74.360</td>
</tr>
<tr>
<td>Length 4</td>
<td>490.171</td>
<td>79.492</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>17.357</td>
<td>3.122</td>
</tr>
<tr>
<td>LD (D)</td>
<td>64.291</td>
<td>13.632</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFWs</td>
<td>89.674</td>
<td>4.030</td>
</tr>
<tr>
<td>EFCs</td>
<td>.569</td>
<td>.125</td>
</tr>
<tr>
<td>CVU</td>
<td>.825</td>
<td>.099</td>
</tr>
<tr>
<td>Content</td>
<td>23.143</td>
<td>2.415</td>
</tr>
<tr>
<td>Organization</td>
<td>15.238</td>
<td>2.229</td>
</tr>
</tbody>
</table>

*Note.* NTA: $N = 42$, except Speed 2, 4, ND, Length 2, and 4, $N = 41$; TA: $N = 43$. 

124
The statistics provide an overview of the performance of the TA group vis-à-vis the NTA group at the start when both groups wrote silently. The differences between the two groups in their baseline performance were further tested in 4.3.4.2.

Table 4.6 presents the descriptive statistics (i.e., means, standard deviations, and minimum and maximum values) for the main task data.

Table 4.6

Descriptive Statistics for Main Task Data

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>NTA</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>40.902</td>
<td>8.040</td>
</tr>
<tr>
<td>Speed 1</td>
<td>9.136</td>
<td>2.548</td>
</tr>
<tr>
<td>ND</td>
<td>.921</td>
<td>.041</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length 1</td>
<td>356.000</td>
<td>59.171</td>
</tr>
<tr>
<td>Length 2</td>
<td>375.738</td>
<td>66.010</td>
</tr>
<tr>
<td>Length 3</td>
<td>482.024</td>
<td>80.813</td>
</tr>
<tr>
<td>Length 4</td>
<td>506.000</td>
<td>89.264</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>2.038</td>
<td>.297</td>
</tr>
<tr>
<td>LD (D)</td>
<td>66.913</td>
<td>12.112</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFWs</td>
<td>89.788</td>
<td>4.008</td>
</tr>
<tr>
<td>EFCs</td>
<td>.554</td>
<td>.130</td>
</tr>
<tr>
<td>CVU</td>
<td>.814</td>
<td>.083</td>
</tr>
<tr>
<td>Content</td>
<td>22.095</td>
<td>3.392</td>
</tr>
<tr>
<td>Organization</td>
<td>15.738</td>
<td>1.578</td>
</tr>
</tbody>
</table>

Note. NTA: N = 42, except Time, Speed 1, 2, 3, and 4, N = 41; TA: N = 43.

The statistics provide an overview of the performance of the TA group vis-à-vis the NTA group when the two groups wrote under their corresponding conditions. The differences between the two groups in performing the main writing task were further tested in 4.3.5.
4.3.2 Grouping participants based on baseline writing performance

Like previous WMC groupings, the z-scores for baseline writing performance scores were used and the demarcations (lower than -.5, between -.5 and .5, and greater than .5) were followed for grouping participants in the 20 dimensions of writing performance.

Table 4.7 lists the distribution of high-achievers, mid-achievers, and low-achievers, both TA and NTA, in each of the 20 aspects of baseline writing performance.

Table 4.7

Distribution of High-, Mid-, and Low-Achievers in 20 Groupings

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>NTA High (N)</th>
<th>Mid (N)</th>
<th>Low (N)</th>
<th>TA High (N)</th>
<th>Mid (N)</th>
<th>Low (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>8</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Speed 1</td>
<td>9</td>
<td>14</td>
<td>19</td>
<td>15</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>Speed 2</td>
<td>8</td>
<td>13</td>
<td>20</td>
<td>13</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Speed 3</td>
<td>8</td>
<td>15</td>
<td>19</td>
<td>16</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Speed 4</td>
<td>7</td>
<td>17</td>
<td>17</td>
<td>15</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>ND</td>
<td>12</td>
<td>19</td>
<td>10</td>
<td>16</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>Length 1</td>
<td>17</td>
<td>9</td>
<td>16</td>
<td>11</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Length 2</td>
<td>17</td>
<td>10</td>
<td>14</td>
<td>13</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>Length 3</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>13</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Length 4</td>
<td>17</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>15</td>
<td>14</td>
<td>13</td>
<td>8</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>General complexity</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>14</td>
<td>19</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>13</td>
<td>12</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>SV</td>
<td>15</td>
<td>16</td>
<td>11</td>
<td>8</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>LD (D)</td>
<td>14</td>
<td>16</td>
<td>12</td>
<td>8</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>EFWs</td>
<td>15</td>
<td>18</td>
<td>9</td>
<td>11</td>
<td>21</td>
<td>11</td>
</tr>
<tr>
<td>EFCs</td>
<td>16</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>21</td>
<td>14</td>
</tr>
<tr>
<td>CVU</td>
<td>17</td>
<td>13</td>
<td>12</td>
<td>14</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>Content</td>
<td>15</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Organization</td>
<td>15</td>
<td>19</td>
<td>8</td>
<td>9</td>
<td>21</td>
<td>13</td>
</tr>
</tbody>
</table>

Note. NTA: N = 42, except Speed 2, 4, ND, Length 2, and 4, N = 41; TA: N = 43.

To confirm the representativeness of those groupings, Brown-Forsythe and Dunnett’s T3 post hoc pairwise comparisons were conducted. Statistical significance was found in all tests for the 20 groupings based on the 20 aspects of performance, all \( p = .000 \), for all comparisons.

4.3.3 The results of the regression analyses

Table 4.8 presents the results of the regression analyses of WMC Average on main task performance.
As is shown in Table 4.8, WMC Average significantly predicted the L2 learner writers’ main task performance on five temporal measures, Time, $F(1, 82) = 8.106, p = .006$, Speed 1, $F(1, 82) = 4.433, p = .038$, Speed 2, $F(1, 82) = 4.858, p = .030$, Speed 3, $F(1, 82) = 6.025, p = .016$, and Speed 4, $F(1, 82) = 6.368, p = .014$, but not on other measures.

Table 4.8

**Results of Regression Analyses of WMC Average on Main Task Performance**

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Simple linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fluency</strong></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>$F(1, 82) = 8.106, p = .006^*, R = .300, R^2 = .090$, adjusted $R^2 = .079$</td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F(1, 82) = 4.433, p = .038^*, R = .226, R^2 = .051$, adjusted $R^2 = .040$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F(1, 82) = 4.858, p = .030^*, R = .237, R^2 = .056$, adjusted $R^2 = .044$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F(1, 82) = 6.025, p = .016^*, R = .262, R^2 = .068$, adjusted $R^2 = .057$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F(1, 82) = 6.368, p = .014^*, R = .268, R^2 = .072$, adjusted $R^2 = .061$</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>$F(1, 83) = 1.358, p = .247, R = .127, R^2 = .016$, adjusted $R^2 = .004$</td>
</tr>
<tr>
<td>Length 1</td>
<td>$F(1, 83) = 0.142, p = .708, R = .041, R^2 = .002$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td>Length 2</td>
<td>$F(1, 83) = 0.017, p = .898, R = .014, R^2 = .000$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td>Length 3</td>
<td>$F(1, 83) = 0.133, p = .716, R = .040, R^2 = .002$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td>Length 4</td>
<td>$F(1, 83) = 0.257, p = .614, R = .056, R^2 = .003$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td><strong>Syntactic complexity</strong></td>
<td></td>
</tr>
<tr>
<td>General complexity</td>
<td>$F(1, 83) = 1.573, p = .213, R = .136, R^2 = .019$, adjusted $R^2 = .007$</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>$F(1, 83) = 1.112, p = .295, R = .115, R^2 = .013$, adjusted $R^2 = .001$</td>
</tr>
<tr>
<td>SV</td>
<td>$F(1, 83) = 1.222, p = .272, R = .120, R^2 = .015$, adjusted $R^2 = .003$</td>
</tr>
<tr>
<td>LD (D)</td>
<td>$F(1, 83) = 0.172, p = .680, R = .045, R^2 = .002$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
</tr>
<tr>
<td>EFWs</td>
<td>$F(1, 83) = 1.686, p = .198, R = .141, R^2 = .020$, adjusted $R^2 = .008$</td>
</tr>
<tr>
<td>EFCs</td>
<td>$F(1, 83) = 1.476, p = .228, R = .132, R^2 = .017$, adjusted $R^2 = .006$</td>
</tr>
<tr>
<td>CVU</td>
<td>$F(1, 83) = 2.005, p = .161, R = .154, R^2 = .024$, adjusted $R^2 = .012$</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F(1, 83) = 0.077, p = .782, R = .030, R^2 = .002$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$F(1, 83) = 2.934, p = .090, R = .185, R^2 = .034$, adjusted $R^2 = .023$</td>
</tr>
</tbody>
</table>

* means statistical significance was found on the simple linear regression of WMC Average on a measure.

Table 4.9 presents the results of the regression analyses of WMC One on main task performance.

As is shown in the table, WMC One, the operation span, significantly predicted the time spent on the main task only, among the 20 measures of main task writing performance, $F(1, 82) = 6.461, p = .013$. 

127
Table 4.9

Results of Regression Analyses of WMC One on Main Task Performance

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Simple linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fluency</strong></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>$F(1, 82) = 6.461, p = .013^*$, $R = .270$, $R^2 = .073$, adjusted $R^2 = .062$</td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F(1, 82) = 1.676, p = .199$, $R = .142$, $R^2 = .020$, adjusted $R^2 = .008$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F(1, 82) = 1.855, p = .177$, $R = .149$, $R^2 = .022$, adjusted $R^2 = .010$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F(1, 82) = 2.839, p = .096$, $R = .183$, $R^2 = .033$, adjusted $R^2 = .022$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F(1, 82) = 2.957, p = .089$, $R = .187$, $R^2 = .035$, adjusted $R^2 = .023$</td>
</tr>
<tr>
<td>ND</td>
<td>$F(1, 83) = 0.666, p = .417$, $R = .089$, $R^2 = .008$, adjusted $R^2 = -0.004$</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td></td>
</tr>
<tr>
<td>Length 1</td>
<td>$F(1, 83) = 1.607, p = .208$, $R = .138$, $R^2 = .019$, adjusted $R^2 = .007$</td>
</tr>
<tr>
<td>Length 2</td>
<td>$F(1, 83) = 1.078, p = .302$, $R = .113$, $R^2 = .013$, adjusted $R^2 = .001$</td>
</tr>
<tr>
<td>Length 3</td>
<td>$F(1, 83) = 0.137, p = .712$, $R = .041$, $R^2 = .002$, adjusted $R^2 = -0.010$</td>
</tr>
<tr>
<td>Length 4</td>
<td>$F(1, 83) = 0.070, p = .791$, $R = .029$, $R^2 = .001$, adjusted $R^2 = -0.011$</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>$F(1, 83) = 1.394, p = .241$, $R = .129$, $R^2 = .017$, adjusted $R^2 = .005$</td>
</tr>
<tr>
<td>General complexity</td>
<td>$F(1, 83) = 0.881, p = .351$, $R = .102$, $R^2 = .011$, adjusted $R^2 = -0.001$</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>$F(1, 83) = 0.535, p = .466$, $R = .080$, $R^2 = .006$, adjusted $R^2 = -0.006$</td>
</tr>
<tr>
<td>SV</td>
<td>$F(1, 83) = 0.630, p = .430$, $R = .087$, $R^2 = .008$, adjusted $R^2 = -.004$</td>
</tr>
<tr>
<td>LD (D)</td>
<td>$F(1, 83) = 0.041, p = .841$, $R = .022$, $R^2 = .000$, adjusted $R^2 = -.012$</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td></td>
</tr>
<tr>
<td>EFWs</td>
<td>$F(1, 83) = 1.361, p = .247$, $R = .127$, $R^2 = .016$, adjusted $R^2 = .004$</td>
</tr>
<tr>
<td>EFCs</td>
<td>$F(1, 83) = 1.832, p = .180$, $R = .147$, $R^2 = .022$, adjusted $R^2 = .010$</td>
</tr>
<tr>
<td>CVU</td>
<td>$F(1, 83) = 1.375, p = .244$, $R = .128$, $R^2 = .016$, adjusted $R^2 = .004$</td>
</tr>
<tr>
<td>Content</td>
<td>$F(1, 83) = 0.994, p = .322$, $R = .109$, $R^2 = .012$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td>Organization</td>
<td>$F(1, 83) = 1.517, p = .222$, $R = .134$, $R^2 = .018$, adjusted $R^2 = .006$</td>
</tr>
</tbody>
</table>

* means statistical significance was found on the simple linear regression of WMC One on a measure.

Table 4.10 presents the results of the regression analyses of WMC Two on main task performance.

As is shown, WMC Two significantly predicted the learners’ main task performance on all five time-related measures, Time, $F(1, 82) = 4.776, p = .032$, Speed 1, $F(1, 82) = 9.110, p = .003$, Speed 2, $F(1, 82) = 9.893, p = .002$, Speed 3, $F(1, 82) = 9.491, p = .003$, Speed 4, $F(1, 82) = 10.224, p = .002$, one complexity measure, Length 4, $F(1, 83) = 4.113, p = .046$, and the organization measure, $F(1, 83) = 4.079, p = .047$. It should be noted that the correlation for other length measures, Length 2 in particular, demonstrated a tendency approaching significance.
Table 4.10

Results of Regression Analyses of WMC Two on Main Task Performance

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Simple linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>$F(1, 82) = 4.776, p = .032^*, R = .235, R^2 = .055$, adjusted $R^2 = .044$</td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F(1, 82) = 9.110, p = .003^*, R = .316, R^2 = .100$, adjusted $R^2 = .089$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F(1, 82) = 9.893, p = .002^*, R = .328, R^2 = .108$, adjusted $R^2 = .097$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F(1, 82) = 4.911, p = .003^*, R = .322, R^2 = .104$, adjusted $R^2 = .093$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F(1, 82) = 10.224, p = .002^*, R = .333, R^2 = .111$, adjusted $R^2 = .100$</td>
</tr>
<tr>
<td>ND</td>
<td>$F(1, 83) = 2.033, p = .158, R = .155, R^2 = .024$, adjusted $R^2 = .012$</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
</tr>
<tr>
<td>Length 1</td>
<td>$F(1, 83) = 3.057, p = .084, R = .188, R^2 = .036$, adjusted $R^2 = .024$</td>
</tr>
<tr>
<td>Length 2</td>
<td>$F(1, 83) = 3.810, p = .054, R = .210, R^2 = .044$, adjusted $R^2 = .032$</td>
</tr>
<tr>
<td>Length 3</td>
<td>$F(1, 83) = 3.444, p = .067, R = .200, R^2 = .040$, adjusted $R^2 = .028$</td>
</tr>
<tr>
<td>Length 4</td>
<td>$F(1, 83) = 4.113, p = .046^*, R = .217, R^2 = .047$, adjusted $R^2 = .036$</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>$F(1, 83) = 0.752, p = .388, R = .095, R^2 = .009$, adjusted $R^2 = -.003$</td>
</tr>
<tr>
<td>General complexity</td>
<td>$F(1, 83) = 1.771, p = .187, R = .145, R^2 = .021$, adjusted $R^2 = .009$</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>$F(1, 83) = 1.705, p = .195, R = .142, R^2 = .020$, adjusted $R^2 = .008$</td>
</tr>
<tr>
<td>SV</td>
<td>$F(1, 83) = 1.715, p = .194, R = .142, R^2 = .020$, adjusted $R^2 = .008$</td>
</tr>
<tr>
<td>LD (D)</td>
<td>$F(1, 83) = 0.496, p = .483, R = .077, R^2 = .006$, adjusted $R^2 = -.006$</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>EFW</td>
<td>$F(1, 83) = 1.034, p = .312, R = .111, R^2 = .012$, adjusted $R^2 = .000$</td>
</tr>
<tr>
<td>EFCs</td>
<td>$F(1, 83) = 0.151, p = .698, R = .043, R^2 = .002$, adjusted $R^2 = -.010$</td>
</tr>
<tr>
<td>CVU</td>
<td>$F(1, 83) = 1.760, p = .188, R = .144, R^2 = .021$, adjusted $R^2 = .009$</td>
</tr>
<tr>
<td>Content</td>
<td>$F(1, 83) = 2.034, p = .158, R = .155, R^2 = .024$, adjusted $R^2 = .012$</td>
</tr>
<tr>
<td>Organization</td>
<td>$F(1, 83) = 4.079, p = .047^*, R = .216, R^2 = .047$, adjusted $R^2 = .035$</td>
</tr>
</tbody>
</table>

* means statistical significance was found on the simple linear regression of WMC Two on a measure.

Table 4.11 presents the results of the regression analyses of baseline task writing performance on main task writing performance.

As is shown in Table 4.11, all the 20 aspects of baseline writing performance scores predicted the corresponding main task performance scores well (all $p = .000$). The statistically significant correlations indicated that the baseline performance scores could be included as covariates for later ANCOVA analyses of group differences in main task performance.
Table 4.11

Results of Regression Analyses of Baseline Task on Main Task Performance

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Simple linear regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>$F (1, 83) = 59.508, p = .000^*, R = .648, R^2 = .421, \text{adjusted } R^2 = .413$</td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F (1, 83) = 83.849, p = .000^*, R = .711, R^2 = .506, \text{adjusted } R^2 = .500$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F (1, 82) = 98.212, p = .000^*, R = .740, R^2 = .548, \text{adjusted } R^2 = .542$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F (1, 83) = 91.510, p = .000^*, R = .726, R^2 = .527, \text{adjusted } R^2 = .522$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F (1, 82) = 106.928, p = .000^*, R = .754, R^2 = .569, \text{adjusted } R^2 = .564$</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
</tr>
<tr>
<td>Complex</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td>$\text{ND}$ $F (1, 83) = 77.389, p = .000^*, R = .695, R^2 = .483, \text{adjusted } R^2 = .476$</td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F (1, 83) = 54.786, p = .000^*, R = .631, R^2 = .398, \text{adjusted } R^2 = .390$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F (1, 83) = 71.359, p = .000^*, R = .682, R^2 = .465, \text{adjusted } R^2 = .459$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F (1, 83) = 66.505, p = .000^*, R = .667, R^2 = .445, \text{adjusted } R^2 = .438$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F (1, 83) = 82.764, p = .000^*, R = .709, R^2 = .502, \text{adjusted } R^2 = .496$</td>
</tr>
<tr>
<td>ND</td>
<td>$F (1, 83) = 32.885, p = .000^*, R = .533, R^2 = .284, \text{adjusted } R^2 = .275$</td>
</tr>
<tr>
<td>Length 1</td>
<td>$F (1, 83) = 79.405, p = .000^*, R = .699, R^2 = .489, \text{adjusted } R^2 = .483$</td>
</tr>
<tr>
<td>Length 2</td>
<td>$F (1, 83) = 48.924, p = .000^*, R = .609, R^2 = .371, \text{adjusted } R^2 = .363$</td>
</tr>
<tr>
<td>Length 3</td>
<td>$F (1, 83) = 17.662, p = .000^*, R = .419, R^2 = .175, \text{adjusted } R^2 = .166$</td>
</tr>
<tr>
<td>Length 4</td>
<td>$F (1, 83) = 50.754, p = .000^*, R = .616, R^2 = .379, \text{adjusted } R^2 = .372$</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td></td>
</tr>
<tr>
<td>Kurtosis</td>
<td></td>
</tr>
<tr>
<td>Speed 1</td>
<td>$F (1, 83) = 197.239, p = .000^*, R = .839, R^2 = .704, \text{adjusted } R^2 = .700$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>$F (1, 83) = 83.736, p = .000^*, R = .709, R^2 = .502, \text{adjusted } R^2 = .496$</td>
</tr>
<tr>
<td>Speed 3</td>
<td>$F (1, 83) = 25.514, p = .000^*, R = .485, R^2 = .235, \text{adjusted } R^2 = .226$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>$F (1, 83) = 21.812, p = .000^*, R = .456, R^2 = .208, \text{adjusted } R^2 = .199$</td>
</tr>
<tr>
<td>CVU</td>
<td>$F (1, 83) = 30.151, p = .000^*, R = .516, R^2 = .266, \text{adjusted } R^2 = .258$</td>
</tr>
</tbody>
</table>

* means statistical significance was found on the simple linear regression of baseline task performance scores on main task performance scores.

4.3.4 Prior differences between the two groups

4.3.4.1 Normality tests on baseline writing scores

A series of normality tests was conducted. The values for skewness and kurtosis between -2 and +2 were considered acceptable in order to establish normal univariate distribution (George & Mallery, 2010). Based on this benchmark, the scores that needed a recheck of normality included: 1) Speed 1 for the NTA group (Kurtosis = 3.093), 2) Speed 2 for the NTA group (Kurtosis = 4.098), 3) Speed 3 for the NTA group (Kurtosis = 2.597), 4) Speed 4 for the NTA group (Kurtosis = 3.385), 5) Length 1 for the NTA group (Kurtosis = 2.160), 6) Length 3 for the NTA group (Kurtosis = 2.030), 7) D for
the NTA group (Kurtosis = 5.133) and 8) EFWs for the TA group (Kurtosis = 2.700). More cases of violation of normal distribution were found in the Shapiro-Wilk and Kolmogorov-Smirnov tests. Taken together the results of both the two tests, normality may not be assumed in the following dataset: 1) Speed 1 and Speed 2 for the NTA group, 2) ND for the NTA group, 3) Length 1 for the NTA group, 4) Subclausal Complexity for the NTA group, 5) EFWs for the TA group, 6) EFCs for the TA group, 7) CVU for the NTA group, 8) Content for the TA group, 9) Content for the NTA group, 10) Organization for the TA group, and 11) Organization for the NTA group. I will show how I dealt with the outliers when I present the results of the corresponding ANCOVA analyses, considering the homogeneity of variance for the ANCOVA analyses concomitantly.

4.3.4.2 Results of t-tests on baseline data

A series of Independent-Samples t-tests was conducted on all the baseline scores to compare the performance of the NTA group and the TA group to examine if there were prior differences between the two groups.

Table 4.12 presents the results of the t-tests, all \( N (TA) = 43, N (NTA) = 42 \).

As Table 4.12 shows, there was no statistically significant difference between the NTA group and the TA group in all the 20 measures prior to implementation of the experimental writing task.

Given the absence of prior between-group differences in WMC Average, WMC One, WMC Two, and the vast majority of the baseline performance data, and given the significant predictive power of the baseline data on main task data, a decision was made to use the baseline data and the WMC scores as two covariates for ANCOVA analyses.
Table 4.12

Results of t-Tests Run on Baseline Data

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>-2.281</td>
<td>83</td>
<td>.025</td>
</tr>
<tr>
<td>Speed 1</td>
<td>1.283</td>
<td>83</td>
<td>.203</td>
</tr>
<tr>
<td>Speed 2</td>
<td>1.047</td>
<td>82</td>
<td>.298</td>
</tr>
<tr>
<td>Speed 3</td>
<td>1.330</td>
<td>83</td>
<td>.187</td>
</tr>
<tr>
<td>Speed 4</td>
<td>1.092</td>
<td>82</td>
<td>.278</td>
</tr>
<tr>
<td>ND</td>
<td>0.457</td>
<td>82</td>
<td>.649</td>
</tr>
<tr>
<td>Length 1</td>
<td>-0.712</td>
<td>83</td>
<td>.478</td>
</tr>
<tr>
<td>Length 2</td>
<td>-0.910</td>
<td>82</td>
<td>.366</td>
</tr>
<tr>
<td>Length 3</td>
<td>-0.607</td>
<td>83</td>
<td>.546</td>
</tr>
<tr>
<td>Length 4</td>
<td>-.775</td>
<td>82</td>
<td>.441</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>-1.766</td>
<td>74.760**</td>
<td>.081</td>
</tr>
<tr>
<td>General complexity</td>
<td>-2.152</td>
<td>83</td>
<td>.034</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>-0.836</td>
<td>83</td>
<td>.406</td>
</tr>
<tr>
<td>SV</td>
<td>-1.516</td>
<td>83</td>
<td>.133</td>
</tr>
<tr>
<td>LD (D)</td>
<td>-0.960</td>
<td>83</td>
<td>.340</td>
</tr>
<tr>
<td>EFWs</td>
<td>-1.758</td>
<td>83</td>
<td>.082</td>
</tr>
<tr>
<td>EFCs</td>
<td>-1.633</td>
<td>83</td>
<td>.106</td>
</tr>
<tr>
<td>CVU</td>
<td>-0.281</td>
<td>77.122**</td>
<td>.779</td>
</tr>
<tr>
<td>Content</td>
<td>-1.288</td>
<td>82.010**</td>
<td>.201</td>
</tr>
<tr>
<td>Organization</td>
<td>-1.037</td>
<td>83</td>
<td>.303</td>
</tr>
</tbody>
</table>

* p < .0025 after the Bonferroni adjustment.
** means equal variances not assumed.

It is noted that the WMC scores did not significantly predict most of the main task performance measures. However, since there were no significant differences between the NTA group and the TA group in the WMC scores, using the WMC scores as a covariate would only slightly affect the results of the between-groups analyses of the main effects of TA, and using it as a covariate for analysing all dependent variables facilitated statistical operations and in particular helped examine the interactions of WMC (Average, 1, and 2) with Condition when WMC (Average, 1, and 2) was used as a covariate.

4.3.5 Differences between the two groups in their main task performance

4.3.5.1 The normality of main writing task data

A series of normality tests was then conducted on the participants’ main task scores concerning the 20 measures. The Skewness and Kurtosis scores indicated that normality may not apply to the scores for Speed 2 (Kurtosis = 2.142), Speed 3 (Kurtosis = 2.797),
and Speed 4 (Kurtosis = 3.395) for the NTA group, and for D for the NTA group (Kurtosis = 4.582). The Kolmogorov-Smirnov and the Shapiro-Wilk tests suggested additional data sets did not meet normal distribution criteria: 1) Speed 3 for the NTA group, 2) subclausal complexity scores for the TA group, 3) EFW data for the TA group, 4) Organization scores for the TA group, and 5) Organization scores for the NTA group.

I shall show how I dealt with the outliers when I present the results of the corresponding ANCOVA analyses, considering the homogeneity of variance for the ANCOVA analyses concomitantly.

In what follows, I shall report the results of my analyses of the main effects of Condition and the interactional effects of the two covariates, namely, baseline scores and WMC (WMC Average as a covariate first, then WMC One, and then WMC Two). The analyses followed the procedures outlined from Steps 7) to 14) in Section 4.1, which has provided an overview of how the ANCOVA analyses would be operated.

4.3.5.2 The ANCOVAs with WMC Average as one of the covariates

In this section, I shall present the results of my analyses of the main effects of Condition and the interactional effects of baseline scores and WMC Average.

4.3.5.2.1 Initial ANCOVA analyses—WMC Average

A series of initial one-way ANCOVAs was conducted, controlling for baseline writing performance and WMC Average, using a custom model for the operation which included Condition, the covariates, and the two covariate × Condition interaction terms. The results indicated that there was no covariate × Condition interaction on 15 of the 20 dependent variables.

These 15 dependent variables included four fluency measures, namely, Time 2, Condition × WMC, $F (1, 78) = 0.410, p = .524, \eta^2 = .005$, Condition × Time 1, $F (1, 78) = 0.004, p = .951, \eta^2 = .000$, Speed One 2, Condition × WMC, $F (1, 78) = 0.345, p$
=.558, $\eta^2 = .004$, Condition $\times$ Speed One 1, $F (1, 78) = 2.820, p = .097, \eta^2 = .035$, Speed Three 2, Condition $\times$ WMC, $F (1, 78) = 0.407, p = .525, \eta^2 = .005$, Condition $\times$ Speed Three 1, $F (1, 78) = 3.020, p = .086, \eta^2 = .037$, and ND 2, Condition $\times$ WMC, $F (1, 78) = 0.000, p = .995, \eta^2 = .000$, Condition $\times$ ND 1, $F (1, 78) = 1.272, p = .263, \eta^2 = .016$.

Eight of the 15 dependable variables were complexity measures, namely, Length One 2, Condition $\times$ WMC, $F (1, 79) = 0.662, p = .418, \eta^2 = .008$, Condition $\times$ Length One 1, $F (1, 79) = 1.434, p = .235, \eta^2 = .018$, Length Two 2, Condition $\times$ WMC, $F (1, 78) = 0.712, p = .401, \eta^2 = .009$, Condition $\times$ Length Two 1, $F (1, 78) = 2.863, p = .095, \eta^2 = .035$, Length Three 2, Condition $\times$ WMC, $F (1, 79) = 0.447, p = .506, \eta^2 = .006$, Condition $\times$ Length Three 1, $F (1, 79) = 0.000, p = .999, \eta^2 = .000$, Length Four 2, Condition $\times$ WMC, $F (1, 78) = 0.365, p = .547, \eta^2 = .005$, Condition $\times$ Length Four 1, $F (1, 78) = 0.432, p = .513, \eta^2 = .006$, Syntactic Complexity 2, Condition $\times$ WMC, $F (1, 79) = 0.381, p = .539, \eta^2 = .005$, Condition $\times$ Syntactic Complexity 1, $F (1, 79) = 2.416, p = .124, \eta^2 = .030$, Subclausal Complexity 2, Condition $\times$ WMC, $F (1, 79) = 1.944, p = .167, \eta^2 = .024$, Condition $\times$ Subclausal Complexity 1, $F (1, 79) = 2.636, p = .108, \eta^2 = .032$, SV 2, Condition $\times$ WMC, $F (1, 79) = 0.545, p = .463, \eta^2 = .007$, Condition $\times$ SV 1, $F (1, 79) = 1.913, p = .171, \eta^2 = .024$, and D 2, Condition $\times$ WMC, $F (1, 79) = 1.217, p = .273, \eta^2 = .015$, Condition $\times$ D 1, $F (1, 79) = 1.754, p = .189, \eta^2 = .022$.

Also among the 15 dependent variables were two accuracy measures, namely, EFWs 2, Condition $\times$ WMC, $F (1, 79) = 1.353, p = .248, \eta^2 = .017$, Condition $\times$ EFWs 1, $F (1, 79) = 0.867, p = .355, \eta^2 = .011$, and EFCs 2, Condition $\times$ WMC, $F (1, 79) = 1.667, p = .200, \eta^2 = .021$, Condition $\times$ EFWs 1, $F (1, 79) = 0.412, p = .523, \eta^2 = .005$, and the measure of content, namely, Content 2, Condition $\times$ WMC, $F (1, 79) = 0.003, p = .960, \eta^2 = .000$, Condition $\times$ Content 1, $F (1, 79) = 0.039, p = .844, \eta^2 = .000$. 
There was significant covariate × Condition interaction for five dependent variables. For Speed Two, significant covariate × Condition interaction was found between Speed Two 1 and Condition, $F(1, 77) = 4.121, p = .046, \eta^2 = .051$, but was not found between WMC and Condition, $F(1, 77) = 0.430, p = .514, \eta^2 = .006$. For Speed Four, significant covariate × Condition interaction was found between Speed Four 1 and Condition, $F(1, 77) = 4.330, p = .041, \eta^2 = .053$, but was not found between WMC and Condition, $F(1, 77) = 0.539, p = .465, \eta^2 = .007$. For General Complexity 2, significant covariate × Condition interaction was found between General Complexity 1 and Condition, $F(1, 79) = 6.942, p = .010, \eta^2 = .081$, but not found between WMC and Condition, $F(1, 79) = 1.016, p = .317, \eta^2 = .013$. For CVU 2, significant covariate × Condition interaction was found between CVU 1 and Condition, $F(1, 79) = 7.547, p = .007, \eta^2 = .087$, but not found between Condition and WMC, $F(1, 79) = 0.315, p = .576, \eta^2 = .004$. Differently, for Organization 2, significant covariate × Condition interaction was found between WMC and Condition, $F(1, 79) = 3.945, p = .050, \eta^2 = .048$, but was not found between Condition and Organization 1, $F(1, 79) = 0.758, p = .387, \eta^2 = .010$.

Among the 20 variables, the assumption of the homogeneity of variance was met for 13 dependent variables, namely, Time 2, $F(1, 82) = 3.924, p = .051$, ND 2, $F(1, 82) = 3.252, p = .075$, Length One 2, $F(1, 83) = 2.820, p = .097$, Length Two 2, $F(1, 82) = 3.428, p = .068$, Syntactic Complexity 2, $F(1, 83) = 1.031, p = .313$, General Complexity, $F(1, 83) = 0.095, p = .759$, Subclausal Complexity, $F(1, 83) = 1.205, p = .276$, SV, $F(1, 83) = 0.005, p = .946$, D 2, $F(1, 83) = 0.002, p = .961$, EFWs 2, $F(1, 83) = 0.121, p = .729$, EFCs 2, $F(1, 83) = 0.133, p = .716$, CVU 2, $F(1, 83) = 0.029, p = .866$, and Content 2, $F(1, 83) = 0.726, p = .397$.

The assumption of the homogeneity of variance was not met for seven of the dependent variables including Speed One 2, $F(1, 82) = 7.797, p = .007$, Speed Two 2,
4.3.5.2.2 Further analyses of the 15 measures where no interaction was found

In this section I shall present the results of my further analyses of the 11 dependent variables where neither type of covariate × Condition interaction was found in the initial ANCOVA analyses and where the assumption of homogeneity of variance was met, before I present the results of my further analyses of the four dependent variables where neither type of covariate × Condition interaction was found but where the assumption of homogeneity of variance was not met.

4.3.5.2.2.1 The 11 measures where the assumption of homogeneity of variance was met

For the 11 dependent variables where neither type of covariate × Condition interaction was found in the initial ANCOVA tests and where the assumption of homogeneity of variance was met, a series of ANCOVAs was re-run, keeping one covariate × Condition interaction term while excluding the other Condition, alternatively. The ANCOVAs run with baseline performance × Condition interaction term showed no baseline performance × Condition interaction on all the 11 dependent variables, namely, Time 2, $F(1, 79) = 0.034, p = .854, \eta^2 = .000$, ND 2, $F(1, 79) = 1.294, p = .259, \eta^2 = .016$, Length One 2, $F(1, 80) = 1.479, p = .227, \eta^2 = .018$, Length Two 2, $F(1, 79) = 2.987, p = .088, \eta^2 = .036$, Syntactic Complexity 2, $F(1, 80) = 2.441, p = .122, \eta^2 = .030$, Subclausal Complexity 2, $F(1, 80) = 2.418, p = .124, \eta^2 = .029$, SV 2, $F(1, 80) = 1.976, p = .164, \eta^2 = .024$, D 2, $F(1, 80) = 1.598, p = .210, \eta^2 = .020$, EFWs 2, $F(1, 80) = 0.674, p = .414, \eta^2 = .008$, EFCs 2, $F(1, 80) = 0.384, p = .537, \eta^2 = .005$, and Content 2, $F(1, 80) = 0.041, p = .841, \eta^2 = .001$. In all these tests, the assumption of the homogeneity of variance was met.
The ANCOVAs run with the WMC × Condition interaction term showed no WMC × Condition interaction on all the 11 variables including Time 2, $F(1, 79) = 0.446$, $p = .506$, $\eta^2 = .006$, ND 2, $F(1, 79) = 0.005$, $p = .942$, $\eta^2 = .000$, Length One 2, $F(1, 80) = 0.697$, $p = .406$, $\eta^2 = .009$, Length Two 2, $F(1, 79) = 0.807$, $p = .372$, $\eta^2 = .010$, Syntactic Complexity 2, $F(1, 80) = 0.380$, $p = .539$, $\eta^2 = .005$, Subclausal Complexity 2, $F(1, 80) = 1.719$, $p = .194$, $\eta^2 = .021$, SV 2, $F(1, 80) = 0.590$, $p = .445$, $\eta^2 = .007$, $D$ 2, $F(1, 80) = 1.055$, $p = .307$, $\eta^2 = .013$, EFWs 2, $F(1, 80) = 1.165$, $p = .284$, $\eta^2 = .014$, EFCs 2, $F(1, 80) = 1.654$, $p = .202$, $\eta^2 = .020$, and Content 2, $F(1, 80) = 0.004$, $p = .952$, $\eta^2 = .000$. In all these tests, the assumption of the homogeneity of variance was met except for Time, $F(1, 82) = 3.979$, $p = .049$.

A series of ANCOVAs was then run, without the interaction terms, using the full factorial mode, and pairwise comparisons were also conducted after each ANCOVA operation, following the LSD procedure. The assumption of homogeneity of variance was re-checked. The assumption was still met for all these 11 dependent variables, namely, Time 2, $F(1, 82) = 3.772$, $p = .056$, ND 2, $F(1, 82) = 3.643$, $p = .060$, Length One 2, $F(1, 83) = 2.273$, $p = .135$, Length Two 2, $F(1, 82) = 1.908$, $p = .171$, Syntactic Complexity 2, $F(1, 83) = 0.595$, $p = .443$, Subclausal Complexity 2, $F(1, 83) = 1.008$, $p = .318$, SV 2, $F(1, 83) = 0.007$, $p = .932$, $D$ 2, $F(1, 83) = 0.003$, $p = .958$, EFWs 2, $F(1, 83) = 0.150$, $p = .700$, EFCs 2, $F(1, 83) = 0.150$, $p = .699$, and Content 2, $F(1, 83) = 0.707$, $p = .403$.

Table 4.13 lists the descriptive statistics of the main task data for the 11 measures, the main effects of Condition, the results of pairwise comparisons following the LSD procedure, and the effect sizes.
Table 4.13

**Descriptive Statistics and Results of ANCOVAs for 11 Variables**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>F</td>
<td>df</td>
</tr>
<tr>
<td>Time</td>
<td>37.558 (9.277, 38.754)</td>
<td>40.902 (8.040, 39.649)</td>
<td>0.364</td>
<td>1, 80</td>
</tr>
<tr>
<td>ND</td>
<td>.906 (.055, .904)</td>
<td>.921 (.041, .922)</td>
<td>5.916</td>
<td>1, 80</td>
</tr>
<tr>
<td>Length 1</td>
<td>341.090 (40.769, 343.600)</td>
<td>356.000 (59.171, 353.500)</td>
<td>1.306</td>
<td>1, 81</td>
</tr>
<tr>
<td>Length 2</td>
<td>362.630 (44.204, 366.400)</td>
<td>374.800 (66.549, 370.900)</td>
<td>0.245</td>
<td>1, 80</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>1.987 (.288, 2.016)</td>
<td>2.038 (.297, 2.008)</td>
<td>0.019</td>
<td>1, 81</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>7.791 (1.178, 7.847)</td>
<td>8.186 (.924, 8.129)</td>
<td>2.327</td>
<td>1, 81</td>
</tr>
<tr>
<td>Syntactic variety</td>
<td>15.511 (3.104, 15.715)</td>
<td>16.214 (3.273, 16.006)</td>
<td>0.203</td>
<td>1, 81</td>
</tr>
<tr>
<td>LD (D)</td>
<td>61.554 (9.459, 62.228)</td>
<td>66.913 (12.112, 66.223)</td>
<td>4.492</td>
<td>1, 81</td>
</tr>
<tr>
<td>EFWs</td>
<td>88.404 (5.344, 89.161)</td>
<td>89.787 (4.008, 89.013)</td>
<td>0.064</td>
<td>1, 81</td>
</tr>
<tr>
<td>EFCs</td>
<td>.545 (.124, .561)</td>
<td>.554 (.130, .538)</td>
<td>1.285</td>
<td>1, 81</td>
</tr>
<tr>
<td>Content</td>
<td>20.977 (3.158, 21.181)</td>
<td>22.095 (3.392, 21.886)</td>
<td>1.176</td>
<td>1, 81</td>
</tr>
</tbody>
</table>

Note. For Time, ND, and Length 2, TA: N = 43, NTA: N = 41; for all other measures, TA: N = 43, NTA: N = 42. *p < .05.

4.3.5.2.2.2 The four measures where the assumption of homogeneity of variance was not met

4.3.5.2.2.2.1 Speed One

For Speed One, one outlier was removed from the baseline dataset, from the NTA group (Participant 97, Speed 1 = 18.150). This removal yielded Skewness = .884, Kurtosis = .915, and normality was also met via both the Kolmogorov-Smirnov (\(df = 41, p = .176\)) and Shapiro-Wilk tests (\(df = 41, p = .067\)). One outlier was removed from the main task dataset, from the TA group (Participant 36, Speed 1 = 17.348), and the corresponding main task Speed 1 value for Participant 97 in the NTA group was also removed as an outlier (Speed 1 = 17.000). The normality of the main task data was
achieved, Skewness = .847, Kurtosis = 1.015 for the NTA group, and Skewness = -.119, Kurtosis = -.305 for the TA group. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both indicated no violation of normality, for the TA group, $df = 42, p = .200$ and $p = .766$, and for the NTA group, $df = 40, p = .200$ and $p = .086$.

An ANCOVA was then run with both interaction terms, on the normally distributed dataset. The assumption of homogeneity of variance was not met, $F (1, 80) = 6.746, p = .011$. However, the ratio of the greater group variance ($SD$ squared) to that of the smaller group variance ($SD$ squared) (1.134) was below the four times benchmark, indicating that further analysis based on such variances was safe (Roberts & Russo, 2014). The ANCOVA test indicated that there was significant Speed One $\times$ Condition interaction, $F (1, 76) = 4.198, p = .044, \eta^2 = .052$, but there was no significant WMC $\times$ Condition interaction, $F (1, 76) = 0.119, p = .731, \eta^2 = .002$, for the modified data.

Then, an ANCOVA was re-run, with only the Speed One $\times$ Condition interaction term while excluding the WMC $\times$ Condition interaction term. The assumption of homogeneity of variance was still not met, $F (1, 80) = 6.212, p = .015$. This ANCOVA test confirmed the significant interaction between baseline Speed One and Condition, $F (1, 77) = 4.164, p = .045, \eta^2 = .051$.

Table 4.14 lists the results of this analysis, which indicated no statistically significant difference between the two groups in Speed 1.

This finding was confirmed by a $t$-test run with the new main task data, yielding $t = .932, df = 80, p = .354$. 

139
Table 4.14

**Results of ANCOVA for Speed One after Normalization**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>9.387 (2.106, 9.218)</td>
<td>8.939 (2.243, 9.246)</td>
<td>0.006</td>
</tr>
<tr>
<td>NTA</td>
<td>1, 77</td>
<td>.938</td>
<td>vs. TA</td>
</tr>
</tbody>
</table>

Note. TA: N = 42, NTA: N = 40.

*p < .05.

A scatterplot of main task Speed One with baseline task Speed One by Condition was drawn to examine how the effects of TA differed for participants with different baseline Speed One.

Figure 4.1 shows the scatterplot.

![Scatterplot of main task Speed One with baseline task Speed One by Condition](image)

Figure 4.1 Baseline Task Speed One × Condition Interaction
Figure 4.1 indicated that although overall TA did not affect the number of words produced per minute, it had a differential effect on slow and fast L2 writers. Fast student writers, under the TA condition, tended to write more slowly, while slow student writers, when thinking aloud, appeared to write faster.

I then conducted a univariate test with Condition and Grouping (which contained the high, mid, and low subgroups grouped based on baseline Speed One) as two independent variables, the main task Speed One as a dependent variable, the baseline task Speed One scores and WMC Average as two covariates, keeping the Condition × Grouping interaction term only, to examine if there was significant Condition × Grouping interaction. Follow-up pairwise comparisons were also carried out, following the Bonferroni procedure, to examine if there was any significant difference between the subgroups within the same condition and across the NTA/TA conditions. The assumption of homogeneity of variance was met, $F(5, 76) = 1.469, p = .210$. No significant interaction between Condition and Grouping was found, $F(2, 74) = 1.458, p = .239, \eta^2 = .038$.

Table 4.15 lists the descriptive statistics of the three Speed One subgroups in relation to Condition. The table provides an overview of the between-groups differences within the same condition and across conditions.

### Table 4.15

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subgroups</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>8.824a</td>
<td>.860</td>
<td>7.111</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.382a</td>
<td>.449</td>
<td>8.487</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>9.370a</td>
<td>.583</td>
<td>8.209</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.033a</td>
<td>.688</td>
<td>6.663</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>9.391a</td>
<td>.419</td>
<td>8.556</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>10.241a</td>
<td>.731</td>
<td>8.784</td>
</tr>
</tbody>
</table>

*Note. Condition 1 = NTA, 2 = TA; Subgroup 1 = High, 2 = Mid, 3 = Low.

*Covariates appearing in the model are evaluated at the following values: Speed One 1 = 8.998, WMC = .780.*
Within the same (either NTA or TA) condition, pairwise comparisons of groups showed that there was no significant difference between different Speed One groups in main task Speed One, controlling for prior differences in Speed One and WMC Average (for the NTA condition, all $p = 1.000$, and for the TA condition, $p = .321$ for High versus Mid; $p = .244$ for High versus Low, and $p = .916$ for Mid versus Low). Across the two different conditions, pairwise comparisons of the same levels of groups also indicated no significant difference between the high Speed One groups, $p = .278$ (the TA condition slightly slower), the mid Speed One groups, $p = .989$ and between the low Speed One groups, $p = .191$ (the TA condition slightly faster).

Figure 4.2 shows the results of the between- and within-condition comparisons.

![Figure 4.2 Subgroup Between- and Within-Condition Comparisons on Speed One](image)
A further analysis comparing the two conditions at seven different levels of baseline Speed One scores (5.460, 6.284, 7.293, 8.855, 10.665, 12.225, and 13.325, which corresponded to the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for baseline Speed One scores), including both baseline Speed One and WMC Average as covariates, showed that no significant difference across the two conditions existed at any percentile point, \( p = .094, p = .128, p = .239, p = .957, p = .206, p = .091, \) and \( p = .070, \) respectively.

### 4.3.5.2.2.2 Speed Three

For Speed 3, one outlier was removed from the baseline dataset, from the NTA group (Participant 97, Speed 3 = 24.150). This removal yielded Skewness = .727 and Kurtosis = .651, for the NTA group, and normality examined via both the Kolmogorov-Smirnov and Shapiro-Wilk tests was also met (\( df = 41, p = .200 \) and \( df = 41, p = .187, \) respectively). The corresponding main task Speed 3 value for Participant 97 in the NTA group was also removed as an outlier (Speed 3 = 24.115). Two more outliers were removed from the main task dataset, one from the NTA group (Participant 11, Speed 3 = 21.080), and the other from the TA group (Participant 36, Speed 3 = 21.000). The normality of the main task data was achieved, Skewness = .203 and Kurtosis = -.711 for the NTA group, and Skewness = -.156 and Kurtosis = -.353 for the TA group. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both indicated no violation of normality, for the TA group, \( df = 42, p = .200 \) and \( p = .716, \) and for the NTA group, \( df = 39, p = .200 \) and \( p = .120. \)

An ANCOVA was then run with both interaction terms, on the normally distributed dataset. The assumption of homogeneity of variance was not met, \( F (1, 79) = 8.513, p = .005, \) but the ratio of the greater group variance (SD squared) to that of the smaller group variance (SD squared) was below the four times benchmark \( (2.968^2/2.526^2 = 1.381), \) indicating that the ANCOVA analysis based on such variances was safe (Roberts...
& Russo, 2014). The ANCOVA test indicated neither significant baseline Speed Three \( \times \) Condition interaction, \( F(1, 75) = 0.913, p = .343, \eta^2 = .012 \), nor significant WMC \( \times \) Condition interaction, \( F(1, 75) = 0.254, p = .616, \eta^2 = .003 \), for the modified data. Then, another ANCOVA was run, with only the baseline Speed Three \( \times \) Condition interaction term, to examine if any significant interaction existed between baseline Speed Three and Condition. The assumption of homogeneity of variance was still not met, \( F(1, 79) = 7.964, p = .006 \). Again no significant interaction between baseline Speed Three and Condition was found, \( F(1, 76) = 0.814, p = .370, \eta^2 = .011 \). Still another ANCOVA was run, with only the WMC \( \times \) Condition interaction term. The assumption of homogeneity of variance was still not met, \( F(1, 79) = 6.511, p = .013 \). No statistically significant WMC \( \times \) Condition interaction was found either, \( F(1, 76) = 0.148, p = .702, \eta^2 = .002 \).

An ANCOVA was then run, with baseline Speed Three and WMC as two covariates, using the full factorial mode, with no interaction terms. The assumption of the homogeneity of variance was still not met, \( F(1, 79) = 6.474, p = .013 \). However, the ratio of the group variances, when calculated (the standard deviation for the TA group \( SD = 2.968 \) squared divided by the standard deviation for the NTA group \( SD = 2.526 \) squared), was found to be 1.381, well below the four times benchmark, indicating that further analysis based on such variances was safe (Roberts & Russo, 2014).

Table 4.16 lists the results of the new ANCOVA analysis with the normally distributed dataset.

A \( t \)-test was also run with the new main task data set to compare the means of the two groups on Speed 3, yielding \( t = 1.144, df = 79, p = .256 \), indicating no statistically significant difference between the two groups on the particular measure.
Table 4.16

**Descriptive Statistics and Results of ANCOVA for Speed Three after Normalization**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>F</td>
</tr>
<tr>
<td>Speed 3</td>
<td>12.543 (2.968, 12.194)</td>
<td>11.839 (2.526, 12.215)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Note. TA: N = 42, NTA: N = 39.*

*p < .05.

4.3.5.2.2.2.3 Length Three

For Length 3, both the baseline and the main task dataset for the two groups already demonstrated fair normality. For safety’s sake, one outlier was removed from the baseline dataset, from the NTA group (Participant 42, Length 3 = 718). This removal improved normality for the data, yielding Skewness = -.154 and Kurtosis = -.730, and both the Kolmogorov-Smirnov ($df = 41, p = .200$) and Shapiro-Wilk tests ($df = 41, p = .567$) showed good normality. The corresponding value for the same participant in the main task dataset was also excluded from analysis (Participant 42, Length 3 = 642). The normality of the main task data was checked, Skewness = .393 and Kurtosis = -.085 for the NTA group, and the Kolmogorov-Smirnov and the Shapiro-Wilk tests both indicated no violation of normality, for the NTA group, $df = 41, p = .200$ and $p = .836$.

An ANCOVA was then run with both interaction terms, on the normally distributed dataset. The assumption of homogeneity of variance was not met, $F (1, 82) = 6.390, p = .013$, but the ratio of the greater group variance ($SD$ squared) to that of the smaller group variance ($SD$ squared), was below the four times benchmark ($77.708^2 \div 59.136^2 = 1.727$), indicating that the ANCOVA analysis based on such variances was safe (Roberts & Russo, 2014). The ANCOVA test indicated neither significant baseline Length Three
× Condition interaction, $F (1, 78) = 0.036, p = .851, \eta^2 = .000$, nor significant WMC × Condition interaction, $F (1, 78) = 0.459, p = .500, \eta^2 = .006$, for the modified data. Then, another ANCOVA was run, with only the baseline Length Three × Condition interaction term, to examine if any significant interaction existed between baseline Length Three and Condition. The assumption of homogeneity of variance was still not met, $F (1, 82) = 6.437, p = .013$. Again no significant interaction between baseline Speed Three and Condition was found, $F (1, 79) = 0.067, p = .796, \eta^2 = .001$. Still another ANCOVA was run, with only the WMC × Condition interaction term. The assumption of homogeneity of variance was still not met, $F (1, 82) = 6.284, p = .014$. No statistically significant WMC × Condition interaction was found either, $F (1, 79) = 0.496, p = .483, \eta^2 = .006$.

A final ANCOVA was then run, with baseline Length Three and WMC as two covariates, using the full factorial mode, with no interaction terms. The assumption of the homogeneity of variance was still not met, $F (1, 82) = 6.345, p = .014$. However, the ratio of the group variances, when calculated (the standard deviation for the NTA group ($SD = 77.708$) squared divided by the standard deviation for the TA group ($SD = 59.136$) squared), was 1.727, below the four times benchmark, indicating that further analysis based on such variances was safe (Roberts & Russo, 2014).

Table 4.17 lists the results of this new analysis.

### Table 4.17

**Descriptive Statistics and Results of ANCOVA for Length Three after Normalization**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>$M$ ($SD$, adjusted $M$)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons $p$ (effect size)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>$F$</td>
</tr>
<tr>
<td>Length 3</td>
<td>453.770</td>
<td>478.120</td>
<td>3.700</td>
</tr>
</tbody>
</table>

Note. TA: $N = 43$, NTA: $N = 41$.

*p < .05.
A t-test was also run with the new main task data set to compare the means of the two groups in Length 3, yielding $t = -1.843$, $df = 83$, $p = .069$, indicating no statistically significant difference between the two groups on the particular measure.

4.3.5.2.2.2.4 Length Four

For Length 4, both the baseline and the main task dataset for the two groups already demonstrated adequate normality. Two ANCOVAs were then run, first with only the baseline Length Four × Condition interaction term, and then with only the WMC × Condition interaction term. The ANCOVA run with the baseline Length Four × Condition interaction term showed no baseline Length Four × Condition interaction, $F(1, 79) = 0.545$, $p = .463$, $\eta^2 = .007$. Neither did the ANCOVA run with the WMC × Condition interaction term show any WMC × Condition interaction, $F(1, 79) = 0.477$, $p = .492$, $\eta^2 = .006$. In both ANCOVA operations, the assumption of homogeneity of variance was still not met, $F(1, 82) = 7.156$, $p = .009$ and $F(1, 82) = 7.283$, $p = .008$, respectively.

An ANCOVA was finally run, without the interaction terms, using the full factorial mode, and pairwise comparisons were also conducted afterwards, following the LSD procedure. The assumption of homogeneity of variance was still not met, $F(1, 82) = 6.698$, $p = .011$. Although the assumption was not met in this operation and the previous two operations, the ratio of the group variances, when calculated (the standard deviation for the NTA group ($SD = 90.261$) squared divided by the standard deviation for the TA group ($SD = 62.306$) squared), was 2.099, below the four times benchmark, indicating that the analyses based on such variances was safe (Roberts & Russo, 2014). The results of this ANCOVA operation indicated no statistically significant difference between the two groups in Length Four, $F(1, 80) = 1.947$, $p = .167$, $\eta^2 = .024$. 

147
Table 4.18 shows the adjusted means in juxtaposition with the original means, the main effects of Condition, the results of all pairwise comparisons, and the effect sizes, for this ANCOVA operation.

Table 4.18

Descriptive Statistics and Results of ANCOVA for Length Four

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>F</td>
</tr>
<tr>
<td>Length 4</td>
<td>479.140 (62.306, 483.700)</td>
<td>505.320 (90.261, 500.600)</td>
<td>1.947</td>
</tr>
</tbody>
</table>

Note. TA: N = 43, NTA: N = 41.
*p < .05.

A t-test was also run, to compare the means of the two groups in Length 4, yielding

\[ t = -1.605, \, df = 73.138, \, p = .113 \]

indicating no statistically significant difference between the two groups in the particular measure.

4.3.5.2.3 The five measures where interaction was found

In the initial ANCOVA tests, baseline task performance and Condition interactions were found on four dependent variables, that is, Speed Two (Speed Two 1 × Condition, \( F(1, 77) = 4.121, \, p = .046, \, η² = .051 \)), Speed Four (Speed Four 1 × Condition, \( F(1, 77) = 4.330, \, p = .041, \, η² = .053 \)), General Complexity (General Complexity 1 × Condition, \( F(1, 79) = 6.942, \, p = .010, \, η² = .081 \)), and CVU (CVU 1 × Condition, \( F(1, 79) = 7.547, \, p = .007, \, η² = .087 \)). The assumption of the homogeneity of variance was not met for Speed Two (\( F(1, 81) = 7.876, \, p = .006 \)), and Speed Four (\( F(1, 81) = 7.223, \, p = .009 \)).

Also, in the initial ANCOVA test, a WMC and Condition interaction was found on Organization (\( F(1, 79) = 3.945, \, p = .050, \, η² = .048 \)).

In what follows, I shall examine these dependent variables one by one.
4.3.5.2.3.1 Speed Two

The datasets for Speed Two were first examined to ensure that normal distribution was met for further ANCOVA analyses. One outlier from the baseline dataset was removed (i.e., Participant 97, from the NTA group, Speed Two = 20.250). This removal improved the Skewness and Kurtosis of the group’s baseline speed two data, yielding Skewness = .832, and Kurtosis = .988. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both also indicated no violation of normality, $df = 40, p = .076$ and $p = .092$. The corresponding Speed Two value for Participant 97 in the main task dataset was also removed (Speed Two = 18.731). Another value was also removed from the main task dataset (i.e., Participant 36, from the TA group, Speed Two = 18.044), so was the corresponding value in the baseline task dataset excluded (i.e., Participant 36, from the TA group, Speed Two = 14.000). The Skewness and Kurtosis of the main task Speed Two data were improved, Skewness = -.211 and Kurtosis = -.484, for the TA group, and Skewness = .843, and Kurtosis = .917, for the NTA group. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both also indicated no violation of normality, $df = 42, p = .200$ and $p = .795$, for the TA group, and $df = 40, p = .200$ and $p = .109$ for the NTA group.

An ANCOVA was then run with both interaction terms, on the normally distributed dataset. The assumption of homogeneity of variance was not met, $F (1, 79) = 6.766, p = .011$, but the ratio of the greater group variance ($SD$ squared) to that of the smaller group variance ($SD$ squared), was below the four times benchmark $(2.435^2/2.227^2 = 1.195)$, indicating that the ANCOVA analysis based on such variances was safe (Roberts & Russo, 2014). The ANCOVA test indicated significant baseline Speed Two $\times$ Condition interaction, $F (1, 75) = 5.699, p = .019, \eta^2 = .071$, but no significant WMC $\times$ Condition interaction, $F (1, 75) = 0.155, p = .695, \eta^2 = .002$, for the modified data. An
ANCOVA was then run, with the WMC × Condition interaction term kept only, using both WMC and baseline Speed Two as two covariates. The assumption of homogeneity of variance was met, $F(1, 79) = 3.570, p = .063$. There was no interaction found, $F(1, 76) = 0.042, p = .839, \eta^2 = .001$.

Given that no significant WMC × Condition interaction was found, a final ANCOVA was run with the new dataset, using the baseline Speed Two data and WMC as the covariates, keeping the Speed Two 1 × Condition interaction term only and excluding the WMC × Condition interaction term, given its nonsignificance. The assumption of homogeneity of variance was still not met, $F(1, 79) = 6.165, p = .015$. The ANCOVA test showed a statistically significant interaction between baseline Speed Two data and Condition, $F(1, 76) = 5.651, p = .020, \eta^2 = .069$, but no statistically significant main effect of Condition, $F(1, 76) = 0.035, p = .853, \eta^2 = .000$.

Table 4.19 shows the adjusted means in relation to the original means, the main effects of Condition, the results of all pairwise comparisons, and effect sizes, for this final operation of ANCOVA.

Table 4.19

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>$M (SD,\text{ adjusted } M)$</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>$F$</td>
</tr>
<tr>
<td>Speed 2</td>
<td>9.981 (2.227, 9.819)</td>
<td>9.436 (2.435, 9.749)</td>
<td>0.035</td>
</tr>
</tbody>
</table>


* $p < .05$.

As Table 4.19 shows, there was no statistically significant difference between the TA group and the NTA group in Speed Two. This finding was confirmed by a $t$-test run with the new main task data set, which yielded $t = 1.105, df = 80, p = .273$. 

150
A scatterplot of main task Speed Two by baseline task Speed Two with separate regression lines for the TA and the NTA group was drawn to examine how the effects of Condition differed for participants with different baseline Speed Two performance. Figure 4.3 shows the scatterplot.

![Scatterplot of main task Speed Two by baseline task Speed Two with separate regression lines for the TA and the NTA group.](image)

**Figure 4.3 Baseline Task Speed Two × Condition Interaction**

As the scatterplot shows, TA appeared to have a different effect on Speed Two for participants who stayed at the high and low ends of the measure. Participants who had written unproductively (as measured by Speed Two) wrote more productively under the TA condition, while those who had written productively (as measured by Speed Two) wrote less productively under the TA condition.

I then conducted a univariate test with Condition and Grouping (which contained the high, mid, and low subgroups grouped based on baseline Speed Two) as two independent variables, the main task Speed Two data as a dependent variable, the
baseline task Speed Two data and WMC as two covariates, keeping the Condition × Grouping interaction term only, to examine if there was significant Condition × Grouping interaction. Follow-up pairwise comparisons were also carried out following the Bonferroni procedure to examine if there was any significant difference between the subgroups within the same condition and across the NTA/TA conditions. The assumption of homogeneity of variance was met, \( F(5, 75) = 1.618, p = .166 \). No significant interaction between Condition and Grouping was found, \( F(2, 73) = 1.979, p = .146, \eta^2 = .051 \).

Table 4.20 lists the descriptive statistics of the three Speed Two groups in relation to Condition.

Table 4.20

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subgroups</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>9.603a</td>
<td>.791</td>
<td>8.027 - 11.179</td>
</tr>
<tr>
<td>2</td>
<td>9.939a</td>
<td>.521</td>
<td>8.900 - 10.978</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.681a</td>
<td>.555</td>
<td>8.574 - 10.788</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.546a</td>
<td>.704</td>
<td>7.143 - 9.950</td>
</tr>
<tr>
<td>2</td>
<td>10.212a</td>
<td>.454</td>
<td>9.306 - 11.118</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10.488a</td>
<td>.656</td>
<td>9.180 - 11.795</td>
<td></td>
</tr>
</tbody>
</table>

Note. Condition 1 = NTA, 2 = TA; Subgroup 1 = High, 2 = Mid, 3 = Low.

\( ^a \) Covariates appearing in the model are evaluated at the following values: WMC = .781, Speed Two 1 = 9.583.

Within the same (either NTA or TA) condition, pairwise comparisons of groups showed that there was no significant difference between different speed groups in main task Speed Two, controlling for prior differences in Speed Two and WMC (for the NTA condition, all \( p = 1.000 \), and for the TA condition, \( p = .141 \) for High versus Mid, \( p = .326 \) for High versus Low, and \( p = 1.000 \) for Mid versus Low). Across the two different conditions, pairwise comparisons of the same levels of groups also indicated no statistically significant difference between the high Speed Two groups, \( p = .146 \) (the TA
condition slightly slower), between the mid Speed Two groups, $p = .692$, and between the low Speed Two groups, $p = .193$ (the TA condition slightly faster).

Figure 4.4 shows the results of the between- and within-Condition comparisons.

![Graph showing the results of the between- and within-Condition comparisons on Speed Two](image)

Figure 4.4 *Subgroup Between- and Within-Condition Comparisons on Speed Two*

A further univariate analysis comparing the two conditions at seven different levels of baseline Speed Two scores (5.867, 6.799, 7.870, 9.545, 11.005, 12.912, and 13.980, which corresponded to the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for baseline Speed Two scores), including both baseline Speed Two and WMC scores as the covariates, showed that significant difference across the two conditions only existed at 5.867, $p = .037$, the difference approached significance at 6.799, $p = .055$, it did not reach significance at most of the percentiles, at 7.870, $p = .124$, at 9.545, $p = .824$, at 11.005, $p = .296$, and at 12.912, $p = .074$, and it approached statistical significance at the highest end of the percentiles, at 13.980, $p = .051$. 

153
4.3.5.2.3.2 *Speed Four*

For Speed Four, one outlier from the baseline dataset was removed (i.e., Participant 97, from the NTA group, Speed Four = 26). This removal improved the Skewness and Kurtosis of the group’s baseline Speed Four data, Skewness = .714, and Kurtosis = .838. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both also indicated no violation of normality, $df = 40, p = .200$ and $p = .291$. The corresponding Speed Four value for Participant 97 in the main task dataset was also removed (Speed Four = 26.385). The Skewness and Kurtosis of the main task Speed Four data were improved, Skewness = .725, and Kurtosis = .801, for the NTA group. The Kolmogorov-Smirnov and the Shapiro-Wilk tests both also indicated no violation of normality, $df = 40, p = .200$ and $p = .105$, for the NTA group.

An ANCOVA was then run with both interaction terms, on the normally distributed dataset. The assumption of homogeneity of variance was not met, $F(1, 80) = 7.269, p = .009$, but the ratio of the greater group variance ($SD$ squared) to that of the smaller group variance ($SD$ squared), was below the four times benchmark ($3.317^2/3.103^2 = 1.143$), indicating that the ANCOVA analysis based on such variances was safe. The ANCOVA test indicated neither significant baseline Speed Four × Condition interaction, $F(1, 76) = 2.309, p = .133, \eta^2 = .029$, nor significant WMC × Condition interaction, $F(1, 76) = 0.684, p = .411, \eta^2 = .009$, for the modified data. Another ANCOVA was then run with the new dataset, using the baseline Speed Four data and WMC as the covariates, keeping the Speed Four 1 × Condition interaction term only while excluding the WMC × Condition interaction term for its nonsignificance. The assumption of the homogeneity of variance was still not met, $F(1, 80) = 6.433, p = .013$. This operation confirmed that there was no significant Speed Four 1 × Condition interaction, $F(1, 77) = 2.031, p = .158$. Still another ANCOVA was run with the new dataset, using the baseline data and WMC
as the covariates, keeping the WMC × Condition interaction term only while excluding the Speed Four 1 × Condition interaction term. The assumption of the homogeneity of variance was still not met, $F (1, 80) = 4.918$, $p = .029$. Again, there was no significant WMC × Condition interaction found, $F (1, 77) = 0.390$, $p = .534$.

A final ANCOVA was then run, with the baseline Speed Four data and WMC as two covariates, using the factorial mode and excluding any interaction terms. The homogeneity of variance was still not met, $F (1, 80) = 5.219$, $p = .025$. The ANCOVA test showed no statistically significant main effect of Condition, $F (1, 78) = 0.001$, $p = .970$, $\eta^2 = .000$.

Table 4.21

Results of ANCOVA for Speed Four after Normalization

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA (N=43)</td>
<td>NTA (N=39)</td>
<td>$F$</td>
</tr>
<tr>
<td>Speed 4</td>
<td>13.438 (3.317, 13.088)</td>
<td>12.683 (3.103, 13.069)</td>
<td>0.001</td>
</tr>
</tbody>
</table>


* $p < .05$.

Table 4.21 shows the adjusted means and the original means, the main effects of Condition, the results of all pairwise comparisons, and effect sizes, for the reoperation of ANCOVA.

In summary, in terms of Speed Four, after normalization of the data, no statistically significant covariate × Condition interaction was found, neither was there any statistically significant main effect of Condition. A $t$-test was also run with the new main task data set, yielding $t = 1.128$, $df = 81$, $p = .263$, confirming the absence of the main effect.
4.3.5.2.3.3 General Complexity

For general complexity, given the significant baseline General Complexity × Condition interaction, $F(1, 79) = 6.942, p = .010, \eta^2 = .081$ and no significant WMC × Condition interaction, $F(1, 79) = 1.016, p = .317, \eta^2 = .013$, an ANCOVA was re-run, using the baseline scores and WMC as two covariates, keeping the baseline General Complexity × Condition interaction term only while excluding the WMC × Condition interaction term. The assumption of the homogeneity of variance was still met, $F(1, 83) = 0.940, p = .760$. There was still a significant interaction between baseline General Complexity and Condition, $F(1, 80) = 6.722, p = .011, \eta^2 = .078$. A confirmative ANCOVA was also run, using only the WMC × Condition interaction term, with WMC and baseline General Complexity as two covariates. The assumption of the homogeneity of variance was still met, $F(1, 83) = 0.288, p = .593$. There was still no significant interaction between the WMC scores and Condition, $F(1, 80) = 0.743, p = .391, \eta^2 = .009$.

Table 4.22 shows the adjusted means and the original means, the main effects of Condition, the results of all pairwise comparisons, and the effect sizes, for the ANCOVA analysis that included only the baseline General Complexity × Condition interaction term.

Table 4.22

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>$M$ (SD, adjusted $M$)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons $p$ (effect size)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>$F$</td>
</tr>
<tr>
<td>General complexity</td>
<td>12.358 (2.434, 12.831)</td>
<td>12.968 (1.980, 12.697)</td>
<td>0.151</td>
</tr>
</tbody>
</table>

*Note. TA: $N = 43$, NTA: $N = 42$. *p < .05.
As Table 4.22 shows, there was no main effect of Condition on main task General Complexity.

A t-test was then run with the main task data set to compare the means of the two groups on General Complexity, yielding $t = -1.265$, $df = 83$, $p = .209$, indicating no statistically significant difference between the two groups on the particular measure.

A scatterplot of main task General Complexity by baseline task General Complexity with separate regression lines for the TA and NTA group was drawn to examine how the effects of Condition differed for participants with different baseline general complexity performance. Figure 4.5 shows the scatterplot.

![Figure 4.5 Baseline General Complexity x Condition Interaction on General Complexity](image)

As the scatterplot shows, TA appeared to have a different effect on general complexity for participants who stayed at the high and low ends of the measure.
Participants whose compositions demonstrated low general complexity tended to write less well in terms of the measure under the TA condition while those whose compositions demonstrated high general complexity tended to write even better in terms of general complexity. In other words, although there was no statistically significant main effect of Condition on general complexity, TA had a differential effect on participants with different levels of baseline general complexity performance, detrimental for those who had a low level of general complexity performance, but facilitative for those who had a high level of that aspect of performance.

I then conducted a univariate test with Condition and Grouping (which contained the high, mid, and low subgroups grouped based on baseline General Complexity) as two independent variables, the main task General Complexity as a dependent variable, the baseline task General Complexity and WMC as two covariates, keeping the Condition × Grouping interaction term only, to examine if there was significant Condition × Grouping interaction. Follow-up pairwise comparisons were also carried out following the Bonferroni procedure to examine the differences between the subgroups within the same condition and across the conditions. The assumption of homogeneity of variance was met, $F(5, 79) = 1.389, p = .238$. A significant interaction between Condition and Grouping was found, $F(2, 77) = 3.135, p = .049, \eta^2 = .075$.

Table 4.23

Descriptive Statistics for Different General Complexity × Condition Groups

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subgroups</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower Bound</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>12.769</td>
<td>.684</td>
<td>11.406</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.495</td>
<td>.404</td>
<td>11.692</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12.834</td>
<td>.583</td>
<td>11.673</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>14.078</td>
<td>.644</td>
<td>12.797</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.483</td>
<td>.419</td>
<td>11.649</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>11.975</td>
<td>.533</td>
<td>10.914</td>
</tr>
</tbody>
</table>

Note. Condition 1 = NTA, 2 = TA; Subgroup 1 = High, 2 = Mid, 3 = Low.
*Covariates appearing in the model are evaluated at the following values: General Complexity 1 = 12.142, WMC = .782.
Within the same (either NTA or TA) condition, pairwise comparisons of groups showed that there was no significant difference between different General Complexity groups in main task General Complexity, controlling for prior differences in General Complexity and WMC (for the NTA condition, all $p = 1.000$, and for the TA condition, $p = .138$ for High versus Mid, $p = .123$ for High versus Low, and $p = 1.000$ for Mid versus Low). However, across the two different conditions, pairwise comparisons of the same levels of groups indicated a significant difference for the high General Complexity group, $p = .049$, in favour of the TA condition, though the differences between the mid General Complexity groups and between the low General Complexity groups did not achieve significance, $p = .984$ and $p = .139$, both in favour of the NTA condition.

Figure 4.6 shows the results of the between- and within-Condition comparisons.
A further analysis comparing the two conditions at seven different levels of baseline General Complexity scores (8.802, 9.316, 10.367, 11.703, 13.702, 15.191, and 16.595, which corresponded to the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for baseline General Complexity scores) also showed that significant differences between the two conditions existed at the higher ends of the percentiles, at 15.191, \( p = .026 \) and at 16.595, \( p = .017 \), the difference approached significance at the lowest percentile level, 8.802, \( p = .053 \), and it did not reach significance at other percentile levels, at 9.316, \( p = .074 \), at 10.367, \( p = .193 \), at 11.703, \( p = .912 \), and at 13.702, \( p = .084 \). The results confirmed the finding that TA helped high General Complexity L2 writers with the very measure but prevented low General Complexity L2 writers from keeping up the same measure.

4.3.5.2.3.4 Correct Verb Use

For CVU, given that the initial ANCOVA analysis found significant baseline CVU × Condition interaction, \( F (1, 79) = 7.547, p = .007, \eta^2 = .087 \) but no significant WMC × Condition interaction, \( F (1, 79) = 0.315, p = .576, \eta^2 = .004 \), another ANCOVA was run, using the baseline CVU scores and WMC as two covariates, keeping the baseline CVU × Condition interaction term only while excluding the WMC × Condition interaction term. The assumption of the homogeneity of variance was still met, \( F (1, 83) = 0.040, p = .842 \). There was still a significant interaction between the baseline general complexity scores and Condition, \( F (1, 80) = 7.461, p = .008, \eta^2 = .085 \). A confirmative ANCOVA was also run, using only the WMC × Condition interaction term, with WMC and baseline CVU as two covariates. The assumption of the homogeneity of variance was still met, \( F (1, 83) = 0.436, p = .511 \). There was still no significant interaction between the WMC scores and Condition, \( F (1, 80) = 0.153, p = .697, \eta^2 = .002 \).

Table 4.24 shows the adjusted means in juxtaposition with the original means, the main effects of Condition, the results of all pairwise comparisons, and the effect sizes,
for the ANCOVA analysis with the baseline CVU × Condition interaction term only.

As Table 4.24 shows, there was no main effect of Condition on main task CVU, $F(1, 80) = 0.595, p = .443, \eta^2 = .007$. The $t$-test run with the main task data set confirmed the lack of reactive effects on CVU, $t = -0.778, df = 83, p = .439$.

Table 4.24

Results of ANCOVA for CVU with Interaction Re-examined

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>$F$</td>
</tr>
<tr>
<td>CVU</td>
<td>.798 (.113, .800)</td>
<td>.814 (.083, .814)</td>
<td>0.595</td>
</tr>
</tbody>
</table>

Note. TA: $N = 43$, NTA: $N = 42$.
*p < .05.

A scatterplot of main task CVU by baseline task CVU, with separate regression lines for the TA and NTA group, was drawn to examine how the effects of Condition differed for participants with different baseline CVU performance. Figure 4.7 shows the scatterplot.

As the scatterplot shows, TA appeared to have a different effect on CVU for participants who stayed at the high and low ends of the measure. Participants whose compositions demonstrated a low level of CVU tended to write worse than they had achieved silently in terms of the measure under the TA condition while those whose compositions demonstrated a high level of CVU tended to write with a better level of CVU. In other words, although there was no statistically significant main effects of Condition on CVU, TA had a differential effect on participants with different levels of CVU, being detrimental for those who had a low level while being facilitative for those who stayed at a high level.
I then conducted a univariate test with Condition and Grouping (which contained the high, mid, and low subgroups grouped based on baseline CVU) as two independent variables, the main task CVU as a dependent variable, the baseline task CVU and WMC as two covariates, keeping the Condition × Grouping interaction term only, to examine if there was significant Condition × Grouping interaction. Follow-up pairwise comparisons were also carried out following the Bonferroni procedure to examine if there was any significant difference between the subgroups within and across the conditions. The assumption of homogeneity of variance was met, $F (5, 79) = 0.782, p = .565$. No significant interaction between Condition and Grouping was found, $F (2, 77) = 2.344, p = .103, \eta^2 = .057$. Table 4.25 lists the descriptive statistics of the three CVU groups in relation to Condition.

Figure 4.7 Baseline CVU and Condition Interaction
Table 4.25

Descriptive Statistics for Different CVU × Condition Groups

<table>
<thead>
<tr>
<th>Condition</th>
<th>Subgroups</th>
<th>Mean</th>
<th>Std. Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower Bound</td>
<td>Upper Bound</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>.796</td>
<td>.031</td>
<td>.735</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.804</td>
<td>.024</td>
<td>.757</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.847</td>
<td>.041</td>
<td>.766</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>.825</td>
<td>.030</td>
<td>.765</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>.799</td>
<td>.021</td>
<td>.757</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>.771</td>
<td>.033</td>
<td>.706</td>
</tr>
</tbody>
</table>

Note. Condition 1 = NTA, 2 = TA; Subgroup 1 = High, 2 = Mid, 3 = Low.
*Covariates appearing in the model are evaluated at the following values: WMC = .782, CVU 1 = .823.

Within the same (either NTA or TA) condition, pairwise comparisons of groups showed that there was no significant difference between different CVU groups in main task CVU, controlling for prior differences in CVU and WMC (all \( p = 1.000 \) except for \( p = .942 \) in the case of High versus low under the TA condition). However, across the two different conditions, pairwise comparisons of the same levels of groups indicated a significant decline under the TA condition for the low CVU group, \( p = .038 \), though the differences between the high CVU groups and between the mid CVU groups did not achieve significance, \( p = .355 \) and \( p = .878 \). Figure 4.8 shows the results of the between-and within-Condition comparisons.

A further analysis comparing the two conditions at seven different levels of baseline CVU scores (.645, .696, .768, .833, .886, .934, and .947, which corresponded to the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for baseline CVU scores) showed that the difference between the two conditions was significant at the lower ends of the percentiles, at .645, \( p = .006 \), at .696, \( p = .008 \), and at .768, \( p = .034 \), but was not significant at the middle percentile and the higher ends of the percentiles, at .833, \( p = .680 \), at .886, \( p = .302 \), at .934, \( p = .090 \), and at .947, \( p = .070 \). The results showed that in terms of CVU, the differential effects of TA consisted mainly in its significantly detrimental effects on low CVU L2 writers.
4.3.5.2.3.5 Organization

In the initial ANCOVA analysis, the organization data showed a statistically significant WMC × Condition interaction, $F (1, 79) = 3.945, p = .050, \eta^2 = .048$, but no statistically significant baseline Organization × Condition interaction, $F (1, 79) = 0.758, p = .387, \eta^2 = .010$. For organization, both the baseline and main task datasets were first examined for normality. It transpired that the Skewness and Kurtosis values were all acceptable but the p values of the Kolmogorov-Smirnov and the Shapiro-Wilk tests were all significant. This may be attributable to the limited range of organization scores (from 10 to 18). No outliers were identified. Given the results of the initial ANCOVA analysis, another ANCOVA was run, using baseline Organization and WMC as two covariates, keeping the WMC × Condition interaction term only and excluding the baseline Organization × Condition interaction term. The assumption of the homogeneity of
variance was still not met, \( F(1, 83) = 11.914, p = .001 \), but the ratio of the greater group variance to the smaller variance \( (2.527^2/1.578^2 = 2.565) \) was below 4, indicating that further analysis based on such variances was safe. The ANCOVA test indicated that there was still significant interaction between WMC scores and Condition, \( F(1, 80) = 4.591, p = .035, \eta^2 = .054 \).

Table 4.26 shows the adjusted means in juxtaposition with the original means, the main effects of Condition, the results of all pairwise comparisons, and the effect sizes, for this reoperation.

Table 4.26

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA</td>
<td>14.605 (2.527, 14.699)</td>
<td>F=5.657, df=1, 80, p=.020*</td>
<td>NT A vs. TA</td>
</tr>
<tr>
<td>NTA</td>
<td>15.738 (1.578, 15.628)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. TA: N = 43, NTA: N = 42. *p < .05.

As Table 4.26 shows, there was a main effect of Condition on main task Organization, \( F(1, 80) = 5.657, p = .020, \eta^2 = .066 \). A t-test run on the main task data set comparing the TA and the NTA groups confirmed the existence of such a main effect, \( t = -2.486, df = 70.687, p = .015 \).

The significant WMC and Condition interaction means that the effect of Condition should be further examined by referring to the mediating role of WMC. A scatterplot of main task Organization by baseline task Organization with separate regression lines for the TA and NTA group was drawn to examine how the effects of Condition on Organization differed for participants with different WMC.

Figure 4.9 shows the scatterplot.
As the scatterplot shows, while TA overall appeared to be detrimental to Organization, its detrimental effect was mainly felt by low-WMC L2 writers, and it was much more detrimental to low-WMC L2 writers than it was facilitative to high-WMC L2 writers.

I then conducted a univariate test with Condition and Grouping (which contained the high, mid, and low subgroups grouped based on their WMC Average) as two independent variables, the main task Organization as a dependent variable, the baseline task Organization scores and WMC as two covariates, keeping the Condition × Grouping interaction term only, to examine if there was significant Condition × Grouping interaction. Follow-up pairwise comparisons were also carried out following the Bonferroni procedure to examine if there was any significant difference between the subgroups within the same condition and across the conditions. The assumption of
homogeneity of variance was met, $F(5, 79) = 1.370, p = .245$. Significant interaction between Condition and Grouping was found, $F(2, 77) = 5.568, p = .006, \eta^2 = .126$.

Table 4.27 lists the descriptive statistics of the WMC subgroups relating to Condition.

Within the NTA condition, pairwise comparisons showed that there was no significant difference in main task Organization between the three WMC subgroups, controlling for prior differences in Organization and WMC ($p = .833$ for High versus Mid; $p = .803$ for High versus Low, and $p = 1.000$ for Mid versus Low, with the higher-WMC participants having slightly better organization). Within the TA condition, pairwise comparisons showed no significant difference in main task Organization between the high-WMC and mid-WMC groups, $p = 1.000$, in favour of the high-WMC group, but significant difference between the high-WMC group and the low-WMC group, $p = .022$, and the Mid WMC group and the low-WMC group, $p = .003$, disadvantaging the low-WMC participants. Also, between the two different conditions, pairwise comparisons of the same levels of WMC groups indicated a significant decline under the TA condition for the low WMC group, $p = .000$, though the difference between the high WMC groups and that between the mid WMV groups did not achieve significance, $p = .468$ and $p = .904$, both in favour of the NTA condition.

Figure 4.10 shows the results of the between- and within-condition comparisons.
A further analysis comparing the two conditions at seven different levels of WMC (.509, .556, .725, .809, .889, .951, and .968), which corresponded to the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles for WMC scores) also showed that the difference between the two conditions was significant at the lower ends of the percentiles, at .509, \( p = .004 \), at .556, \( p = .003 \), and at .725, \( p = .004 \), approached significance at .809, \( p = .058 \), but was not significant at the higher ends of the percentiles, at .889, \( p = .600 \), at .951, \( p = .829 \), and at .968, \( p = .720 \). The results confirmed that the detrimental effect of TA on Organization was mainly felt by low-WMC L2 writers.

4.3.5.2.4 Summary

Table 4.28 summarizes the finalized descriptive statistics and results of the ANCOVAs and pairwise comparisons for the 20 measures.
Table 4.28

**Finalized Descriptive Statistics and Results of ANCOVAs and Pairwise Comparisons**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>F</td>
</tr>
<tr>
<td>Time</td>
<td>37.558</td>
<td>40.902</td>
<td>0.364</td>
</tr>
<tr>
<td>Speed 1</td>
<td>9.387</td>
<td>8.939</td>
<td>0.006</td>
</tr>
<tr>
<td>Speed 2</td>
<td>9.981</td>
<td>9.436</td>
<td>0.035</td>
</tr>
<tr>
<td>Speed 3</td>
<td>12.543</td>
<td>11.839</td>
<td>0.002</td>
</tr>
<tr>
<td>Speed 4</td>
<td>13.438</td>
<td>12.683</td>
<td>0.001</td>
</tr>
<tr>
<td>ND</td>
<td>906 (.921)</td>
<td>906 (.921)</td>
<td>5.916</td>
</tr>
<tr>
<td>Length 1</td>
<td>341.090</td>
<td>356.000</td>
<td>1.306</td>
</tr>
<tr>
<td>Length 2</td>
<td>362.630</td>
<td>374.800</td>
<td>0.245</td>
</tr>
<tr>
<td>Length 3</td>
<td>453.770</td>
<td>478.120</td>
<td>3.700</td>
</tr>
<tr>
<td>Length 4</td>
<td>479.140</td>
<td>505.320</td>
<td>1.947</td>
</tr>
<tr>
<td>General complexity</td>
<td>12.358</td>
<td>12.968</td>
<td>0.151</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>1.987</td>
<td>2.038</td>
<td>0.019</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>7.791</td>
<td>8.186</td>
<td>2.327</td>
</tr>
<tr>
<td>SV</td>
<td>15.511</td>
<td>16.214</td>
<td>0.203</td>
</tr>
<tr>
<td>LD (D)</td>
<td>61.554</td>
<td>66.913</td>
<td>4.492</td>
</tr>
<tr>
<td>CVU</td>
<td>798 (.814)</td>
<td>798 (.814)</td>
<td>0.595</td>
</tr>
<tr>
<td>EFWs</td>
<td>88.404</td>
<td>89.787</td>
<td>0.064</td>
</tr>
<tr>
<td>EFCs</td>
<td>.545</td>
<td>.544</td>
<td>1.285</td>
</tr>
<tr>
<td>Content</td>
<td>20.977</td>
<td>22.095</td>
<td>1.176</td>
</tr>
<tr>
<td>Organization</td>
<td>14.605</td>
<td>15.738</td>
<td>5.657</td>
</tr>
</tbody>
</table>

*Note. For Time, ND, Length 2, Length 3, and Length 4, TA: N = 43, NTA: N = 41; for Speed 1, TA: N = 42, NTA: N = 40; for Speed 2, and Speed 3, TA: N = 42, NTA: N = 39; for Speed 4, TA: N = 43, NTA: N = 42. For Speed 1, Speed 2, Speed 3, Speed 4, and Length 3, the outliers were removed. For Speed 2, General Complexity, and CVU, significant baseline performance × Condition interaction was found. For Organization, significant WMC × Condition interaction was found. *p < .05.
Overall, controlling for baseline performance and WMC, TA significantly decreased three measures, namely, ND, $F(1, 80) = 5.916, p = .017, \eta^2 = .069$, lexical diversity ($D$), $F(1, 81) = 4.492, p = .037, \eta^2 = .053$, and Organization, $F(1, 80) = 5.657, p = .020, \eta^2 = .066$. Also, controlling for baseline performance and WMC, TA tended to have a negative effect on one more measure, Length Three, $F(1, 80) = 3.700, p = .058, \eta^2 = .044$.

A statistically significant interaction was found between baseline performance and Condition for Speed One, $F(1, 77) = 4.164, p = .045, \eta^2 = .051$, and for Speed Two, $F(1, 76) = 5.651, p = .020, \eta^2 = .069$, when the outliers in both the Speed One and Speed Two datasets were removed. Significant interaction was also found on General Complexity, $F(1, 80) = 6.722, p = .011, \eta^2 = .078$, and on CVU, $F(1, 80) = 7.461, p = .008, \eta^2 = .085$. In further analyses grouping participants based on their baseline performance and especially comparing main task performance across the two conditions at seven points of percentiles for baseline scores, the interaction on Speed One was not supported. The interaction on Speed Two was not supported by the grouping analyses either, and the group difference was only significant at the lowest percentile, the 5th percentile (i.e., at 5.867, $p = .037$), and approached significance at the second lowest percentile, the 10th percentile (i.e., at 6.799, $p = .055$). This indicated that while TA did not affect Speed Two overall, it worked for the slowest Speed Two L2 writers; that is, it was associated with their writing significantly more words per minute.

The interaction on General Complexity was corroborated by the grouping analyses, where a significant interaction between Condition and Grouping was found, $F(2, 77) = 3.135, p = .049, \eta^2 = .075$, and where the high General Complexity group under the TA condition significantly exceeded their counterparts who wrote under the NTA condition, $p = .049$. The significant interaction was confirmed when a comparison was made
between the two conditions at seven different levels of baseline General Complexity scores, where the difference between the two conditions reached significance at the two highest percentiles, the 90th and the 95th percentiles (i.e., at 15.191, \( p = .026 \), and at 16.595, \( p = .017 \)) and approached significance at the lowest percentile level, at 8.802, \( p = .053 \), indicating that TA did differentially affect high and low General Complexity performers, especially the former, promoting the performance of the former in the very measure while not exerting much influence on that of the latter.

The interaction on CVU was not supported by the initial grouping analyses, where no main interaction effect between Condition and Grouping was found, \( F(2, 77) = 2.344, \ p = .103, \ \eta^2 = .057 \). However, a finer comparison of subgroups indicated that the performance of the low CVU group under the TA condition was significantly overshadowed by their counterparts who wrote under the NTA condition, \( p = .038 \), their prior differences in the measure and WMC being controlled. The significant interaction was further confirmed when comparisons were made between the two conditions at seven different levels of baseline CVU scores, where significant differences between the two conditions were found at the three lowest percentiles, the 5th, 10th, and 25th percentiles (i.e., at .645, \( p = .006 \), at .696, \( p = .008 \), and at .768, \( p = .034 \)) but not at all the other points. This indicated that TA did differentially affect high and low CVU performers in that it significantly undermined the performance of the latter in the very measure.

A statistically significant interaction was also found between WMC Average and Condition on Organization, \( F(1, 80) = 4.591, \ p = .035, \ \eta^2 = .054 \). The interaction was corroborated by the grouping analyses, where a significant interaction between Condition and Grouping was found, \( F(2, 77) = 5.568, \ p = .006, \ \eta^2 = .126 \). A finer comparison of subgroups indicated that, under the TA condition, the low Organization
group was significantly outperformed by both the mid and high WMC groups ($p = .022$ for High versus Low, and $p = .003$ for Mid versus Low), though the three groups did not differ under the NTA condition. The comparison also indicated that, across the two conditions, the low WMC group showed a significant decline in Organization under the TA condition, when compared with their counterparts in the NTA condition, $p = .000$. The significant interaction was further specified when a comparison was made between the two conditions at seven different levels of WMC scores, where the difference between the two conditions was significant at the three lowest percentiles, the 5th, 10th, and 25th percentiles (i.e., at .509, $p = .004$, at .556, $p = .003$, and at .725, $p = .004$) and it approached significance at the 50th percentile, at .809, $p = .058$, but not at all the other points. This indicated that TA did differentially affect high and low WMC performers in that it significantly undermined the performance of the latter in Organization.

The above interaction effects indicated that TA had differential effects on L2 writers of different aspects of proficiency. The differential effects were not obvious in terms of writing speed, given that it only significantly increased Speed Two of those who belonged to the lowest five percent. The interaction was more marked in terms of General Complexity, where previously high achievers demonstrated a significant boost in the measure, while previously low achievers did not show much influence. The interaction was also noticeable in terms of CVU, where previously low performers used verb forms even significantly worse, while previously high performers did not show much influence.

TA also had differential effects on L2 writers of different WMC in that under the TA condition, L2 writers with low WMC performed significantly worse than they had achieved in organizing compositions while those with high WMC did not appear to be influenced by engaging in TA.
4.3.5.3 The ANCOVAs with WMC One as a covariate

Following the above procedures, ANCOVAs and follow-up analyses were conducted on the main task data, this time with baseline task data and WMC One as two covariates.

Table 4.29 summarizes the final descriptive statistics and results of the ANCOVAs and pairwise comparisons for the 20 measures, after interaction relations and normality of data were considered.

Overall, controlling for baseline task performance and WMC One, TA significantly decreased three measures, namely, ND, $F(1, 80) = 5.905, p = .017, \eta^2 = .069$, lexical diversity ($D$), $F(1, 81) = 4.499, p = .037, \eta^2 = .053$, and Organization, $F(1, 80) = 5.094, p = .027, \eta^2 = .060$. Also, controlling for baseline performance and WMC One, the negative effect of TA on Length Three approached statistical significance, $F(1, 80) = 3.680, p = .059, \eta^2 = .044$. These results were similar to those found when WMC Average was used as a covariate.

Interactions between baseline task performance and Condition were also found, when both WMC One and baseline performance were used as two covariates. The interaction between baseline Speed One and Condition approached statistical significance, $F(1, 77) = 3.859, p = .053, \eta^2 = .048$, and that between baseline Speed Two and Condition was found to be significant, $F(1, 76) = 5.187, p = .026, \eta^2 = .064$, when the outliers in both the Speed One and Speed Two datasets were removed. Statistically significant interactions were also found between baseline General Complexity and Condition, $F(1, 80) = 6.839, p = .011, \eta^2 = .079$, and between baseline CVU and Condition, $F(1, 80) = 7.826, p = .006, \eta^2 = .089$. In terms of WMC One, a significant interaction was found between WMC One and Condition on Organization, $F(1, 80) = 5.532, p = .021$. 

173
Table 4.29

Results of Analyses with WMC One as a Covariate

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>F</td>
</tr>
<tr>
<td>Time</td>
<td>37.558</td>
<td>40.902</td>
<td>0.306</td>
</tr>
<tr>
<td></td>
<td>(9.277, 38.788)</td>
<td>(8.040, 39.612)</td>
<td></td>
</tr>
<tr>
<td>Speed 1</td>
<td>9.387</td>
<td>8.939</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>(2.106, 9.212)</td>
<td>(2.243, 9.249)</td>
<td></td>
</tr>
<tr>
<td>Speed 2</td>
<td>9.981</td>
<td>9.436</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>(2.227, 9.813)</td>
<td>(2.435, 9.752)</td>
<td></td>
</tr>
<tr>
<td>Speed 3</td>
<td>12.543</td>
<td>11.839</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>(2.968, 12.189)</td>
<td>(2.526, 12.220)</td>
<td></td>
</tr>
<tr>
<td>Speed 4</td>
<td>13.438</td>
<td>12.683</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(3.317, 13.083)</td>
<td>(3.103, 13.074)</td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>.906</td>
<td>.921</td>
<td>5.905</td>
</tr>
<tr>
<td></td>
<td>(.055, .904)</td>
<td>(.041, .922)</td>
<td></td>
</tr>
<tr>
<td>Length 1</td>
<td>341.090</td>
<td>356</td>
<td>1.292</td>
</tr>
<tr>
<td></td>
<td>(40.769, 343.6)</td>
<td>(59.171, 353.4)</td>
<td></td>
</tr>
<tr>
<td>Length 2</td>
<td>362.630</td>
<td>374.800</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>(44.204, 366.4)</td>
<td>(66.549, 370.9)</td>
<td></td>
</tr>
<tr>
<td>Length 3</td>
<td>453.770</td>
<td>478.120</td>
<td>3.680</td>
</tr>
<tr>
<td></td>
<td>(59.136, 454.8)</td>
<td>(77.708, 477.1)</td>
<td></td>
</tr>
<tr>
<td>Length 4</td>
<td>479.140</td>
<td>505.320</td>
<td>1.925</td>
</tr>
<tr>
<td></td>
<td>(62.306, 483.7)</td>
<td>(90.261, 500.5)</td>
<td></td>
</tr>
<tr>
<td>General complexity</td>
<td>12.358</td>
<td>12.968</td>
<td>0.154</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>2.434, 12.832</td>
<td>(1.980, 12.697)</td>
<td></td>
</tr>
<tr>
<td>SubcLHoral complexity</td>
<td>1.987, 2.038</td>
<td>0.020</td>
<td>1.81</td>
</tr>
<tr>
<td>SV</td>
<td>7.791</td>
<td>8.186</td>
<td>2.320</td>
</tr>
<tr>
<td></td>
<td>(.178, 7.847)</td>
<td>(.924, 8.129)</td>
<td></td>
</tr>
<tr>
<td>LD (D)</td>
<td>15.511</td>
<td>16.214</td>
<td>0.198</td>
</tr>
<tr>
<td></td>
<td>(3.104, 15.717)</td>
<td>(3.273, 16.004)</td>
<td></td>
</tr>
<tr>
<td>CVU</td>
<td>61.554</td>
<td>66.913</td>
<td>4.499</td>
</tr>
<tr>
<td></td>
<td>(.9459, 62.226)</td>
<td>(12.112, 66.225)</td>
<td></td>
</tr>
<tr>
<td>EFWs</td>
<td>.798</td>
<td>.814</td>
<td>0.597</td>
</tr>
<tr>
<td></td>
<td>(.113, .800)</td>
<td>(.083, .814)</td>
<td></td>
</tr>
<tr>
<td>EFCs</td>
<td>88.404</td>
<td>89.787</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>(5.344, 89.153)</td>
<td>(4.008, 89.021)</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>.545</td>
<td>.554</td>
<td>1.298</td>
</tr>
<tr>
<td></td>
<td>(.124, .560)</td>
<td>(.130, .538)</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>20.977</td>
<td>22.095</td>
<td>1.146</td>
</tr>
<tr>
<td></td>
<td>(3.158, 21.188)</td>
<td>(3.392, 21.879)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.605</td>
<td>15.738</td>
<td>5.094</td>
</tr>
<tr>
<td></td>
<td>(2.527, 14.748)</td>
<td>(1.578, 15.648)</td>
<td></td>
</tr>
</tbody>
</table>

Note. For Time, ND, Length 2, Length 3, and Length 4, TA: N = 43, NTA: N = 41; for Speed 1, TA: N = 42, NTA: N = 40; for Speed 2, and Speed 3, TA: N = 42, NTA: N = 39; for Speed 4, TA: N = 43, NTA: N = 39; for all other measures, TA: N = 43, NTA: N = 42. For Speed 1, Speed 2, Speed 3, Speed 4, and Length 3, the outliers were removed. For Speed 2, General Complexity, and CVU, significant baseline data × Condition interaction was found. For Organization, significant WMC × Condition interaction was found. *p < .05.
The interaction between baseline Speed One performance and Condition found in the ANCOVA analyses was not supported by the results of further analyses which grouped the participants based on their baseline Speed One measures and which compared main task performance across the two conditions at seven points of percentiles for the baseline Speed One data. The interaction of another speed measure, Speed Two, with Condition was not supported by the grouping analyses either, but the percentile analyses showed that the difference between the TA and the NTA groups was significant at the lowest percentile, the 5th percentile (i.e., at 5.867, \( p = .046 \)). This indicated that while TA did not affect Speed Two overall, it impacted the slowest Speed Two L2 writers, who produced significantly more words per minute.

The interaction with General Complexity was corroborated by the grouping analyses, where significant interaction between Condition and Grouping was found, \( F (2, 77) = 3.110, \ p = .050, \ \eta^2 = .075 \), and the high General Complexity group under the TA condition significantly exceeded their counterparts who wrote under the NTA condition, \( p = .049 \). The significant interaction was further confirmed when comparisons were made between the two conditions at seven different levels of baseline General Complexity scores. Significant difference between the two conditions was found at the two highest percentiles, the 90th and the 95th percentiles (i.e., at 15.191, \( p = .025 \), and at 16.595, \( p = .016 \)), in favour of the TA condition. In contrast, the difference between the two conditions approached significance at the lowest percentile level, at 8.802, \( p = .052 \), in favour of the NTA condition. These results suggest that TA did differentially affect high and low General Complexity performers, promoting the performance of the former and potentially impeding that of the latter.

The interaction on CVU was not supported by the initial grouping analyses, where no main interaction effect between Condition and Grouping was found, \( F (2, 77) = 2.572 \),
\[ p = 0.083, \eta^2 = 0.063. \] However, a finer comparison of subgroups indicated that the performance of the low CVU group under the TA condition was significantly overshadowed by their counterparts who wrote under the NTA condition, \( p = 0.032 \), controlling for prior differences in the measure and WMC One. The significant interaction was further confirmed when comparisons were made between the two conditions at seven different levels of baseline CVU scores, where significant differences between the two conditions were found at the three lowest percentiles, the 5th, 10th, and 25th percentiles (i.e., at .645, \( p = 0.005 \), at .696, \( p = 0.007 \), at .768, \( p = 0.031 \)) but not at all the other points. This indicated that TA did differentially affect high and low CVU performers, given that it significantly undermined the performance of the latter but did not exert much influence on the performance of the former.

The WMC One \( \times \) Condition interaction on Organization was corroborated by the grouping analyses, where a significant interaction between Condition and Grouping was found, \( F(2, 77) = 4.806, p = 0.011, \eta^2 = 0.111 \). A finer comparison of subgroups indicated that, across the two conditions, the low WMC One group showed a significant decline in Organization under the TA condition, when compared with their counterparts in the NTA condition, \( p = 0.000 \). The significant interaction was further confirmed when comparisons were made between the two conditions at seven different levels of WMC One scores, where significant differences between the two conditions were found at the lower end of the percentile spectrum, the 5th, 10th, 25th, and 50th percentiles (i.e., at .252, \( p = 0.003 \), at .381, \( p = 0.002 \), at .580, \( p = 0.002 \), and at .733, \( p = 0.035 \)). These results indicated that TA did differentially affect high and low WMC One performers given that it significantly undermined the performance of the latter in Organization while having minimal effect on that of the former.
The above interactions indicate that TA had differential effects on L2 writers. The differential effects were not obvious in terms of writing speed, given that TA only significantly increased Speed Two of those who belonged to the lowest five percent of the Speed Two group. The interaction was more conspicuous in terms of General Complexity, where previously high achievers were boosted in the measure, while previously low achievers did not show much influence from TA. The interaction was equally noticeable in terms of CVU, where low performers’ performance in use of verb forms significantly deteriorated, while high performers did not show much influence.

TA also had differential effects on L2 writers of different WMC One. Under the TA condition, L2 writers with low WMC One performed significantly worse in organizing their compositions while those with high WMC One were not affected.

4.3.5.4 The ANCOVAs with WMC Two as a covariate

Similarly, ANCOVAs and follow-up analyses were conducted on the main task data, with the baseline task data and WMC Two as two covariates.

Table 4.30 summarizes the finalized descriptive statistics and results of the ANCOVAs and pairwise comparisons for the 20 measures, which were obtained after interaction relations and normality of data were considered.

Overall, controlling for baseline performance and WMC Two, TA significantly decreased three measures, namely, ND, \( F(1, 80) = 5.985, p = .017, \eta^2 = .070 \), LD \( (D) \), \( F(1, 80) = 4.500, p = .037, \eta^2 = .053 \), and Organization, \( F(1, 81) = 5.113, p = .026, \eta^2 = .059 \). Also, controlling for baseline performance and WMC Two, the effect of TA on Length Three, which was negative, approached statistical significance, \( F(1, 80) = 3.889, p = .052, \eta^2 = .046 \).

Significant interactions were found between WMC Two and Condition on Length Three, \( F(1, 80) = 6.185, p = .015, \eta^2 = .072 \), Length Four, \( F(1, 79) = 5.320, p = .024 \),
### Table 4.30

**Results of Analyses with WMC Two as a Covariate**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>$M$ (SD, adjusted $M$)</th>
<th>ANCOVA</th>
<th>Pairwise comparisons $p$ (effect size)</th>
<th>NTA vs. TA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TA</td>
<td>NTA</td>
<td>$F$</td>
<td>$df$</td>
</tr>
<tr>
<td>Time</td>
<td>37.558 (9.277, 38.744)</td>
<td>40.902 (8.040, 39.658)</td>
<td>0.367</td>
<td>1, 80</td>
</tr>
<tr>
<td>Speed 1</td>
<td>9.387 (2.106, 9.228)</td>
<td>8.939 (2.243, 9.228)</td>
<td>0.000</td>
<td>1, 77</td>
</tr>
<tr>
<td>Speed 2</td>
<td>9.981 (2.227, 9.826)</td>
<td>9.436 (2.435, 9.732)</td>
<td>0.065</td>
<td>1, 76</td>
</tr>
<tr>
<td>Speed 3</td>
<td>12.543 (2.968, 12.194)</td>
<td>11.839 (2.526, 12.215)</td>
<td>0.002</td>
<td>1, 77</td>
</tr>
<tr>
<td>Speed 4</td>
<td>13.438 (3.171, 13.100)</td>
<td>12.683 (3.103, 13.055)</td>
<td>0.008</td>
<td>1, 78</td>
</tr>
<tr>
<td>ND</td>
<td>.906 (.055, .904)</td>
<td>.921 (.041, .922)</td>
<td>5.985</td>
<td>1, 80</td>
</tr>
<tr>
<td>Length 1</td>
<td>341.090 (40.769, 343.5)</td>
<td>356.000 (59.171, 353.5)</td>
<td>1.363</td>
<td>1, 81</td>
</tr>
<tr>
<td>Length 2</td>
<td>362.630 (44.204, 365.100)</td>
<td>374.800 (66.549, 370.200)</td>
<td>0.326</td>
<td>1, 79</td>
</tr>
<tr>
<td>Length 3</td>
<td>453.770 (59.136, 456.800)</td>
<td>482.020 (80.813, 478.700)</td>
<td>3.889</td>
<td>1, 80</td>
</tr>
<tr>
<td>Length 4</td>
<td>479.140 (62.306, 483.700)</td>
<td>505.320 (90.261, 500.600)</td>
<td>2.099</td>
<td>1, 79</td>
</tr>
<tr>
<td>General complexity</td>
<td>12.358 (2.434, 12.839)</td>
<td>12.968 (1.980, 12.698)</td>
<td>0.164</td>
<td>1, 80</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>1.987 (2.88, 2.016)</td>
<td>2.038 (.297, 2.008)</td>
<td>0.022</td>
<td>1, 81</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>7.791 (1.178, 7.848)</td>
<td>8.186 (.924, 8.127)</td>
<td>2.270</td>
<td>1, 81</td>
</tr>
<tr>
<td>SV</td>
<td>15.511 (3.104, 15.718)</td>
<td>16.214 (3.273, 16.003)</td>
<td>0.194</td>
<td>1, 81</td>
</tr>
<tr>
<td>LV ($D$)</td>
<td>61.554 (9.459, 62.104)</td>
<td>66.913 (12.112, 66.100)</td>
<td>4.500</td>
<td>1, 80</td>
</tr>
<tr>
<td>CVU</td>
<td>.798 (.113, .800)</td>
<td>.814 (.083, .814)</td>
<td>0.545</td>
<td>1, 80</td>
</tr>
<tr>
<td>EFWs</td>
<td>88.404 (5.344, 89.164)</td>
<td>89.787 (4.008, 89.010)</td>
<td>0.072</td>
<td>1, 81</td>
</tr>
<tr>
<td>EFCs</td>
<td>.545 (.124, .561)</td>
<td>.554 (.130, .538)</td>
<td>1.420</td>
<td>1, 81</td>
</tr>
<tr>
<td>Content</td>
<td>20.977 (3.158, 21.175)</td>
<td>22.095 (3.392, 21.892)</td>
<td>1.237</td>
<td>1, 81</td>
</tr>
<tr>
<td>Organization</td>
<td>14.605 (2.527, 14.719)</td>
<td>15.738 (1.578, 15.621)</td>
<td>5.113</td>
<td>1, 81</td>
</tr>
</tbody>
</table>


For Speed 1, Speed 2, Speed 3, Speed 4, and Length 3, the outliers were removed.

For Speed 2, Length 2, General Complexity, and CVU, significant baseline performance × Condition interaction was found. For Length 3, Length 4 and $D$, significant WMC Two × Condition interaction was found.

*p < .05.
\( \eta^2 = .063, \) and \( D, F (1, 80) = 5.447, p = .022, \eta^2 = .064. \)

The interaction between baseline Speed One performance and Condition was not supported by the results of further analyses that grouped participants based on their baseline Speed One performance and that compared main task performance across the two conditions at seven points of percentiles for baseline Speed One data. The interaction on another speed measure, Speed Two, with Condition was not supported by the grouping analyses either, but was found significant at the lowest percentile, the 5\(^{th}\) percentile (i.e., at 5.867, \( p = .042 \)). This indicated that while TA did not affect Speed Two overall, it impacted the slowest Speed Two L2 writers, who wrote significantly more words per minute.

The interaction between baseline Length Two and Condition was corroborated by the grouping analyses, where a significant interaction between Condition and Grouping was found, \( F (2, 76) = 3.642, p = .031, \eta^2 = .087. \) Finer comparisons of subgroups within the same conditions indicated that the three Length Two subgroups differed significantly in terms of Length Two under the NTA condition (i.e., high vs. mid and high vs. low), but they were close in this measure under the TA condition.

Comparisons of subgroups across the two conditions indicated that the Length Two performance of the high achievers significantly dropped under the TA condition, compared with that of their counterparts who wrote under the NTA condition, \( p = .015. \) The significant interaction was further confirmed when comparisons were made between the two conditions at seven different levels of baseline Length Two scores. The difference in Length Two between the two conditions achieved significance at the highest percentile point, the 95th percentile (i.e., at 453.750, \( p = .039 \)) and approached significance at the second highest percentile point, the 90\(^{th}\) percentile (i.e., 424.50, \( p = .051 \)). This indicated that TA did differentially affect L2 writers in terms of Length
Two, causing the small portion of high baseline task Length Two performers to write significantly less productively while not exerting influence on low performers.

The interaction on General Complexity was corroborated by the grouping analyses, which found a significant interaction between Condition and Grouping, $F(2, 77) = 3.319$, $p = .041$, $\eta^2 = .079$. Further comparisons of the subgroups indicated that the high General Complexity group under the TA condition significantly exceeded their counterparts who wrote under the NTA condition, $p = .042$. The significant interaction was confirmed when comparisons were made between the two conditions at seven different levels of baseline General Complexity scores. Significant differences between the two conditions were found at the two highest percentiles, the 90th and the 95th percentiles (i.e., at 15.191, $p = .021$, and at 16.595, $p = .013$), in favour of the TA condition, and at the lowest percentile, the 5th percentile (at 8.802, $p = .045$), in favour of the NTA condition. The results indicated that TA did differentially affect high and low General Complexity performers, promoting the performance of the former while exerting significantly negative influence particularly on the very lowest performers.

The interaction on CVU was not supported by the initial grouping analyses, which found no main interaction effect between Condition and Grouping, $F(2, 77) = 2.457$, $p = .092$, $\eta^2 = .060$. However, finer comparisons of subgroups indicated that the performance of the low CVU group under the TA condition was significantly overshadowed by that of their counterparts who wrote under the NTA condition, $p = .033$, when the prior differences in the measure and WMC Two were controlled. The significant interaction was confirmed when comparisons were made between the two conditions at seven different levels of baseline CVU scores. Significant differences between the two conditions were found at the three lowest percentiles, the 5th, 10th, and 25th percentiles (i.e., at .645, $p = .005$, at .696, $p = .007$, at .768, $p = .032$) but not at all
the other points. This indicated that TA did differentially affect high and low CVU performers given that it significantly undermined the performance of the latter but did not exert much influence on the former.

The interaction between WMC Two and Condition on Length Three was not corroborated by the grouping analyses, which found no significant interaction between Condition and Grouping, $F(2, 76) = 2.692, p = .074, \eta^2 = .066$. However, further analyses comparing the subgroups of WMC Two indicated that the number of syllables produced by L2 writers with high WMC Two significantly dropped compared with that produced by their counterparts at the same WMC Two level of WMC Two who wrote under the NTA condition, $p = .012$. The significant WMC Two × Condition interaction was confirmed when comparisons were made between the two conditions at seven different levels of WMC Two percentiles. The difference in Length Three between the two conditions achieved significance at the higher ends of the percentile spectrum (i.e., at .870, $p = .017$, at .926, $p = .002$, at .976, $p = .002$, and at .981, $p = .002$). This indicated that TA did differentially affect L2 writers who had different levels of WMC Two, in terms of Length Three, given that it caused those with high WMC Two to produce significantly fewer syllables, while not exerting influence on those with low levels of WMC Two in producing syllables.

The interaction between WMC Two and Condition on Length Four was not corroborated by the grouping analyses, which found no significant interaction between Condition and Grouping, $F(2, 76) = 2.340, p = .103, \eta^2 = .058$. However, further analyses comparing the subgroups of WMC Two indicated that the number of total syllables produced by L2 writers with high WMC Two significantly dropped compared with that produced by their counterparts at the same WMC Two level of WMC Two who wrote under the NTA condition, $p = .044$. The significant interaction was confirmed.
when comparisons were made between the two conditions at seven different levels of WMC Two percentiles. The difference in Length Four between the two conditions achieved significance at the higher ends of the percentile spectrum (i.e., at .926, \( p = .012 \), at .976, \( p = .008 \), and at .981, \( p = .008 \)). This indicated that TA did differentially affect L2 writers at different WMC Two levels, in terms of Length Four, given that it caused those with high levels of WMC Two to produce significantly fewer total syllables, while not exerting influence on the number of total syllables produced by those with low levels of WMC Two.

The interaction between WMC Two and Condition on \( D \) was not corroborated by the grouping analyses, which found no significant interaction between Condition and Grouping, \( F (2, 77) = 1.416, p = .249, \eta^2 = .035 \). However, further analyses comparing the subgroups of WMC Two indicated that the lexical diversity measure \( D \) shown by L2 writers with high WMC Two significantly dropped compared to that displayed by their counterparts at the same level of WMC Two who wrote under the NTA condition, \( p = .013 \). The significant interaction was confirmed when comparisons were made between the two conditions at seven different levels of WMC Two percentiles. The difference in \( D \) between the two conditions achieved significance at the higher ends of the percentile spectrum (i.e., at .870, \( p = .013 \), at .926, \( p = .002 \), at .976, \( p = .002 \), and at .981, \( p = .002 \)). This indicated that TA did differentially affect L2 writers of different WMC Two levels, in terms of their lexical diversity \( D \), given that it caused those with high levels of WMC Two to significantly resort to repetitive words while not exerting much influence on those with low WMC Two levels in this regard.

In summary, the above interaction effects indicated that TA had differential effects on L2 writers at different levels of several different dimensions of writing performance, controlling for their prior differences in the dimensions of performance and WMC Two.
The differential effects were not obvious in terms of writing speed, given that it only significantly increased Speed Two of those who belonged to the slowest five percent. The interaction was seen in terms of Length Two, where those who previously wrote the most words wrote significantly less, while the others were little affected. The interactions on General Complexity and CVU were different. In terms of General Complexity, under the TA condition, the performance of the highest achievers was significantly boosted in the measure, while that of the lowest achievers declined significantly. In terms of CVU, under the TA condition, previously low CVU performers did significantly worse, while previously high performers were not affected.

TA also had more differential effects in relation to WMC Two than to WMC Average and WMC One. Under the TA condition, L2 writers with high WMC Two performed significantly worse while those with low WMC Two did not show much influence, in terms of Length Three, Length Four, and $D$.

4.4 Summary

Table 4.31 summarizes the effects of TA on L2 writers’ performance in writing a narrative task.

The effects were found when a series of ANCOVAs and concomitant analyses were conducted, first with WMC Average, then with WMC One, and finally with WMC Two, as one of the covariates. These effects were confirmed via a series of ANCOVA tests excluding the WMC measures as a covariate where their correlations with main task data were not significant. The effects were also confirmed via a $t$-test run on the main task data. In summary, controlling for baseline performance and average WMC, or WMC One, or WMC Two, TA had a statistically significant negative effect on three measures: ND, lexical diversity (measured by $D$), and organization, and tended to affect Length Three as well.
Table 4.31

Effects of TA on L2 Writers’ Performance in Writing a Narrative Task

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Direction of effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Time</td>
</tr>
<tr>
<td></td>
<td>Less time</td>
</tr>
<tr>
<td>Speed 1</td>
<td>Slower speed</td>
</tr>
<tr>
<td>Speed 2</td>
<td>Slower speed</td>
</tr>
<tr>
<td>Speed 3</td>
<td>Slower speed</td>
</tr>
<tr>
<td>Speed 4</td>
<td>Slower speed</td>
</tr>
<tr>
<td>ND</td>
<td>More dysfluencies*</td>
</tr>
<tr>
<td>Complexity</td>
<td>Length 1</td>
</tr>
<tr>
<td></td>
<td>Shorter text</td>
</tr>
<tr>
<td>Length 2</td>
<td>Shorter text</td>
</tr>
<tr>
<td>Length 3</td>
<td>Shorter text</td>
</tr>
<tr>
<td>Length 4</td>
<td>Shorter text</td>
</tr>
<tr>
<td>General complexity</td>
<td>Increased general complexity</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>Increased syntactic complexity</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>decreased subclausal complexity</td>
</tr>
<tr>
<td>SV</td>
<td>Decreased syntactic variety</td>
</tr>
<tr>
<td>LD (D)</td>
<td>Decreased lexical diversity*</td>
</tr>
<tr>
<td>Accuracy</td>
<td>CVU</td>
</tr>
<tr>
<td></td>
<td>Decreased accuracy</td>
</tr>
<tr>
<td>EFW</td>
<td>Increased accuracy</td>
</tr>
<tr>
<td>EFCs</td>
<td>Increased accuracy</td>
</tr>
<tr>
<td>Content</td>
<td>Decreased content</td>
</tr>
<tr>
<td>Organization</td>
<td>Decreased organization*</td>
</tr>
</tbody>
</table>

* the effects reached statistical significance.

Using baseline performance scores and WMC Average or WMC One or WMC Two as two covariates, statistically significant interaction between baseline performance and Condition was found on Speed Two, General Complexity, and CVU, and additionally on Length Two only in the case where WMC Two was one of the covariates. Previously high achievers benefited from TA in General Complexity, and previously low achievers suffered from TA in General Complexity and CVU. In a different direction, previously high Length Two L2 writers were significantly less productive in terms of Length Two, but previously low Speed Two L2 writers were lifted in terms of this speed.

As for the interaction between WMC and Condition, a statistically significant interaction was found between WMC Average and Condition and between WMC One and Condition on Organization. In both interactions, low WMC L2 writers performed significantly worse in organizing texts. The interaction between WMC Two and Condition was not found in Organization, but was significant in a range of other aspects.
of performance, namely, Length Three, Length Four, and Lexical Diversity (measured by $D$). In all these cases, high WMC Two L2 writers performed significantly worse.

4.5 Reactivity and Task Effects

The results from the quantitative analyses were then compared with those from Yang et al.’s (2014) study, in which 95 university sophomores who were volunteers recruited from around 20 classes comprising all non-English-major sophomores around the university (approximately 600) completed a baseline argumentative writing task silently and then completed a similar argumentative main task under a silent, a nonmetacognitive think-aloud, or a metacognitive think-aloud condition. The ranges of significant reactive effects (i.e., what aspects of performance were significantly affected) found in the two studies were first compared. The effect sizes were also examined. Finally, a series of ANCOVAs was run to compare the performance data of the two TA groups in the two studies to see if the two TA groups in the two studies performed statistically differently, using their respective corresponding baseline writing performance data as the covariate.

4.5.1 The different ranges of influence

Table 4.32 shows the descriptive statistics and the results of the ANCOVA analyses run on Yang et al.’s data, juxtaposed with the results of pairwise comparisons following the Bonferroni procedure.

In Yang et al.’s study, overall quality was examined, which was scored based on Jacob et al.’s scoring scheme, but the subscores of its components, namely, content (30%), organization (20%), vocabulary (20%), language use (25%), and mechanics (5%), were not examined.

For the sake of comparison, Table 4.32 also presents the descriptive statistics and the results of the ANCOVA analyses run on the subscores of content and organization.
### Descriptive Statistics and the Results of ANCOVAs Run on Yang et al.’s Data

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>M (SD, adjusted M) for 3 think-aloud conditions</th>
<th>ANCOVA</th>
<th>Pairwise comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NTA</td>
<td>NMTA</td>
<td>MTA</td>
</tr>
<tr>
<td>Time</td>
<td>59.226</td>
<td>62.594</td>
<td>84.938</td>
</tr>
<tr>
<td></td>
<td>(13.889, (15.330, (26.996,</td>
<td>(63.856, (62.183, (82.691,</td>
<td></td>
</tr>
<tr>
<td>Speed 3**</td>
<td>12.245</td>
<td>11.153</td>
<td>8.501</td>
</tr>
<tr>
<td></td>
<td>(3.595, (2.626, (2.902,</td>
<td>(11.848, (11.233, (8.606,</td>
<td></td>
</tr>
<tr>
<td>ND</td>
<td>0.069</td>
<td>0.092</td>
<td>0.095</td>
</tr>
<tr>
<td></td>
<td>(0.042, (0.067, (0.048,</td>
<td>(0.065, (0.091, (0.100,</td>
<td></td>
</tr>
<tr>
<td>General complexity</td>
<td>13.223</td>
<td>14.308</td>
<td>13.379</td>
</tr>
<tr>
<td></td>
<td>(2.050, (2.451, (2.468,</td>
<td>(13.417, (14.112, (13.324,</td>
<td></td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>1.975</td>
<td>2.126</td>
<td>1.938</td>
</tr>
<tr>
<td></td>
<td>(0.280, (0.339, (0.340,</td>
<td>(1.995, (2.088, (1.951,</td>
<td></td>
</tr>
<tr>
<td>Subcralual complexity</td>
<td>8.834</td>
<td>8.995</td>
<td>9.084</td>
</tr>
<tr>
<td></td>
<td>(1.288, (1.228, (0.994,</td>
<td>(8.904, (8.999, (9.030,</td>
<td></td>
</tr>
<tr>
<td>SV</td>
<td>12.225</td>
<td>11.562</td>
<td>11.406</td>
</tr>
<tr>
<td></td>
<td>(2.704, (2.154, (2.861,</td>
<td>(12.759, (11.406, (12.283,</td>
<td></td>
</tr>
<tr>
<td>LD (D)</td>
<td>66.125</td>
<td>66.507</td>
<td>66.774</td>
</tr>
<tr>
<td></td>
<td>(15.713, (12.578, (12.898,</td>
<td>(67.520, (66.455, (66.146,</td>
<td></td>
</tr>
<tr>
<td>EFCs</td>
<td>0.551</td>
<td>.590</td>
<td>.581</td>
</tr>
<tr>
<td></td>
<td>(0.114, (0.102, (0.123,</td>
<td>(0.546, (0.573, (0.584,</td>
<td></td>
</tr>
<tr>
<td>CVU</td>
<td>0.857</td>
<td>0.881</td>
<td>.899</td>
</tr>
<tr>
<td></td>
<td>(0.808, (0.807, (0.898,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>20.71</td>
<td>20.688</td>
<td>22.094</td>
</tr>
<tr>
<td></td>
<td>(4.459, (4.185, (4.343,</td>
<td>(20.845, (20.809, (21.464,</td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>14.807</td>
<td>13.781</td>
<td>14.875</td>
</tr>
<tr>
<td></td>
<td>(1.662, (2.166, (1.540,</td>
<td>(14.845, (13.853, (14.715,</td>
<td></td>
</tr>
<tr>
<td>Overall quality</td>
<td>67.581</td>
<td>68.844</td>
<td>72.656</td>
</tr>
<tr>
<td></td>
<td>(9.528, (10.925, (9.093,</td>
<td>(68.887, (69.157, (70.919,</td>
<td></td>
</tr>
</tbody>
</table>

*Note.* NTA: all \(N = 31\), except Dysfluencies, \(N = 30\); NMTA: all \(N = 32\), except Dysfluencies, \(N = 29\); MTA: all \(N = 32\), except Dysfluencies: \(N = 31\). A statistically significant baseline score \(\times\) Condition interaction was found on Speed 3, \(F(2, 89) = 3.338, p = .040, \eta^2_p = .070\). Also, the interaction between baseline dysfluencies and Condition approached statistical significance, \(F(2, 84) = 2.992, p = .056, \eta^2_p = .066\). *\(p < .05\).* **indicates the baseline score \(\times\) Condition interaction reached statistical significance.
Table 4.32 shows that in Yang et al.’s study that used argumentative writing tasks, TA increased dysfluencies significantly, $F(2, 84) = 5.863, p = .004, \eta^2 = .122$, and tended to affect Organization negatively, $F(2, 89) = 3.137, p = .048, \eta^2 = .066$. In comparison, in this study, which used cognitively less demanding narrative tasks with outlines given, the same type of verbal reporting also increased dysfluencies significantly. It also affected organization negatively, but to a significant degree. The most notable difference, however, was that TA significantly affected lexical variety in this study.

4.5.2 The different scales of influence

Table 4.33 contrasts the effect sizes of TA in the two experiments. To make this contrast plausible, the data in Yang et al.’s study were reanalysed just to compare the NTA and the TA groups, using the LSD procedure.

As can be seen from Table 4.33, the sizes of the effects of TA were greater with the argumentative writing task than with the narrative writing task as far as dysfluencies and organization were concerned, where significant effects were found, but the effect sizes for both tasks were between the medium and large levels and were very close ($d = 0.676$ vs. $0.569$, $\eta^2_p = .111$ vs. .069 for Dysfluencies and $d = 0.563$ vs. $0.520$, $\eta^2_p = .077$ vs. .066 for Organization). Additionally, on SV, the effect of TA was statistically significant, had a medium to large effect size, and was much greater for the argumentative writing task while it did not achieve significance for the narrative writing task and was below the small effect size benchmark ($d = 0.586$ vs. $0.099$ and $\eta^2_p = .077$ vs. .002). The exception in this comparison of studies was the effect of TA on lexical variety which was significant for the narrative writing task and had an effect size approaching the medium effect size benchmark, while that for the argumentative writing task was negligible ($d = 0.462$ vs. $0.055$ and $\eta^2_p = .053$ vs. .001).
Table 4.33

**Contrast of the Effect Sizes of TA in Two Studies**

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Narrative writing task</th>
<th>Argumentative writing task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direction of effects</td>
<td>Direction of effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Less time</td>
<td>Less time</td>
</tr>
<tr>
<td>Speed 3</td>
<td>Lower speed</td>
<td>Lower speed</td>
</tr>
<tr>
<td>Dysfluencies</td>
<td>More dysfluencies</td>
<td>More dysfluencies*</td>
</tr>
<tr>
<td>Complexity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>General complexity</td>
<td>Increased complexity</td>
<td>Increased complexity</td>
</tr>
<tr>
<td>Syntactic complexity</td>
<td>Increased complexity</td>
<td>Increased complexity</td>
</tr>
<tr>
<td>Subclausal complexity</td>
<td>Decreased complexity</td>
<td>Increased complexity</td>
</tr>
<tr>
<td>SV</td>
<td>Decreased variety</td>
<td>Decreased variety</td>
</tr>
<tr>
<td>LD (D)</td>
<td>Decreased diversity*</td>
<td>Decreased diversity</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVU</td>
<td>Decreased accuracy</td>
<td>Increased accuracy</td>
</tr>
<tr>
<td>EFCs</td>
<td>Increased accuracy</td>
<td>Increased accuracy</td>
</tr>
<tr>
<td>Content</td>
<td>Worse content</td>
<td>Worse content</td>
</tr>
<tr>
<td>Organization</td>
<td>Worse organization*</td>
<td>Worse organization*</td>
</tr>
</tbody>
</table>

*means the effects reached statistical significance.

For measures where neither experiment reported significant effects, some differences in effect size were found between the two experiments, the general trend being that the effect sizes for the argumentative writing task were greater than those for the narrative writing task. The aspects of writing performance that demonstrated this trend included Speed Three (one being within the small to medium range, and the other being negligible, \(d = 0.299\) vs. \(0.009\) and \(ŋ²_p = .023\) vs. \(.000\)), General Complexity (one being within the small to medium range, and the other being negligible, \(d = 0.372\) vs. \(0.087\) and \(ŋ²_p = .034\) vs. \(.002\)), and Syntactic Complexity (one being within the small to medium range, and the other being negligible, \(d = 0.342\) vs. \(0.032\) and \(ŋ²_p = .029\) vs. \(.000\)). The effect size difference was narrower for EFCs, where, although the effect size was still greater for
the argumentative writing task, both effect sizes were within the small to medium range ($d = 0.325$ vs. $0.246$ and $\eta^2_p = .026$ vs. $.016$).

In contrast, little difference in effect size was found for Time, where the effect sizes reported in both experiments were negligible ($d = 0.083$ vs. $0.136$ and $\eta^2_p = .002$ vs. $.005$, with the effect sizes for the narrative writing task being slightly greater). Not much difference was found for CVU, either, where the effect sizes from both experiments were almost identical and were both negligible ($d = 0.163$ vs. $0.167$ and $\eta^2_p = .007$ vs. $.007$). It should be noted that given the almost equivalent effect sizes, the effect for the narrative writing task on CVU was negative while that for the argumentative writing task was positive.

The effect size for the narrative writing task exceeded that for the argumentative writing task in Subclausal Complexity ($d = 0.332$ vs. $0.093$ and $\eta^2_p = .028$ vs. $.002$), one falling within the small to medium range, with the effect being negative, and the other being negligible, with the effect being very slightly positive. Similar difference was also found for Content ($d = 0.238$ vs. $0.001$ and $\eta^2_p = .014$ vs. $.000$).

4.5.3 Task and Condition interaction

I then conducted a series of ANCOVA tests, using both Condition (TA vs. NTA) and Task (argumentative vs. narrative) as two fixed factors, and the main task performance scores as dependent variables, to examine if there existed any significant Task × Condition interaction. No significant interaction was found, indicating that the effects of Condition on task performance did not vary significantly across the two tasks, for Time, $F (1, 143) = 3.007, p = .085, \eta^2_p = .021$, for Speed 3, $F (1, 140) = 3.315, p = .071, \eta^2_p = .023$, for dysfluencies, $F (1, 141) = 0.327, p = .568, \eta^2_p = .002$, for Subclausal Complexity, $F (1, 144) = 2.118, p = .148, \eta^2_p = .014$, for SV, $F (1, 144) = 0.002, p = .967, \eta^2_p = .000$, for Lexical Variety, $F (1, 144) = 1.944, p = .165, \eta^2_p = .013$, for CVU, $F (1,
144) = 1.832, \( p = .178, \eta^2_p = .013 \), for EFCs, \( F(1, 144) = 2.948, \ p = .088, \eta^2_p = .020 \), for Content, \( F(1, 144) = 0.771, \ p = .381, \eta^2_p = .005 \), and for Organization, \( F(1, 144) = 0.025, \ p = .873, \eta^2_p = .000 \). Significant Task × Condition interaction was found for General Complexity, \( F(1, 144) = 5.182, \ p = .024, \eta^2_p = .035 \), and for Syntactic Complexity, \( F(1, 144) = 4.050, \ p = .046, \eta^2_p = .027 \).

![Figure 4.11 Interaction between Task and Condition for General Complexity](image)

Figure 4.11 Interaction between Task and Condition for General Complexity

The interaction between Task and Condition for General Complexity was illustrated in Figure 4.11. As is shown in Figure 4.11, the effects of TA differed across the two tasks in terms of the direction of effects and the magnitude of difference. For the argumentative task, TA increased general complexity, while for the narrative task, it decreased general complexity. Also, the difference in general complexity was close across tasks under the silent condition, but it became significantly greater under the TA condition, with TA increasing the general complexity of the argumentative essays.
written by the participants in Yang et al.’s (2014) study but decreasing that of the narrative compositions written by the participants in this study.

The interaction between Task and Condition as regards Syntactic Complexity is illustrated in Figure 4.12.

![Figure 4.12 Interaction between Task and Condition for Syntactic Complexity](image)

**Figure 4.12 Interaction between Task and Condition for Syntactic Complexity**

As is shown, TA increased the syntactic complexity of the argumentative essays written by the participants in Yang et al.’s (2014) study but it decreased that of the narrative compositions written by the participants in the current study. Also, although under the silent condition the difference in general complexity between the participants was in favor of those writing the narrative, under the TA condition, it was in favor of those writing the argumentative and the difference in syntactic complexity grew significantly greater.
4.5.4 Summary

In terms of the aspects of performance where statistically significant effects of TA were found, the most noticeable difference between the two experiments lay in lexical variety. In addition, the results of the comparison and contrast of the effect sizes indicated that TA may have caused greater reactive effects for those writing the argumentative essay than for those writing the narrative, on most of the measures examined (i.e., SV, Speed Three, General Complexity, Syntactic Complexity, Dysfluencies, Organization, and EFCs). Conversely, it appeared to have caused greater effects for the participants in this study on only three of the measures examined (Lexical Diversity, Subclausal Complexity and Content), most noticeably in Lexical Diversity (the measure $D$). Further statistical analyses showed that the effects of TA differed significantly across the two tasks for the two batches of participants on two complexity measures (i.e., General Complexity and Syntactic Complexity), both pointing to the greater facilitative effects of TA when it was associated with writing the argumentative essay than its detrimental effects when writing the narrative. The reasons for the differences could be task-specific and may be related to participants as well, which I shall discuss in some detail in the discussion chapter.
CHAPTER 5 REACTIVITY: RESULTS FROM QUALITATIVE ANALYSES OF STUDENT REFLECTIONS

This chapter presents the results of my qualitative analyses of the answers that the 43 participants who thought aloud gave to the third reflection question (i.e., Do you think thinking aloud affected the way you wrote the essays?), including their elaborations. In the examples given below, English translations, which were literal, were offered in brackets, in italics.

5.1 The Extent and Negativity of Effects that Participants Felt

Participants reported different degrees of reactivity they felt. Among the 43 participants who thought aloud, 39 reported that the TA task affected their writing processes.

Twenty-three of them appeared certain, stressing that TA did affect writing or affected writing greatly. Participant 10 wrote rather resolutely:

这个必须有影响，说没影响那绝对是大假话。写作是一个复杂的过程，审题、构思、过渡、遣词造句，仔细揣摩，边写边说出自己脑中所想定然会对思考产生不小的影响，更何况被九秒规则束缚着，没法自由发挥，自在思考。

(This must have had influence. To say there was no influence is absolutely a big lie. Writing is a complex process, involving analysing the task, outlining, transition, diction and sentence construction, and careful thinking. Saying what is in one’s mind while writing definitely affected writing greatly. Besides, there was the 9-second constraint. You couldn’t write to your most freely, you couldn’t think freely.)

Another participant, Participant 90, wrote in a very dramatic way emphasizing the great effects of TA on him/her:
大大大大大地影响我写了！至少没给我机会检查文法错误 泪目… 作文写得太锉了… [不知道“锉”字怎么写…] 把注意力放在说上怎么写…啊…

(Greatly greatly greatly greatly greatly affected my writing! At least no opportunity given to me to check grammar mistakes, tearful eyes … It was too poorly written … [Don’t know how to write the Chinese character for “too poorly”]. Attention was paid to saying, How could I write … ah…)

Among these 23 participants who offered an affirmative answer, only one answered that there were positive effects as well as negative ones, in contrast to the rest of them who all reported negative effects. Participant 99 wrote:

影响是一定的，但不好说是好是坏。

首先，精确表达了我的所想，使该部分不易流失，有利于写长篇。

其次，说本身就不太习惯，写作时心态不如平时，思路受阻。

(That it affected my writing is definite, but it is hard to say whether the effect was good or bad.

Firstly, I was able to accurately express what I thought. It helped me keep my thoughts in mind. It was good for writing long texts.

Then, reporting while writing I was not very used to. My state of mind during writing was not as good as usual. My route of thinking was blocked.)

Among the 39 participants who thought TA affected writing, five of them replied that TA, to a certain extent, affected writing or had some influence. Participant 98 reported:

有一些影响。在描述想法的过程中，因为有边写边说的潜意识（书写过程经常长于 9 秒，为保证录音连贯，只能写什么说什么），所以在构思时，
不能及时将想法转化为文字，而必须先说后写。难免对书写整个过程造成延长时间等影响。

(Had some influence. In the process of describing thoughts, because of the subconscious awareness of writing while speaking (the writing process was often longer than 9 seconds, so to ensure that the recording was continuous, one could only report what he/she wrote). So in planning and writing, one could not convert thoughts into words in a timely way, but had to speak before writing. Unavoidably, this caused effects on the whole writing process such as delay.)

Among the 39 participants who thought TA affected writing, 11 of them stated that TA had some influence but it was not big or that it had slight influence. Participant 17 wrote:

有一定影响。因为整个写作/录音过程中，为了更好地表达自己的思路，我不得不加入大量解释性的内容，也存在一定的“为说而说”的问题。不过，整体上影响并不是很大，因为毕竟说出口自己所想，也不会打断自己的思路，所以倒也没有很大的影响，水平基本发挥出来了。

可惜的就是，第一次做这种录音，所以有些紧张。

(Had certain effects, because in the whole writing/recording process, to better express my own thinking, I had to add a large amount of explanatory content, and to a certain extent, there was the problem of “speaking for speaking’s sake”. However, on the whole the effect was not large, because it was reporting what I thought, and it wouldn’t interrupt my own course of thinking. Therefore, it did not have a large effect. I wrote basically in conformity with my level. The pity was that, because it was the first time for me to do such recording, I was somewhat nervous.)
Among the 43 participants who thought aloud, only four answered no effects.

Participant 25 wrote,

感觉没有影响，因为我觉得个人在写作思考过程中其实就是在脑海中的自
言自语，现在不过是说出来了，因而我觉得没有影响，并且边说边写这一
方式，日常写作的时候，也是这样的，不过就是声音小些或者在心里说罢
了，因而不会有什么影响的，仅就个人而言。

(Feel there was no effect, because I think one’s writing and thinking process is in
fact like speaking to oneself in his brain and now he only articulates it. So I feel
there was no effect. And this way of saying while writing is what one does in daily
writing except that the voice is small or it is saying in one’s heart. So there was
no effect, only personally speaking.)

Participant 55 talked about his/her increasing familiarity:

我感觉边说边写没有太大影响吧，刚开始有些不习惯，一会儿过后就没什
么了，感觉边说边写应该是挺不错的方法，只是需要一个单独的空间，防
止外界干扰。

(I feel reporting while writing had no big effect. At the beginning I was not used
to it, but after a while I felt at ease. I feel it is a fairly good method, but one needs
an exclusive space to prevent exterior interference.)

Participant 102 did not blame TA for his/her inability to write:

其实说的过程不是很影响写的过程，主要写不出来是自己很少用英语，很
基本的词汇，语法结构都忘了。很多句子的遣词造句更是一塌糊涂，所以
我不觉得是说影响了写。

(In fact the process of reporting did not affect the process of writing very much. I
could not write mainly because I seldom used English. I forgot very basic
vocabulary and grammatical structures. For many sentences, the choices of words and sentence constructions were simply a disaster. So I did not feel reporting affected my writing.)

Still another participant, Participant 94, answered that there was no (negative) effect, and he/she wrote on and stated the facilitative effects:

没有，说出来更有利于刺激大脑去想，注意力也更加集中，只是说出来更慢了，一定程度影响上速度，但说出来总体好处大，有利于表达出所想。

(No. Speaking out was good for stimulating the brain to think. The attention was more focused, except that speaking made it slow, and affected the speed to a certain degree. But, on the whole, speaking out had bigger benefits. It was facilitative to expressing thoughts.)

Altogether there were four out of the 43 participants who mentioned the positive effects of TA (Participants 99, 94, 18 and 83). For example, Participant 18 stated that TA was possibly able to stimulate him/her to generate new ideas. Participant 83 stated,

我个人感觉说写作文是一个很不错的形式，能将内心所想及时说出，不致于遗漏掉观点，便于表达情感。

(I personally feel that reporting thoughts while writing a composition is a very good way. One can tell one’s inner thoughts in real time so that he/she will not leave out them. It is good for expressing emotions.)

5.2 The Reasons for the Effects

As for the reasons for the influence of TA, 15 participants mentioned that TA diverted their attention. Their specific statements included “一心不可二用” (One heart cannot be used dually, Participant 22), “如果是自己默写，可能会静下心注重思考内容” (Had I written it silently, I could have settled down to focus on thinking about the
content, Participant 23), “我会有一部分精力分散到说上” (I had part of my energy diverted to saying, Participant 24), “必然有分心的影响” (Definitely has the effect of diverting attention, Participant 32), “有可能太注重说的话会影响思考” (Possibly paying too much attention to saying would affect thinking, Participant 35), “说的时候无法集中精力思考” (When saying, I could not focus my attention on thinking, Participant 40), “说也转移了我的注意力，使我不能专心想语句和词汇” (Saying also diverted my attention, making it hard for me to focus on thinking about sentences and words, Participant 43), “同时动用说和写有点耗费注意力” (Engaging in saying and writing at the same time somewhat consumed attention, Participant 67), “由于不习惯说、写同步，会分一定的精力在说上导致不能全神贯注于写” (Because I was not used to speaking and writing simultaneously, part of my attention was diverted to speaking so that I could not put all my heart into writing, Participant 78) and “觉得精力集中不了” (Felt I couldn’t focus, Participant 100).

Eight participants mentioned that they were not used to TA while writing. Participant 100 stated, “平常都是光写和思考, 写又加上说, 实在不适应” (Usual writing is writing and thinking only. Writing plus speaking I really couldn’t adjust myself to it). Participant 41 wrote, “一般写作的时候是不说话的，这回一说话不太习惯，想法很受影响” (Usually when writing one doesn’t speak. This time I spoke while writing and I was not used to it. My thoughts were greatly affected), but he/she admitted that “后面习惯了以后还好一些” (It became better when I got used to it later). Participant 69, who stated that he/she had never engaged in such an activity, thought the effect of TA was negligible when he/she was used to it as he/she wrote, though he/she felt at the beginning he/she could not adjust himself/herself well to it.
Five participants mentioned their nervousness or change in their state of mind. For example, Participant 31 reported, “因为要求不停地说，就有些慌乱，写的错误率就上升了” (Because I was required to speak continuously, I was somewhat put in a fluster, and my error rate increased). Participant 43 admitted, “在录音过程中因为声音不能断也有些紧张和影响思考” (In the course of recording, because vocalization must be continuous, I was somewhat nervous and this affected thinking). Participant 99 also admitted that his state of mind was disturbed.

Three participants mentioned the difficulty in coordinating concurrent reporting and writing. Participant 18 stated,

有时候说着说着就忘记该怎么写了……说影响了写，但写也有点影响了说，写的时候基本都是写已经想好的也就是说了的，但会对下一步的思考有影响。反过来，说的时候，就有心口不统一的现象。

(Sometimes when I spoke on and on I forgot how I should write ... Speaking affected writing, but writing also affected speaking somehow. When I wrote I basically wrote what I had thought over and reported, but this would affect my next step of thinking. On the contrary, when I spoke, there was the phenomenon of “think one way and speak another”.)

Participant 67 also stated a similar problem, as he/she wrote, “可能写的时候说的不流畅，说的时候写的不佳” (Possibly when I wrote I did not speak fluently; when I spoke, I did not write well).

Participant 100 also complained, “考虑了说就忘了写，考虑了写又忘了要写什么” (When I considered speaking I forgot writing, when I considered writing I forgot what I wanted to write).
Six participants mentioned the pressure that the 9-second constraint or having to speak continuously caused. For example, Participant 98 mentioned, “书写过程经常长于 9 秒,为保证录音连贯只能写什么说什么” (The writing process was often longer than 9 seconds. To ensure that the recording was continuous, I could only report what I wrote). Participant 27 wrote, “心里还装着不能 9 秒不说话，思维都凌乱了” (Besides, I kept it in mind that I could not keep silent for more than 9 seconds. My thinking was in disorder then.)

Three participants mentioned the extra pressure and load of TA. Participant 26 stated, “我一直逼自己一定要说,给自己造成了一定的压力和负担” (I have been forcing myself to speak, which caused a certain level of pressure and load on me). Participant 36 stated, “因为忙得不可开交,所以写得更盲目而以致于随意了” (Because I was overwhelmingly busy, I wrote so blindly that I wrote randomly). Participant 98 mentioned that keeping the task requirement in mind affected him/her, as he/she had “边写边说的潜意识” (the subconscious awareness of writing while speaking).

Four participants mentioned the effects of some external factors. Participant 14 wrote, “说写实验的形式也让自己很容易被外部因素干扰” (The approach of the speaking and writing experiment also made me myself very easily disturbed by external factors). Participant 55 felt he/she needed independent space to prevent external interference. Participant 65 also stated his/her preference for a quiet surrounding. Similarly, participant 77 wrote, “写作需要一个安静的环境才会心静开阔，思考更好些” (Writing requires a quiet surrounding and then you can calm down and think broadly and think better).

Two participants mentioned the greater speed of thinking than of speaking. Participant 27 wrote, “因为想得比说得快很多，要把想法都说出来就已经忘了自己
当初要写什么”（Because thinking was much faster than speaking, when I finished reporting all my thoughts, I had already forgotten what I had intended to write at the start）。Participant 92 also wrote, “我觉得思维比想法快，如果我把每个想法写出来必然会影响我的思考速度以致影响写作进度，换句话说就是，也许我有一个很好的想法或遣词造句接下来，但很可能就在说中遗忘了”（I feel thinking was faster than ideas. If I wrote every idea down, this would definitely affect the speed of my thinking and consequently the progress of my writing. In other words, perhaps I had got a very good idea or sentence or word, but I was very likely to forget it in speaking）。

Similarly, Participant 75 wrote that thinking and language were hard to synchronize.

Participant 35 analysed what was required to do in conducting TA. He/She wrote, “很多时候你要想办法表达（说）出来，你还要发音，读、说出来，但直接写的话就不用了”（Many times you would manage to express (articulate) it, you had to pronounce it, read it, and tell it aloud, but you don’t need this when you write it down directly）.

Two participants mentioned voice interference. Participant 65 commented, “这个过程不仅用口，用脑，还要多加一个感知—声音，这干扰了我的思绪”（This process not only used the mouth, and the brain, but also one extra sense—voice, which disturbed my thinking）。Likewise, participant 51 mentioned that he/she had to pay attention to his/her voice level and the microphone, which distracted his/her attention.

One participant, Participant 31, mentioned her tiredness, as she wrote, “边说边写有点累，到后面就更累了，对写的影响还是挺大的”（Speaking while writing was a bit tiring, and became more tiring in the latter part of writing, which had a fairly big effect）。

5.3 The Aspects of Writing that were Affected

The 43 participants mentioned various aspects of writing that TA affected.
Twenty-one participants mentioned the effects on their thinking, in their words, “思考空间” (the space for thinking), “构思” (conceptions), “整体思考” (holistic thinking), “下一步的思考” (the next step of thinking), “思路” (the course of thinking), “忘记” (forgetting), and “思维深度” (the depth of thinking). Participant 23, for example, reflected that “说的话会有一个想法到表达的过程，可能会打断思路” (verbalization entailed a process from thinking to expression, which would probably interrupt the train of thought). Participant 35 also felt that the way that he/she thought with his/her brain seemed different. Participant 41 admitted, “想法很受影响，前半部分在说的过程中脑子基本不转，都不知道该写些什么，后面习惯了以后还好一些。” (My thinking was very much affected. In the first half, during speaking my brain basically did not run, and I did not know what I should write. Later when I got used to it, it turned a bit better.)

Participant 53 recognized the forgetting effect, as he/she wrote, “很多东西已经想到，但顾及到要说出，在说的时候，不知不觉就忘记了，自己一个人想的时候，还能想得更深入，说反而有种坏节奏，没想到太深就说出，然后写了。” (Many things you already thought of, but you needed to verbalize them. When you verbalized them, you forgot them subconsciously. When thinking alone, you could think more deeply. Speaking indeed had a bad rhythm. You did not think deeply before you told it and then wrote it down.)

Participant 62 reported the constraint of TA on the quality of thinking, as he/she wrote,
Participant 65 similarly remarked on the constraint that TA put on his/her thinking. He/She wrote “而且因为一边说一边写很干扰写作，我也不能全身心写作，有的想法被蒙蔽了” (Because speaking while writing disturbed writing very much, I could not put all my heart to writing and some of my thoughts were concealed).

Participant 77 also reflected, 说会有时思维停滞，无法写下一句，或是等时间较久，即使想了很多，但感觉都不太好，迟迟未下笔。（When speaking, sometimes I got my thinking at a standstill. I could not write the next sentence. Or I had to wait for a relatively long interval. Even though I thought a lot, I did not feel good and was hesitant to write.)

Participant 79 mentioned the effects of speaking on specific conception, and complained that “无法整体思考，思路容易打断” (I could not think holistically and the course of thinking was easily interrupted) and “表达出来的具体内容无法与安静写作时精心组织的内容相比” (The specific content that I expressed could not compare with the content I carefully organized when I wrote quietly). Participant 83 also reported, “写的时候大脑思考又要说出来，难免有一些地方想的不全面…… 精彩的地方没有想出多少，要一直说话，就不能静下心去思考这些” (When I wrote, I used my brain to think and I had to verbalize my thoughts. Unavoidably, I couldn’t think...
thoroughly in some places … I did not think of many wonderful points. I had to speak on, so I could not calm down and think like that). Likewise, participant 89 found that he/she did not think sufficiently. Participant 99 also felt he was not as “witty” as before and his very good thinking often vanished.

Word choice was the most frequently mentioned specific aspect of writing performance since 14 participants reported that they did not engage in as much search for better words as they had silently done. Participant 22 stated that he/she was not able to use better substitutes. Participant 39 complained that speaking greatly affected searching for appropriate words because the human brain needed time for searching. Participant 49, similarly, reported that speaking during writing affected the time he/she would have used for thinking about more difficult words and collocations as well as complex sentences. Participant 56 wrote, “写的时候有时会想出多种表达方式，从中选择最好的，边说边写的话就会跳过这样的环节(有时)” (In writing sometimes you think of many ways of expression and choose the best from them, but in the speaking and writing approach, you would skip over such procedure (sometimes)). Similarly, Participant 57 reported, “说话会影响思考，导致文中很多简单词汇出现了很多次” (Speaking would affect thinking, causing many simple words to appear many times in the composition). Participant 65 also admitted, “在写作中，只是想起最简单最常用的词，没有新意” (In such writing, I only thought of the simplest and most common words. There was no novelty). Participant 67 also pointed out the relative difficulty in substituting “good words”. The following is Participant 83’s account:

有点影响词汇的选择。当口说手写的时候，有点快，不能充分斟酌词汇的选择，在词汇运用上有所欠缺。
(Somewhat affected choice of words. When speaking with the mouth and writing by hand, it was a bit fast, and you couldn’t fully weigh and consider choice of words. There was a certain level of deficiency in use of words.)

Also related to the effect on lexis was a forgetting effect. For example, Participant 57 mentioned that due to TA, words he/she had been normally familiar with escaped him/her for a moment. Participant 62 exemplified this effect, as he/she wrote,

比如“想，希望”要换成英文的话头脑中一下子有很多想法，说出一种似乎把别的就忘记了。

(For example, if you want to express “want” or “hope” in English, momentarily you have a lot of ideas in mind, but to tell one aloud you seem to forget others.)

Participant 79 had another elaboration of this effect,

比如，在往常写作中，大脑会闪现出许多备用，可用的词汇，但由于需要说，则大脑还须分配精力于说出其中的词汇，那末有一些一闪而过的就会忘记。

(For example, in usual writing, many spare, usable words flash into one’s brain, but because I had to speak, my brain had to allocate attention to telling the words out, and then some that flashed through my brain would be forgotten.)

Ten participants mentioned the effects of TA on carefully considering syntax. Participant 51, for example, reported that,

当自己在说的时候，嘴巴中在说，脑子中反映不出下面要描述的句子，所以更多的写的是一些简单句，少了很多复杂句或者连句等等。

(When I myself spoke, my mouth was moving, but my brain could not reflect the next sentence to be written. So in more cases I wrote some simple sentences, but much fewer complex or connected sentences.)
Participant 65 also admitted that his/her sentence structures became very simple. Participant 70 recognized simpler syntax too, referring to the forgetting effect, as he/she wrote, “比如一些句子一说出来就会忘了怎么样才能使它更高级” (For example, some sentences once you said aloud, you forgot how to upgrade it).

Seven participants mentioned the effect of TA on the writing speed. Participant 57 remarked, “如果没有说话，应该会写得更快” (If I had not spoken, I would have written faster). Participant 90 answered,

录音的时候思维几乎处于停滞状态，这是因为本人表达能力实在差，口语的时钟频率和思考的时钟频率不知道差几个数量级……为了说专门调低了思考的时钟频率。

(When I did the recording my thinking was almost at a standstill. This was because my ability to express myself was really poor. I did not know how many orders of magnitude of difference there were between my oral clock frequency and that of my thinking. To speak, I especially turned down the clock frequency of my thinking.)

Participant 94 also stated, “说出来更慢了，一定程度影响上速度” (Telling it aloud makes writing slow; it affected the speed of writing to a certain extent).

Still three participants mentioned the effects of TA on errors. For example, Participant 90 mentioned that he/she had no opportunity to check grammar mistakes.

Another four participants mentioned the effects of TA on organization related aspects of performance. Participant 24 remarked,

我在平常的写作中没有自语的习惯，所以让我写的时候我会有一部分的精力分散到说上，使我在写的时候，总体的安排受了影响。
(I don’t have the habit of speaking to myself in normal writing. So when I spoke, I allocated a part of my attention to speaking, which affected the overall arrangement when I wrote.)

Participant 70 also replied, “比如一些句子一说出来就会 … 忘了前后连贯，逻辑也不太顺畅” (For example, some sentences once you told aloud … you forgot coherence, and the logic was not smooth as well).

One participant, Participant 49, mentioned that his/her style of writing seemed to have been changed to colloquialism due to TA, as he/she wrote,

我感觉写得像讲口语, 讲流水账, 不像一篇作文了, 而且说的时候一些如 “然后，然后” 会反映在作文中写了“and，and”。

(I feel like writing in an oral style, or keeping a running account, rather than writing a composition. Besides, when I said things like “and, and”, they would find themselves in my composition as “and, and”.)

Still another participant, Participant 65, admitted a change in his/her emotions—“感情不够真挚” (my feelings were not genuine).

5.4 Summary

The participants’ responses to the reactivity question reflected that 1) most of them felt their writing performance and/or writing processes had been affected, 2) far more participants thought the effects of TA were negative than those who viewed it positively, 3) the cognitive load may be the primary reason for the reactivity, among a multitude of reasons expressed by the participants in their reflections, which covered the dual demand for attention or dual processing, unfamiliarity with the task, nervousness and pressure, difficulty in coordinating concurrent reporting and writing, voice interference, and tiredness, and 4) TA reportedly may have affected a wide range of aspects of their writing, especially their thinking processes and word choice. It should be noted that they
may have been subjective (and have exaggerated) in their answers. When participants modified the effects of TA with the Chinese character “很” (literally, “very” or “very much”), it may not mean “greatly” really. That is because of a Chinese habit of using the adverb. Indeed, some participants, though mentioning that TA affected them very much, did not actually state many effects in their following elaborations. The participants’ accounts, though informative, should be viewed with some caution.
CHAPTER 6 VERIDICALITY: RESULTS FROM COMPARISON OF REPORTS

This chapter presents the results from my analyses of CVRs, in comparison to RVRs, regarding the thoughts or thinking processes that occurred at points of revision.

Altogether, the 38 think-aloud participants whose CVRs and RVRs were valid identified and reported on a total of 517 instances of revision retrospectively, averaging 13.61 individually. Three types of incomplete CVRs and two types of complete CVRs in relation to the points of revision were categorized in my reiterative comparisons of the CVRs and RVRs: 1) lack of report on what one had intended to write, 2) lack of report on what one actually wrote, 3) lack of report on metacognitive thinking, 4) complete CVRs for revisions that may not involve much metacognitive thinking, and 5) complete CVRs for revisions that involved metacognitive thinking. In what follows, I shall illustrate these types of incomplete and complete CVRs.

In all the examples given below, to indicate a participant’s revisions in a composition segment, insertions are shown in bold type and deletions are shown with strikethroughs; in all the verbal report excerpts, the words in italics are the literal translations of a participant’s original think-alouds in Chinese and those in normal font represent a participant’s original think-alouds in English. Pauses in seconds are indicated in parentheses and complementary information or my comments are given in square brackets. Transcripts directly related to revisions were underlined in the CVR excerpts.

6.1 Lack of Report on What One had Intended to Write

Incomplete reporting of what one had intended to write was identified in the case of a false start where one ceased to report an idea and started with another idea but failed to complete the original idea. The relevant revision occurred when one felt that he/she
was unable to express the original idea competently or simply because he/she found a better idea.

Example 1 (Participant 10):

As a result, he got bad scores in the final test. After that, he recognized that it was a dangerous signal. If he couldn’t get a good score in the test, he wasn’t able to graduate from the school. This was a big problem. It meant So he began to study hard and finally got a good score in the test.

His CVRs concerning the revision:

这相当于, (3s) 它意味着, It means, means, mean, (8s) [inhalation] 唉, 不行。

*This was equal to, (3s) It meant, It means, means, mean, (8s) [inhalation] oh, not OK.*

His RVRs concerning the revision:

我还是想写“意味着他美好的出路……”，然后发现这句我不会写，就换了一句。

*I had wanted to write “meant his marvelous future... [would be destroyed]”, but I found I could not write on, and then I wrote another sentence instead.*

The CVR segment corresponding to the revision was coded as a case for incomplete CVR because it did not contain any information on what he/she had intended to write on (i.e., *his marvelous future*, according to his/her RVRs), which, it could be inferred, he/she was conscious of at the time of CVR and therefore such should have been reported concurrently. The lack of reporting was related to what had happened in the 8-second pause in his concurrent report, between “It means, means, mean” and the inhalation.

Example 2 (Participant 43)

Revision:
Now, he works in of course, he is a lawyer, and he does him work well.

His CVRs concerning the revision:

[Read the Chinese instructions] Now, [inhalation] now he works, now he: works in, [make clicks] 不好说呀, now he is a lawyer, now, now, of course, c-o-u-r-s-e, he is a lawyer.

[Read the Chinese instructions] Now, [inhalation] now he works, now he: works in, [make clicks] Not easy to express it, now he is a lawyer, now, now, of course, c-o-u-r-s-e, he is a lawyer.

His RVRs concerning the revision:

这儿本来想写他在律师事务所工作，但是，但是律师事务所这个词不会写，所以还是写了 he is a lawyer.

Here I had wanted to write he works in a law firm, but, but I did not know the English word for “law firm”, so I wrote he is a lawyer.

Participant 43 started to write about his father’s present in English immediately after he referred to the instructions for the second paragraph. In his CVR excerpt, he apparently failed to report what he had planned very quickly in his mind prior to his utterance “Not easy to express it”. He may have planned while, or shortly after, he read the instructions, and he still kept it in his mind while he wrote.

Only two cases of this kind were found although there were quite a number of false starts, because the participants appeared to be able to report what they had intended to write almost exactly in the overwhelming majority of cases. It should be noted that such incomplete reporting was identified by reference to the RVRs for what one had intended to write. It is possible that there were missing cases in the RVRs due to memory decay in recalling.
6.2 Lack of Report on the Revised and the Revision

This type of incompleteness was associated with additive or substitutive revisions when the participant did not mention the word that he/she added or used as a substitute or the word that he/she replaced.

Example 1 (Participant 17)
Revision:
In the my childhood, I was deeply influenced by my father in many ways …

His CVRs concerning the revision:
I I am deeply influ, 就是在我的成长过程中, through my childhood, in my childhood, 在我的童年中, in my childhood, 哦，我在各方面都深深地受到了我的父亲的影响, I am deeply influenced, deeply influenced, influenced by my father, by my father, on many, on many, 哎呀，天哪，many ways, 在很多方面我说不出来那我就说从很多种方式吧，那我就说 in many ways, in many ways, [inhale] 我在各个方面都，哦，都受深深地受到了，都深深地受到了我父亲的影响……
I I am deeply influ, in the process of my growing up, through my childhood, in my childhood, in my childhood, in my childhood, oh, in every aspect I was deeply influenced by my father, I am deeply influenced, deeply influenced, influenced by my father, by my father, on many, on many, ah, my god, many ways, in many aspects I can’t express, I will say in many ways, I will say in many ways, in many ways, [inhale] I in every aspect, oh, was deeply influenced, deeply influenced by my father …

His RVRs concerning the revision:
可能因为是想突出我和我父亲这种关系吧，改成了 my.
Maybe because I wanted to emphasize such a relationship between my father and me, I changed it to my.

Throughout Participant 17’s concurrent verbal report, he did not mention the word the, which he revised to my. That he had put the down was likely a result of a slip of his pen as he spoke rather fluently during his reporting, which indicated that there might not be room for the metacognitive thinking he reported retrospectively. In other words, his RVRs should be considered with some caution.

Example 2 (Participant 57)

Revision:

My father was a person who takes highly responsibility for his work, so he has no spare time to relax himself.

Her CVRs concerning the revision:

嗯，my father，take，have，对，我的父亲是一个对工作很负责的人，所以他的业余生活很单调无聊，我希望他以后怎么样。My father was a, My father was, 嗯, My father was, My father was a person, who take, who takes, who takes, 对工作高度负责, res-pon-sibility, responsibility, who takes highly responsibility for, his work, 嗯, so he had so he had no so he has no spare time, 哦, to, so he has no spare time to, relax, 哦，然后，以后我希望他的生活能够，在以后的时间里……

Oh, my father, take, have, right, my father is a person who is very responsible for his work, so his spare time is very monotonous and boring, I hope later he will do this and that. My father was a, My father was, oh, My father was, My father was a person, who take, who takes, who takes, highly responsible for his work, res-pon-sibility, responsibility, who takes highly responsibility for, his work, oh,
so he had so he had no so he has no spare time, oh, to, so he has no spare time to.

relax, oh, then, later I hope his life can, in future …

His RVRs concerning the revision:

relax 后面加了个 himself，这样感觉……结束得不是那么突然

I added himself right after relax; in this way I did not feel … the ending was that abrupt.

In completing this sentence, Participant 57 made two additive revisions, time and himself. By listening to her rather smooth and continuous verbalization, one could hardly notice when she did the revisions. I could only infer that given her mentioning of the word time, the word time was added during the writing course when she noticed a discrepancy between what she verbalized (i.e., has no spare time to) and what she wrote (i.e., has no spare to), most probably after she finished writing the latter. Himself was not mentioned but was added when the original sentence was finished, between the word relax and the full stop. This meant that she may have done the addition about a second after she did an immediate scan of what she had just finished writing (i.e., to relax.). This swiftness in revision was related to the fact that she may have acquired the collocation relax oneself formulaically.

Example 3 (Participant 70)

Revision:

He said, “whe Every time I went go out, you just play with it, and remember, I will return soon.

Her CVRs concerning the revision:

He said, [inhalation], (5s) every time I went back, 不是, every time, I went out, I go out, you, just, play, with, it, and remember, I will return, soon.
He said, [inhalation], (5s) every time I went back, no, every time, I went out, I go
out, you, just, play, with, it, and remember, I will return, soon.

Her RVRs concerning the revision:

I had wanted to write when he went out blah blah, ah, later I added every time.
This way, it looked, [making clicks], that is, more ... How to put it? A little bit
gooder.

The participant forgot to report what he actually had written, namely, whe, which he
revised, in his concurrent report. The revision happened during the five-second interval
after the inhalation.

Altogether three cases were identified under this category. It is understandable that
when thinking aloud, the participants sometimes may have forgotten to mention one or
two words that were involved in revision. I did not include cases that involved spelling
changes here, where the participants might just make the revisions without reporting the
original or renewed ways of spelling. With spelling corrections included, there should
be more cases when participants missed reporting what they did in revising.

6.3 Lack of Report on Metacognitive Thinking

Lack of reporting on metacognitive thinking concerned 259 out of the 517 cases of
revision reported, where the participants did not report their metacognitive thinking that
underpinned their revising behaviors.

Example 1 (Participant 10):

The revision:

Compared with the past, he is not as strict with me as the past.

His CVRs concerning the revision:
Well, next, (6s) next will be, compare with the past, compare with the past, in terms of comparing with the past, what changes to write about? Let me think, (8s) a little older, (5s) compared with past, (9s) compare with the past, (4s) [making clicks] (3s) well, grey hair, (4s) appears on, (5s) alas, how to say grey hair in English? He had white hairs? [inaudible], oh, forget it, compare with the past, what other aspects? (5s) Let me think, compare with the past, [couch], changes, (9s), oh, I got it, not as strict. He wa- (2s) he is, not, as, strict, with me, as the past ……

His RVRs concerning the revision:

The last but one paragraph, a slip of the pen, I wrote he was as … he is as he was.

This was coded as a case for incomplete CVR and one that indicated a failure to report metacognitive thinking, because the participant arguably should have had thought of the reason for this revision, something like “the past tense was wrong; this describes the present”, which may have flashed across his mind and escaped from his consciousness too quickly for reporting. Specifically, the participant failed to report his thoughts that occurred to him in the two-second pause that happened between wa- and
he is. This case also indicated that the participant’s retrospective verbal report might need further analysis since the revision was, in fact, not merely “a slip of the pen”.

Example 2 (Participant 14):

The revision made:

But Nevertheless, I know his request and anticipation on me has never changed.

His CVRs concerning the revision:

But I know, (5s) but I know, (3s) but 不能这样用, 但 好像不能直接接-
nevertheless, nevertheless, nevertheless, I know his, request on me, I know his, his, requirement, requirement, his request on me, 对我的要求，怎么说呢？his requirements, require 是需要的意思, request, his request, on me, his request, (2s) 和期待, his anticipation, his request and, an-ti-ci-pation, on on me has never changed, has never, changed.

But I know, (5s) but I know, (3s) but cannot be used this way, but may not be used
to directly connect- nevertheless, nevertheless, nevertheless, I know his, request on me, I know his, his, requirement, requirement, his request on me, requirement for me, how should I put it? his requirements, require means need, request, his request, on me, his request, (2s) and anticipation, his anticipation, his request and, an-ti-ci-pation, on on me has never changed, has never, changed.

His RVRs concerning the revision:

因为父亲对于我来说除了要求来说，要求这个词它比较硬一点，比较 ……
但是还是希望用一个比较柔的词，更柔一点的词，更带有父爱的词，所以就想添一个 anticipation.
Because my father to me besides request, the word request is relatively a little bit hard, relatively ... but I hoped to use a relatively soft word, a softer word, a word that carried the love of a father, so I wanted to add anticipation.

Participant 14 added the word anticipation when he finished writing request on me. However, there was no reporting of the metacognitive thinking underlying this behaviour—he wanted to use anticipation to enrich his father's love for him, as revealed by his RVRs. Noticeably, the metacognitive thinking underlying the first revision substituting Nevertheless for But was reported.

Example 3 (Participant 17):

So, he successfully came graduated from high school and went into university. During the time in the university, my father worked studied hard, so after he graduated, he had got the chance to work in the province hospital.

His CVRs concerning the revision:

so, he 他成功地考入了大学 he successfully successfully came into, came into university [exhalation] ↓graduate↓ he he successfully graduated from high school graduated, from high school [making clicks] 哎呀，就这高考不会说，所以要不然就不用这么麻烦了，and (4s) ca- went into, into university. (3s) 嗯，at the time, 然后，during the time, 就是在大学学习的时候, university, (3s) [inhalation] my father, (2s) [exhalation] (2s)↓worked (2s) salary, 知道吗? ↓ my fa:: 就是我, 我爸很努力，很努力学习, (2s) 学习很努力, 然后, so 所以在 他毕业之后, after he, ↓after he↓ graduate, ↓g-r-a, provin↓ (3s) 他, 他毕业后有机会进到省人省医院，但是省医院我不知道怎么说呀。So after he graduated he has, he has a chance, he has, he had the chance, (2s) got,改成 got
吧，他获得了。

So, he was successfully admitted to university, he successfully

came into, came into university (2s) [exhalation] ↓graduate↓ he he successfully
graduated from high school, from high school [making clicks] alas, it
is this gaokao [College Entrance Examination] I can’t say, otherwise it wouldn’t
be so troublesome, and (4s) ca- went into, into university. (3s) oh, at the time, then,
during the time, at the time of university, university, (3s) [inhale] my father,
(2s) [exhalation] (2s)↓worked (2s) salary, know it↓↓ my fa:: Then my, my father
was hardworking, studied very hard, (2s) studied very hard, then, so so after he

graduated, after he, ↓after he↓ graduate, ↓g-r-a, provin↓ (3s) he, he had a chance
to enter the provincial peo- provincial hospital, but I don’t know how to say the

provincial hospital. So after he graduated he has, he has a chance, he has, he had
the chance, (2s) got, change to got ah. He got a, to work, in the, pro:: Alas, don’t
know how to say the provincial hospital. Make it province hospital then, province
hospital. Alas, my god, too sad, too wretched.

His RVRs concerning the revision:

我本来想“他成功地进入了大学”写的 he successfully came into the university,

后来我想 came 不太对，想把 came 改成 went，但是后来我就想想，我把“从

高中毕业”也加上了，就改成 he successfully graduated ... university，就把句

式也改了。[下一个修改] 我本来想说“我父亲学习很刻苦”，但是，但是一

开始写成 my father worked hard，然后后来想了一下，想到我爸爸这会儿还
没有工作，所以我斟酌一下，把它改成了 study。[下一个修改] 我原本写的
had，但是后来一想，had 这个，主动性太强，毕竟是因为你是去就业的嘛，
就是主动性要差一点，所以我就改成了 he got the chance，这样就显得就是
可能是因为我爸爸优秀所以被选中了而不是就是去挑一样的。

I had thought to write he successfully came into the university for “he successfully
erentered university”, but then I did not think came was quite right, and wanted to
change came to went, but later I had another thought and added “graduated from
high school”, and then I changed it into he successfully graduated … university,
and changed the sentence pattern. [The next revision] I wanted to express “my
father studied very hard”, but, but at first I wrote my father worked hard. Then I
thought about it and I though of the fact that my dad did not work at this time. So
I weighed it a little and change it into study. [The next revision] I had written had,
but later when I thought about it, the word had, had too much initiative. After all,
it was that you went out to be employed. That is, the initiative was less. So I
changed it into he got the chance. Then, it appeared not so much that my dad was
there to choose a job as that he was selected because he was excellent.

The participant’s RVRs suggested that the three revisions he made (from came to
graduated, from worked to studied, and from had to got) all involved metacognitive
thinking that was not reported in his CVRs. The revisions, notably, all happened very
swiftly within several seconds, as shown in the transcripts above.

Example 4 (Participant 23):

The revision:

Maybe it’s impossible for him to create a get bigger achievement in the future.

Her CVRs:

Maybe, it’s impossible for him to create, a, a big, erm, (6s) big, erm, 不要写
create, to, (4s) get a:: big, to get bigger, achievement, (5s) achievement in the
future……
Maybe, it’s impossible for him to create a big, erm, big, should not write create, to, get a bigger, achievement in the future……

Her RVRs:

我刚开始没有考虑好到底要该具体要写什么，然后就写错一下，然后改了。

At the start I did not think over what on earth to be written specifically. Then I wrote it wrongly, and then corrected it.

In this excerpt, the participant failed to report what underlay her seemly intuitive revisions. When she reported that she should not have written the word create and when she substituted get for it, she must have engaged in some sort of thinking based on which she made her revision decision. She did not report the thinking, though. Her RVRs did not provide any specific information either but seemed to indicate that the revision was simply due to a slip of the pen.

Example 5 (Participant 56):

The revision:

… b for the reason that the facts of the world influence.

Her CVRs:

… because, for, for the reason, that, the facts, 现实发生的变化, the facts, 现在的, 现在的世界, the facts of, 现在, the facts of the world, 事实, 事实, 然后, 影响了, 影响人的看法。 Influence, 影响, 在一些方面, 一些方式的影响, the facts influence, influence, 哦, 的看法, (5s), 直接就这样, influence.

… because, for, for the reason, that, changes in reality, the facts, present, the present world, the facts of, present, the facts of the world, facts, facts, then, influence, influence people’s views. Influence, influence, in some aspects,
influence in some ways, the facts influence, influence, oh, views, (5s), like this, influence.

His RVRs:

I had thought to use because, but then I found I needed to change it because it had been used too much, and then I used for the reason.

The participant’s CVRs show that he made the revision to for the reason immediately after he wrote only a letter b for because. Apparently, there was no concurrent verbal report on the thinking behind this decision, which should have happened in the very transient interval between the moment when he uttered the word because and the moment when he uttered its substitute for the reason.

Example 6 (Participant 92):

The revision:

But my father wasn’t always the value person in family, not only because of the poor life and less money but also the seventh boy child in the family.

Her CVRs:

But, my father, my father, but but my father, was always, 应该是最不受重视的, was always the, 重视, 想一想, 重视就是, 受到最多重视, (5s) [inhalation], 重视, was, always the, (3s) was not always, wasn’t always the, 嗯, value, value person, the value person, in family. 为什么这么说呢, 应该是, 嗯, (3s), not only because, 嗯, because of, the, because of poor, 嗯, not only because of 嗯, the poor life, not only because of the poor life and, less money, but, my father wasn’t always the value person in family, not only because of the poor life and less, but also, (4s), 哦, but also, 嗯, 还是因为他是, 家里面的,
第七个孩子。**But also, the, seventh, boy, (1s) the seventh child, [inhalation], 嗯，**
in the family, 啊，嗯。

But, my father, my father, but but my father, was always, *should be the least cherished*, was always the, *cherish, let me think, cherish is, receive the most cherishing*, (5s) [inhalation], *cherish*, was, always the, (3s) was not always, wasn’t always the, *erm, value, value person, the value person, in family*. Then why?

*Should be, erm, (3s), not only because, erm, because of, the, because of poor, erm, not only because of *erm*, the poor life, not only because of the poor life and, less money, but, my father wasn’t always the value person in family, not only because of the poor life and less, but also, (4s), oh, but also, *erm, also because he is, the seventh child, in the family*. But also, the, seventh, boy, (1s) the seventh child, [inhalation], *erm*, in the family, *ah, erm*.

Her RVRs:

The first revision:

刚开始写的是 was, 后来, 后来发现, 后来发现, 然后, 只写了是有价值的, 忘记 unvalue 了, 然后就在前面加否定号。

*At the start I wrote was. Later, later I found, later I found, then I found I wrote only the value side and forgot unvalue. So I put the negative expression in the front.*

The second revision:

把 not only 和 because 换了一下。这是因为那个……应该就是 because 在前面，我也不知道当时怎么想的，就写错了。

*I exchanged not only and because. This is because that ... should be in front of because. I don’t know what I thought then and why I wrote wrongly.*
The third revision:

爸爸不是第七个男孩儿，他就是第七个孩子，就用错了。

My father was not the seventh boy, he was the seventh child, I used the word wrongly.

The first revision did not involve much metacognitive thinking but embodied an alteration in planning, which was reported concurrently, as the excerpt shows. For the second, however, the participant should have thought of the parallelism rule entailed in the not only ... but also ... construction, hence her revision. The thinking may have happened after she finished a re-reading of what she had written, during the four-second pause towards the end of the report segment, when she was going to write on with but also. For her change from boy to child, some metacognitive thinking should have happened, within the one-second interval, between her utterance of boy and that of the seventh child. It happened so swiftly that it escaped the participant’s CVRs.

Overall, the above examples show that the participants, when thinking aloud, did not report their metacognitive thinking completely. This may be because they did the associated revisions rather automatically or spontaneously, and the metacognitive knowledge underlying these revisions had become implicit (or fossilized) for them.

6.4 Complete Concurrent Verbal Reports for Revisions not Involving Perceivable Metacognitive Thinking

This category of CVRs involved revisions made on spellings, capitalizations, collocations, or slip-of-the-pen errors that hardly entailed any perceivable metacognitive/metalinguistic thinking. It may also involve certain cases where participants abandoned certain expressions simply because they could not write them or where participants thought of some additional messages and then simply inserted them. In such CVRs, participants made mention of the end products of their thinking (e.g., correct spellings as substitutes) and/or their decisions in their CVRs. This category of
CVRs differed from the previous category in that, arguably (or, inferrably), they did not involve (much) accountable metacognitive thinking or metalinguistic knowledge that should have been reported. These CVRs were therefore thought to be complete. Altogether, 219 cases were found for this category.

Example 1 (Participant 62):

Luckily, my mother and her sisters and brother, my uncle and aunts were really hardworking and did well in their study, especially my mother.

Her CVRs concerning the revision:

Then write about her, her growth, oh, her schooling, oh, luckily, luckily, luckily, [inhalation] my mother, she, she, my mother and her brother and sisters were all hardworking. Really hard, very hardworking, [inhalation], and in terms of study, very good, and oh, did well in their their study, especially my mom, especially, especially my mother, mother.
Her RVRs concerning the revision:

她的兄弟姐妹就是我的，就是我的 uncle and aunts。

*Her brother and sisters, are my uncle and aunts.*

In this revision, Participant 62 substituted *my mother, my uncle and aunts* for *my mother and her sisters and b*. She first wrote the latter, then crossed out the first *and*, added a comma above, and then crossed out *her sisters and b*, and wrote *my uncle and aunts* above. Her revision concerning the substitution, which embodied an idea reformulation, was made almost intuitively (i.e., *write my, not her*) and was implemented rather swiftly. There was no inferable metacognitive thinking involved concerning this substitution. The report was therefore considered to be complete.

However, that she crossed out the first *and* and used a comma instead was related to some reportable metalinguistic knowledge, which concerned the use of the conjunction *and* and that of the comma in concatenating three nouns. She should have thought of this and should have reported it, but, apparently, this knowledge, judged from her swift pauses in the excerpt, might be rather efficiently retrieved so that it escaped her concurrent reports.

Example 2 (Participant 31):

The revisions made:

And *in* in the future, *I don’t* actually I don’t know the thought of my father about his future.

Her CVRs concerning the revision:

将来的话，嗯，他将来的生活。嗯，嗯。And in the future, 又用 and 不好，
in 还大写了，in the future, I think, I, actually, actually I don’t know, 哦，实际上，
实际上，I don’t know, 实际上我不知道，我爸爸，我爸爸对未来的想法，I
don’t know the thought of my father about his future.
As for future, well, his future life, well, well, And in the future, using and repeatedly is not good. In was capitalized, in the future, I think, I, actually, actually I don’t know, well, actually, actually, I don’t know, actually I don’t know, how my father, my father, think about his future, I don’t know the thought of my father about his future.

Her RVRs concerning the revision:

This In 拉掉是因为人家不用大写我给大写了, 划掉 I don, 是因为我加一个 actually。

I deleted this In because no capitalization was needed but I capitalized it. I deleted I don because I wanted to add actually.

The second revision (crossing out I don) was related to a slip-of-the-pen error because she kept verbalizing what she would write (i.e., actually I don’t know) but apparently made a false start (i.e., I don) in haste. Her reporting concerning this revision was considered to be complete because there was negligible inferable metacognitive thinking involved and she actually reported the end product of her revision decision (i.e., actually I don’t know). The first revision (crossing out In) was related to a complete CVR of metacognitive knowledge because the participant reported the reason, though indirectly, for her revision behavior (i.e., In was capitalized).

Example 3 (Participant 22):

The revision made:

and he still live in contryside contryside. Because my brother and sister was are very busy…

His CVRs concerning the revision:

And he, he now, he now, he, ah, he, my father was old, he but he, ah, 他仍然 and, was old and, and, he is still t-i still l-i-v-e in country c-o-n c-o-n-t-r-y countryside.
The first revision concerned the spelling of countryside, which the participant pronounced repeatedly until he finalized its form. Though much cognitive effort was spent in the revision process, no metacognitive knowledge was expected to be reported. The second revision from *Bea* to *Because* also concerned spelling, but it did not take much effort and *Bea* appeared to be an a-slip-of-the-pen error. The CVRs related to these two revisions were perceived as complete.

Overall, under this category, the CVRs could be seen as rather complete because there was not much metacognitive thinking underlying the revision behavior that could or was expected to be reported.
6.5 Complete Concurrent Verbal Reports for Metacognitive Thinking

In 34 cases, the participants were able to report their metacognitive thinking quite completely, and naturally, in greater detail and with more accuracy than was revealed by their RVRs.

The following example shows how one participant completely reported what he had intended to write.

Example 1 (Participant 14):

Revision:

Because of his sad childhood, he tried every means to become a good father.

His CVRs concerning the revision:

Because of his sad childhood, he tried every means to become a good father.
Because of his, (4s), [exhalation], what to write? [making clicks], well, (7s) well, because of his, because, because of his, because of, his sadly childhood, because of his poor, his poor, o, a, his sadly childhood, his sad, what sadly, [indistinguishable voice] because of his sad childhood, (10s) what is thrifty in English? Hardworking, he is very hardworking, because of his sad childhood, his, oh, economical, economical, thrifty, he is very, he is, he is very alas, not right!

I needed to write about the past, (9s) oh, simple, descriptions, right, it was the past, [read the previous sentence], right right right, not the present written here, in that case one part would be skipped over, because of his sad childhood, right right right, then I can directly jump to, his attitude, attitude towards me, when I was a child, because of his child, sad childhood, he tried tried every means, he tried every means, every means, to become, to to become, a good father.

His/Her RVRs concerning the revision:

当时肯定是想要表达 …… 通过另外一个句子来表达这个意思，但是 ……
后来又发现这个句子不太对，不太好，所以就给又改了，改成了 he tried every means to become a good father。当时是 …… 我来想一下, because of his sad childhood, he is very …… 哦！当时是想写，因为他不幸的童年，他养成了什么什么样的性格。但是后来觉得这个不好又给改了。

At that time I must have wanted to express ... express this meaning through another sentence, but ... later I found the sentence was not right, not good, so I changed it, changed it into he tried every means to become a good father. At that time ... let me think, because of his sad childhood, he is very … Oh! I had wanted
to write, because of his sad childhood, he developed a certain character. But later I did not feel it good and changed it.

The participant’s CVRs provided a natural flow of metacognitive thought concerning the decision he made to abandon what he had partly written (i.e., he is ve). A comparison of his CVRs and RVRs showed that the reason for his revision was more completely (and specifically) and more accurately presented in his CVRs (his concern that he should focus on his father’s past in this paragraph; he had wanted to write his father is thrifty) than was answered in his RVRs (the sentence was not right, not good; I had wanted to write, because of his sad childhood, he developed a certain character).

The following example shows how one participant reported her metacognitive thinking completely using some metalanguage:

Example 2 (Participant 26):

Revision:

[He is a good teacher, all his students think highly of him.] He Not only did he teach his students ademetic things ….

Her CVRs concerning the revision:

举一个具体的例子来说，好，就说，行吧，再写，瞎编一个老套路的，他，他不仅，他不仅仅在教育他们知识，就是，he, he not only but also, 就这么用吧，Not only he，就是他教会他同学的知识，not only he teach，his students, [inhalation] 我应该说是学术的知识，ademetic, 就是不仅, not only he teach his, students, teach somebody about something, teach somebody something 吧，就是，something 就是，a-de, 呀，ademetic 这个单词好像不对，ademetic, 哎呀，[making clicks]，学术方面的东西，算了，不会表达这
个，就是，not only 会不会要倒装? 应该是，did 吧，用过去式算了。Not only did he teach his students ademetic things, but also ……

To cite a specific example. Well, then, ok, write then, fabricate a stereotyped one.

He, he, not only, he not only teaches them knowledge, well, he, he not only but also, that's it, Not only he, he teaches his students knowledge, not only he teach his students, [inhalation], I should say it is academic knowledge, ademetic, that is, not only, not only he teach his students, teach somebody about something, teach somebody something. That is, that is something, a-de, oh, the word ademetic appeared wrong, ademetic, alas, [making clicks], something academic, just leave it here, I can’t express this. Then, should not only lead an inverted sentence? It should. Then, did, use the past tense form. Not only did he teach his students ademetic things, but also ……

Her RVRs concerning the revision:

Not only 在前的话是要倒装，所以我把这个 he 划掉。这个 not only 倒装句嘛，应该加一个词，但是我忘记加了，所以就写 did。

Not only, when put in the front, should lead an inversion, so I crossed out this he.

For this not only inversion, one word should be added, but I forgot it, so I wrote did.

As this example shows, the participant was able to completely report her metacognitive thinking (i.e., partial inversion was needed after not only, which should be followed by the auxiliary verb) underlying her revision behavior.

The following example shows how one participant completely reported what underlay her idea reformulation in his think-alouds:

Example 3 (Participant 51):

The revision:
Comparing her past and present,

His CVRs:

Comparing her past and present.

(4s) Comparing her past and present.

write the comparison of the past first. (4s) Comparing her past and present.

His RVRs:

想先写未来，然后再看一下条件是先写比较。

I had wanted to write about the future, but then I read the instructions and understood I needed to write about the comparison first.

Participant 51, in both his concurrent and RVRs, reported the reason for his revision. That is, he realized it was the time that he should write about the comparison paragraph rather than the future paragraph.

The following example shows how one participant completely reported what he thought when he made a revision regarding diction:

Example 4 (Participant 84):

Revision:

Recently, my professor teach talk about culture revolution in China.

His CVRs concerning the revision:

Recently, recently, my, recently, my professor, my, professor, my professor, 最近我的教授给我们讲解了文化大革命的故事。Recently, ↓recently↓, my p-p-r-o-f-e-s-s-s my professor, recently, my professor, [inhale], erm, say, told me, ah, 最近, my professor, my professor, [exhale], teach, professor, my 这个不应该用 teach, my professor, teach 太简单, 太幼稚, my professor, [inhale].
Recently, my professor told us the story about the Cultural Revolution. Recently, my professor, my professor, my professor, [inhilation], erm, say, told me, ah, Recently, my professor, my professor, [exhalation], teach, professor, my _Here_ teach shouldn’t be used. my professor. the word teach is too simple, too naïve, my professor. [inhilation], (3s) talk, t:: talk, about, China, want, see, t-u-r culture revolution, r-e, u-a-t-i-o-n revolution, in, China, ah.

His RVRs concerning the revision:

后来说“我的教授教我们”，“教”的话，teach 这个词也比较简单。我觉得有
点儿太生硬，然后就换成了 talk about，用 talk about 我觉得比较清楚一点。

Then I said “My professor teach us”. As for “teach”, the word, teach was fairly simple. I felt it a bit stiff here and then changed it to talk about. Using talk about,

I felt it a bit clearer.

As the contrast of the above CVRs and RVRs shows, the participant completely reported what he naturally thought of when he changed teach to talk about in his CVRs (too simple, too naïve).

6.6 Summary

Overall, the above analyses indicate that the participants’ CVRs were found not to contain all their ongoing thoughts concerning the revisions they reported on; that is, the participants were not able to report whatever there was in their minds that they heeded. While some lack of reporting may be fragmental or incidental (i.e., failure to report what they had intended to write or what they actually wrote), the lack of reporting of underlying metacognitive thinking was substantial (i.e., failure to report what motivated
quite a number of revisions metalinguistically). This may be because such thinking occurred transitorily and was not easy to capture. The incompleteness of CVRs to this degree did not appear to be crucial, however, if the purpose of getting the CVRs was to understand the revision behaviours, because the missed metacognitive information was able to be inferred from the resources available (e.g., the CVRs and the compositions).
CHAPTER 7 VERIDICALITY: RESULTS FROM QUALITATIVE ANALYSES OF STUDENT REFLECTIONS

This chapter presents the results of my qualitative analyses of the 43 think-aloud participants’ answers to the first and second reflection questions (i.e., Do you think what you reported represents what you thought completely? and Do you think what you reported represents what you thought accurately?) and their elaborations. As is the case in Chapter 5, in the examples given below, English translations were offered in brackets in italics.

7.1 Student Reflections on the Completeness Question
Thirty-seven participants gave valid answers while six participants had an off-topic answer (e.g., they answered whether their thoughts were completely written rather than reported). Among these 37 participants, 18 thought that their CVRs were complete, while 19 thought otherwise.

7.1.1 Answers and reasons for complete reports
In answering the completeness question, the 18 participants insisted on their adherence to the TA instructions and/or made an account of what they completely reported as justifications for complete CVRs.

Participant 23 stated that she reported all her thinking and composing processes:

都差不多说出来了，思考的过程，写作思路都说出来了，可能写作中间有一些想法会遗漏，但应该都会说出来的。

(Almost all was reported. The thinking process and the conceptions for writing were both reported. Perhaps some ideas were missing in writing, but I should have reported them all.)

Participant 31 elaborated how she spoke throughout the task as required:
My thoughts were reported rather completely, because I almost spoke nonstop and essentially had no time to think otherwise.

In the course of writing, before I wrote, I reported my thoughts when I saw the topic, including things like how I planned for it. When I started writing, I basically reported all Chinese and English words and sentences that jumped into my brain. I appeared to enter a mode in which I would report everything that came into my mind so that I even reported irrelevant information like my throat was somewhat sore.

Similarly, participant 39 answered that he/she had reported all his/her thoughts, including conceptions, choice of words, paragraphing, topic considerations, and transitions.

Participant 67 stressed that he/she reported all that stayed in his/her consciousness:

I reported them all. I reported all my first thoughts and logic. But I did not do much planning before I wrote and I wrote immediately after I reported, so my logic was perhaps somewhat disordered. But I got them from the depth of my
consciousness and I directly spoke of them while I wrote them down, especially the part concerning my schooling experiences at the beginning.)

Participant 69’s reply focused on his/her complete reporting of his considerations of words. He wrote,

在写作过程中所有能想到的词汇，包括一些选择词汇与因想不到造成的困扰等均已说出来。但考虑到有一些较高级词汇虽曾在某些场合见到，但由于只有大概印象，致使想用却无法写出，读音也忘记导致未能说出。

(In the course of my writing, I reported all the words that I could think of, including some choices of words, and the trouble I met when I could not find words. But there were some superior words I had ever seen on some occasions, but because I only had rough impression, I could not write them down even though I wanted to use them. I forgot their pronunciations so that I was not able to report them.)

In the above answer, the participant mentioned his/her inability to report some superior (i.e., more sophisticated) words. Given that he/she did not think of them, his/her inability to report them should not be accepted as evidence for incomplete reporting because they were not in his/her mind in the first place.

Participant 84 described how he/she completely reported his/her thoughts while writing the lead-in and middle parts:

基本上都说出来了，因为开始的引题法子比较多，我选择了一个比较新奇的方法，我比较的过程也描述出来了，在中间的写作，因为都是按要求写的，所以想到什么写什么，思路一直很流畅，所以基本念的是自己写的作文内容。
(I basically reported all of them. Because there were a lot of ways to introduce the topic at the beginning, I chose a relatively new method. I also described the process of my comparison. In the middle, because I wrote as required, I wrote whatever I thought. My course of thinking was all the time smooth. So I basically read the contents of the composition I wrote.)

As a final example, Participant 92 answered that he/she reported all his/her thoughts except for “nonthoughts”.

基本都说出来了。有些没说是因为用的比较简单的词或句型，感觉就像选择“I”一样没有想法。做得比较好的是把思路说出来了。

(I basically reported all my thoughts. I did not report some thoughts because I used relatively easy words or sentence structures. I felt I had no thoughts just like I chose “I”. What I did relatively well was that I reported my train of thoughts.)

7.1.2 Answers and reasons for incomplete reports

The 19 Participants reflected on what was missing from their CVRs. They mentioned thoughts that flashed by, or flooded in, or were hard to verbalize, abstract, privacy concerned, or irrelevant to the task. Several of them also mentioned that their subconscious activities were not reported.

Thoughts that formed too rapidly to be caught and verbalized were most frequently mentioned unreported thoughts. Participant 26, for example, explained, “我的瞬时记忆消失了” (my momentary memory vanished), but he played down the effect, “但那些大都是单词的拼法或短语的搭配。那个突然冒出的想法也不是很多” (but those were mostly spellings or collocations, and those thoughts that popped up were not many).

Participant 27 also mentioned flashing thoughts. He/She wrote,

因为有在纸上列作文提纲，看的时候想的过快，来不及说，还想到一些画面，一些事件，他们一闪而过。
(Because I had an outline on the paper. When I looked at it, I thought too fast. I could not speak timely. I also thought of some scenes, some events. They flashed by.)

Participant 53 also talked about some too-fast-to-catch thinking:

感觉边说边写时还是思维的速度比较快，很多想法就直接过去了来不及说，有时候脑子里想了好几句，但说还是得一句一句的来，两个节奏对不上，感觉说的还是不太透彻，只能一定程度上表现想法。

(Feel in thinking aloud the speed of thinking was fast. Many thoughts passed by directly without being reported. Sometimes in my brain I thought of several sentences, but I had to report them one by one. The two did not match in pace. Feel I did not report in depth. My report could only express my thoughts to a certain extent.)

Similarly, Participant 57 admitted his/her failure to report words that flashed across him/her mind and then vanished:

在写的时候对文章框架的想法，接下来些什么的想法都说出来了，但是在对词汇的选择上总觉得有很多内容没有说出，因为脑海里很多词汇一闪就过，然后就想不起来了。

(When I wrote, I reported all my thoughts on the outline of the composition and what I would write next, but I always felt I did not report many contents in terms of word choice, because many words flashed across the sea of my brain and then I could not recall them.)

Several participants noticed the missing of barely verbalizable thoughts, including unconscious behaviours, from their CVRs. Participant 24, for example, replied,
没有。在我写作的时候，有些句子结构上的安排和句子的顺序的安排是很
难用语言表达出来的。还有些修辞是潜意识中选择的。

(No. When I wrote, it was hard to verbalize some arrangements in sentence
structures and sentence orders. Besides, some rhetoric was chosen
subconsciously.)

Participant 49, while mentioning some flashing content concerning his/her father’s
education of him/her, added difficulty in verbalization as another reason for his/her
incomplete reporting. He/She stated, “挑了一些好写的说” (I selectively reported what
was easy to write). He/She offered two examples:

讲爸爸的过去时想说一下他们从前在农村生活很苦但感觉不好说，反正也
不写。讲现在时想说爸爸十分想念我把每天打电话给我问问情况作为一项
任务好像没有说。因为想的时候老是“嗯”过去了，写的时候老是习惯写什
么说什么。

(When I wrote about my father’s past, I wanted to say their life in the countryside
was bitter but I felt it hard to say that. And I did not write about it after all. When
I wrote about my father’s present, I wanted to say my father missed me very much
and undertook calling me as a daily task. I did not seem to report that. I uttered a
“well” when I thought. When I wrote I was used to reporting what I wrote.)

Participant 75, likewise, recognized thoughts that were not verbalization-friendly,
after referring to the instantaneous nature of some thoughts:

有些思维过程有些时候一瞬间就形成了，来不及说出来，一般情况下会补
说出，但很多时候就忘记了。此外，有一些小想法很难用合适语言描述出
来。
(Some thinking processes were sometimes formulated at a moment, more quickly than I could timely report. Generally I would report them afterwards, but many times I forgot them. In addition, some small thoughts were hard to describe with appropriate language.)

Similarly, Participant 65 mentioned the difficulty in verbalizing certain thoughts as well as momentary thoughts:

没有，因为有些想法稍纵即逝，自己没有时间可以将其完全毫无保留地说来……有很多想法是无法用语言来表达的，不能全部重现我当时的想法。

(No, because some thoughts were there one second, gone the next, and I had no time to report them completely and without any reservation ... There were many thoughts that could not be expressed orally. They could not completely represent my thoughts at that time.)

Participant 79 referred to difficulty in verbalization as well, when he/she admitted that he/she was not able to report some thoughts that he/she could not organize well before being interrupted by vocalisation. However, he/she also stated that he/she basically reported all that emerged in his/her mind, including the overall outline and language organization.

Participant 99 also referred to difficulty in verbalization, when he mentioned that he was not able to describe some abstract thoughts:

有的想法比较抽象的，无法描绘 ...... 有的时候想法只是一种感觉，在静下心来表达（写）自己的话时会逐渐显现出来，但若要求 keep speaking，自己的非常好的思路往往会消失。

(Some ideas, which were abstract, I could not describe ... Sometimes ideas were only a kind of feel. When you calmed down and expressed yourself (wrote) they
gradually emerged. But if you were required to keep speaking, your very good ideas could often disappear.)

The above statements indicated Participant 99’s view on the possible inability of TA to catch his abstract thoughts or even feelings, although these statements also contained some reactivity accounts.

Participant 100 also mentioned difficulty in verbalizing thoughts completely, which he/she related to the elusiveness and fleeting nature of thoughts. He/She wrote,

并没有全部说出来，因为有些思维的东西，觉得只可意会而不可言传，它就在脑子里，但是就是表达不出来或不完全。有些想法来得太快去得也快，有些本应该想到的方面却又没考虑到，如果真的要全部都说出来，可能需要很长时间，把各方面都考虑进去，但实际情况不是这样的，往往要求快速。

(I did not report them all, because some thoughts I felt were to be perceived rather than expressed. They were in my mind, but I simply could not express them or express them completely. Some thoughts came fast and went fast. There were some aspects that I should have thought of but I did not. If I had really made an attempt to tell them all, it would have taken a long time, to take every aspect into consideration. But the actual situation was not like that, and one was often required to be fast.)

Then, Participant 62 reported on thoughts that flooded in, as well as thoughts that occurred too fast:

比如有时大脑中会一下冒出多个句型、单词等，来不及一起说出来。还有对文章结构的构思，想得特别快，也来不及说出来。
(For example, sometimes many sentence patterns and words popped up in my mind all at a moment. I could not report them together timely. Then my conception about the structure of the composition, I thought about it very fast and could not report it time.)

From a different perspective, Participant 56 admitted that he/she refrained from reporting some private experiences:

大部分已经说出来了。有一些比较私人的一些经历不知道怎么表达，一闪而过所以没有写出来或者说出来 …… 由于涉及个人隐私，有些印象深刻的记忆会不由自主闪现但不好意思写出来（说出来），就会所想与所说有些差异。

(I told aloud most of them. I did not know to express some personal experiences. They flashed across so I did not write them down or mention them ... Because they involved personal privacy, some impressive memories would flash automatically but I would have been embarrassed if I had written about (told aloud) them. So what I thought was not consistent with what I reported.)

Likewise, Participant 94 mentioned that “有些情况不方便说出” (there were cases that it was not convenient to speak out).

Another participant, Participant 41, noticed that he/she did not report irrelevant information. He/She wrote about his/her distraction, “写作文的时候思维跳到别的地方的一部分没有说出” (When I wrote, I did not report it when my thought jumped to other places).

Finally, one participant, Participant 35, mentioned that sometimes, his/her preoccupation with writing prevented him from reporting all his/her thoughts. He/She
answered, “有些时候写的忘情了，就自然而然不说了” (Sometimes, I wrote from the heart and stopped speaking naturally).

Noticeably, some participants, though answering that they did not report their thoughts completely, elaborated that their perceived incompleteness stemmed from their lack of words or their inability to construct sentences to express their thoughts in English. Participant 25, for example, stated, “比如自己有时候遇上一些地方无法用英文表述的时候，也没有办法把自己的想法完整的说出来” (For example, when I sometimes came across some places I could not express in English, I could not tell my thoughts completely). The problem in his/her answer was that the English words or sentences or ways of expression were not in his/her consciousness in the first place. So this should not be taken as evidence for incomplete reporting of what was heeded as long as he/she reported it in Chinese or reported related thoughts. Participant 35 had an exemplification:

有些时候，你突然想不到一个单词怎么读，但你还是会写，这时就很纠结，可能会忽略过去。

(Sometimes, you suddenly forgot how to pronounce a word, but you still wrote it down. At this time you felt entangled psychologically, and possibly skipped over it.)

Not verbalizing the English word one did not know how to pronounce may not be treated as incomplete reporting because the English pronunciation was not in one’s mind in the first place.

7.1.3 Summary

Overall, judging from the participants’ answers, their CVRs could be rather informative, although not complete due to certain missing elements that primarily concerned flashing by and difficult-to-verbalize thoughts. The types of thoughts missing
from CVRs reported by the participants therefore supported the results of incomplete CVRs found in the previous analyses of the participants’ CVRs against their RVRs.

7.2 Student Reflections on the Accuracy Question

Thirty-two participants gave valid answers while 11 participants had off-topic answers (e.g., Participant 26 answered that the figure and his experiences in his/her composition were fabricated). Among these 32 participants, 27 thought that their CVRs were accurate, while five thought otherwise.

7.2.1 Answers and reasons for accurate reports

The participants insisted on the accuracy of their CVRs by referring to their adherence to the TA instructions or their successful adjustment to the TA task.

Participant 10, for example, stated that he had no time at all to think in one way and report in another. Participant 14 felt that TA was just like speaking to himself, which he had done when writing under normal conditions.

Participant 25 had a more elaborate reply:

(I feel my report was an accurate reflection and was in consistency with what I thought, because I feel the process of writing was to record the results of thinking in the brain selectively, purposively and in an orderly manner, and what I reported was just a description of the miscellaneous thoughts behind this result. The reflection was therefore accurate, and there was no such a thing as unwillingness to speak it out.)
Participant 31 thought her report was accurate even though her thoughts were affected somehow:

对于我来说基本准确反映了我的所想，就如上题中说道好像已经进入了一种特定的想什么就说什么的模式，因此没有考虑什么该说什么不该说。

但是存在一个问题就在于似乎在说的时候，影响了思考，所以可能想的就改变了，但说出来的和被影响之后的想法还是基本一致的。

(To me my reporting basically reflected my thoughts accurately. Just as I had said in answering the first question, I seemed to have entered a specific mode where I spoke out whatever I thought of. So I did not consider what I should report and what I should not.

But one problem that existed seemed to be that when I reported, the reporting affected my thinking. Therefore my thoughts might have been altered. But what I reported was still basically consistent with the altered thoughts.)

Participant 69 was similar in that he/she also mentioned his/her getting used to reporting whatever was in his/her mind:

并没有此种情况存在，事实上在写作过程中，由于习惯于在写的时候说出所想，我个人感受所说均为下意识的对自己想法的反映，难有就算想到也刻意不说这种情况存在。

(There existed no such a case (of inaccurate reporting). In fact, in writing, because I got used to verbalizing what I thought while writing, I personally feel all that I spoke out was a reflection of my thoughts. It was hard to have a case where I intentionally withheld what I thought.)

Participant 79 also justified his/her accurate reporting by his/her full engagement with the TA task. He/She wrote,
我的所说确实反映了我的第一所想，因为需要边说边想边写，所以无法有太多的思维去控制哪些想说哪些想要隐藏，表达出的东西都是很直观的想法，并不掺杂其他东西。

(My reports really reflected my first thinking. Because I needed to speak while thinking and writing, I had no spare thinking to control what to report and what to conceal. What I expressed was all very direct thoughts and did not contain any mixture of other things.)

Participant 39 stated that he/she reported all that was in his/her consciousness. He/She wrote:

想说的都准确反映了，但有时选词自己未意识到。但所想的都准确反映，包括组织、选词，英汉夹杂就是代表。

(What I wanted to speak out was all accurately reflected, but sometimes I was not aware of alternative words. But all that I thought of was accurately reflected, including organization, and diction. The mixture of English and Chinese was evidence.)

Participant 57 also replied to this question with affirmation. Additionally, he/she mentioned his/her re-reading as a strategy to fill the mental gap.

基本是真实想法，包括单词忘记、不会拼写的现象都有提到，但有时在停顿的时候会大脑一片空白，不知道说什么，所以会不断重复已经说过的句子，以给大脑留出思考的时间。

(Basically, I reported all my real thoughts, including my forgetting of words, and my inability to spell, but sometimes when I paused, my mind went blank and I did not know what to report, so I would repeatedly read sentences that I had reported, so that there would be time left for my brain to think.)
Participant 98 mentioned his/her effort to follow the experimental instruction. He/She wrote, in a somewhat formal style:

为保证该写作实验录音文件结果的准确性与客观性，实验者所有想法均毫无保留并准确表达于录音文件中，所有想法均属实。

(To ensure the accuracy and objectivity of the recording for this writing experiment, all the thoughts of the participant were expressed in the recording accurately and without reservation, and all the thoughts were real and true.)

The participants also stressed their candidness in speaking with no scruples. Participant 17, for example, answered that in his planning there was nothing about which he felt awful, and he was not afraid that he would give anything away, given the clear confidentiality terms that applied to this experiment. Participant 18 and Participant 27 answered that they reported whatever they thought of accurately because there were no privacy aspects involved. Participant 23 also replied that there was no place that she was not willing to speak out. Likewise, Participant 43 and Participant 62 mentioned the fact of no privacy concerns as the reason for their accurate reporting, stating “现实情况中也没有什么难言之隐” (there was nothing in the actual situation that it would be awkward to disclose) and “这篇文章并没有牵涉到个人隐私” (This composition did not involve personal privacy), respectively.

7.2.2 Answers and reasons for inaccurate reports

The participants who speculated that their CVRs were inaccurate did not provide cogent answers. Participant 49, for example, referred to the incompleteness of his/her reports. Participant 56’s justification was similarly related to the incompleteness question. Participant 70 answered,

不一定准确，因为说有些影响自己写作，而且说的内容有故意的成分，像是为了思考而说的。
(Possibly inaccurate, because speaking affected my writing in some way. Besides, my reports contained some intentional elements. They seemed to have been spoken out for thinking.)

His negative answer, as can be seen, was partly based on the reactivity issue. And the “intentional elements” would be accurate if they were his/her thoughts at the time of reporting or, in other words, if he/she thought of being that intentional.

Participant 99 was interesting in that he recognized the accuracy of his reports, but he thought that “自己的想法表达出来，本身也会和所想有差异” (One’s thoughts, once spoken out, would in themselves differ from the original thoughts).

Participant 100 based his/her judgment of inaccuracy on the accuracy of his expression. He/She wrote,

不是太准确，但也差得不太多，说的时候，感觉表达不够准确，靠不到最准确的那个意思！本来是那样想的，但却因为有些原因，觉得有些不太习惯这种方式吧！

(Not very accurate, but fairly close. When I spoke, I felt my expression was not accurate enough and could not touch the most accurate meaning! I originally thought that way, but because of some reasons ... I felt I was somewhat not used to this way!)

7.2.3 Summary

Overall, the accuracy of CVRs was supported, judging from the participants’ both positive and negative answers. Several major justifications that can be distilled from their positive answers included their competent and honest undertaking of the TA writing task, and their having no scruples in reporting, due to there being no privacy and embarrassment in the content of their reporting. The several participants who thought
their reports were not sufficiently accurate, in fact, did not provide sound or relevant reasons.
CHAPTER 8 DISCUSSION

This chapter discusses the results of my study in the order in which the research questions were presented (see Chapter 3, p. 68).

8.1 Reactivity on L2 Writing Performance

The question of whether TA had any significant effects on L2 narrative writing performance did not receive a uniform answer. That is, TA had some significant reactive effects on L2 learners’ writing performance, but the effects were only felt in limited aspects. To be specific, TA did not significantly affect 1) time on task; 2) fluency in terms of the speed of L2 writing (i.e., Speed One, Speed Two, Speed Three, and Speed Four); 3) most aspects of complexity including the various lengths of the written text (i.e., Length One, Two, Three, and Four), general complexity, syntactic complexity, subclausal complexity, and SV; 4) accuracy (i.e., CVU, EFWs and EFCs), and 5) Content. It only significantly affected fluency in terms of the increase in dysfluencies, significantly decreased lexical variety, and significantly reduced the quality of organization.

First, TA did not significantly increase the time spent on the L2 narrative writing task. Neither did it significantly reduce the speed of L2 writing, despite a slight slowing down effect across all the four speed measures. This finding is consistent with the results of Yang et al.’s (2014) study, in which no significant effects on time on task and the speed of L2 argumentative writing were found except for a slight decrease in time and the L2 writing speed. The slightly retarding effect is also in accordance with the findings from L1 research. It supports Ransdell’s (1995) similar finding that TA caused college students to slow down significantly in writing a narrative. The relatively short writing time (12 minutes) allowed in Ransdell’s study may have made the slowing effect more salient. The speed-related findings also appear to endorse Janssen et al.’s (1996) finding
that TA significantly increased pause time, although the speed of writing is a measure different from pauses. The temporal aspects of findings, both from this study and from the other studies on reactivity in writing, both L1 and L2, taken together, add to Bowles’s (2010) meta-analysis synthesizing a quite consistent prolongation effect of TA on time on task, and confirm Ericsson and Simon’s (1993) predictions of longer time on task or latency. The insignificantly lower speed of L2 writing found in this experiment, as it was in Yang et al’s study, may be the result of several counterbalancing factors that affected time or speed. The extra effort taken to verbalize thoughts in the process of L2 written production, which may have been spent on recoding of thoughts that were not in a readily verbalizable format, an activity characteristic of Level 2 verbalization (Ericsson & Simon, 1993), would have increased time on task or reduced the writing speed. However, this possible delay may have been offset by the participants’ reluctance to sacrifice their writing speed and therefore their willingness to maintain it at the cost of other aspects of performance. Another counterbalancing factor could be the accelerating effect generated from memorization of planning, which was enhanced by listening to one’s own voice (Penney, 1975; Russo et al., 1989).

Then, TA incurred significantly more dysfluencies; that is, it caused more cross-outs and reformulations. This finding was consistent with Yang et al.’s (2014) finding on the same measure. It showed that the participants, when thinking aloud, did not write as thoughtfully as when they wrote silently, and corrected more words afterwards. In reference to Kellogg’s (1996) model, it indicated a crowding-out effect of TA on the monitoring processes involved in planning, translating, and even execution (considering that L2 learners may not have full automation in spelling, e.g., Ellis & Yuan, 2004), only to be compensated by more post-hoc monitoring processes (i.e., reading and editing processes) launched to keep up the quality of writing (see e.g., Rijlaarsdam & Van den
Bergh’s 1996 compensatory system account of writing processes) (Yang et al., 2014). The displacement of monitoring was made, as Yang et al. (2014) have suggested, because the demand for cognitive processing right after execution could be less. The increased instances of dysfluencies found in this study and in Yang et al.’s (2014) study provide joint evidence for Kellogg’s (1996) prediction that the quality of translating may be disturbed by TA at least. This finding also supported Janssen et al.’s (1996) conclusion that TA disturbed normal writing processes based on an examination of pause lengths and pause locations pertaining to different levels of writing subprocesses. As has been discussed by Yang et al., the fact that more dysfluencies were caused was related to the often negligible status of the measure in scoring L2 writing quality so that this aspect of performance was easily sacrificed. Besides its minor status, the measure itself may be elastic and subject to external interference, and this may have also caused the effect.

As regards the length of the composition, TA did not cause any significant effect, although there appeared a quite consistent direction of effect pointing to somewhat shorter texts. This finding was consistent with the findings of previous studies examining the same/a similar measure that administered different writing tasks such as a description or a semi-guided writing task (i.e., Levy & Ransdell, 1995; Ransdell, 1995; Yanguas & Lado, 2012). Based on Kellogg’s (1996) model, shorter texts, written within a fixed period of time, may point to a constraining effect on overall planning or translating. However, given that the effects on text lengths were not significant, and given that TA did not affect the quality of content significantly, this aspect of effect should not be over-interpreted.

That only the effect of TA on Length 3 (the number of syllables produced) approached statistical significance was related to the significant decline in lexical
diversity. An ANCOVA test on syllables per word controlling for prior differences showed a decline in the measure under the TA condition, adjusted $M$ (TA) = 1.331, $M$ (NTA) = 1.356, $F$ (1, 81) = 3.230, $p = .076$, $\eta^2 = .038$, indicating that, while thinking aloud, the participants tended to use words that contained fewer syllables. This tendency, together with the insignificantly shorter texts in terms of Length One, the number of words produced, made the reduction of Length 3, the number of syllables produced, more salient. Correlation analyses confirmed the relationship between syllables per word/Length Three and lexical diversity, Pearson’s $r = .462/.260$, for the baseline task, and Pearson’s $r = .512/.202$ for the main task, in contrast to a rather small correlation between Length One (the number of words) and lexical diversity, Pearson’s $r = .069$, for the baseline task, and .025 for the main task.

That TA had no statistically significant effect on Length Four (i.e., the total number of syllables produced) was found to be the result of the constraining effect of TA on Length Three (in relation to significantly decreased lexical diversity and insignificantly decreased Length One) being offset by the statistically significant increase in dysfluencies. The correlations between the number of syllables deleted and dysfluencies were strong ($r = .942$ for the baseline data and $r = .920$ for the main task data), and the TA group had greater number of syllables deleted than the NTA group, adjusted $M$ (TA) = 26.763, $M$ (NTA) = 22.615, $F$ (1, 81) = 2.092, $p = .152$, $\eta^2 = .025$. This contribution to Length Four due to increased dysfluencies therefore counterbalanced the decrease in Length Four due to decreased lexical diversity, blocking the extension of the almost significant decreasing effect of TA on Length Three to Length Four.

Consistent with Yang et al.’s (1995) findings, TA did not significantly alter general complexity, syntactic complexity, subclausal complexity, and SV. These nonreactive findings echo Ransdell’s (1995) analogous findings of nonreactivity on complexity
measured by the average length per clause and the total number of clauses. The nonreactivity on general complexity also confirms Yanguas and Lado’s (2012) finding of no effect of TA on the number of words per T-unit. In reference to Kellogg’s (1996) model, it showed that despite the additional demand of CVR, the participants’ thinking processes concerning syntactic translating (e.g., framing plans with syntactic structures, using modification, and verb form choice) were neither inhibited nor facilitated in any significant way. The reason why TA did not affect these aspects of complexity was that they may have been conceived to be rather essential to expression of meaning, which is a foremost concern in writing (e.g., VanPatten, 1990) (Yang et al., 2014). The participants still retained the necessary amount of executive control for these concerns of complexity, even in the case of undertaking the simultaneous task of thinking aloud.

Within aspects of complexity, the exception regarding the effect of TA was lexical variety, where TA showed a significantly negative effect. This effect was also highlighted in the participants’ reflections. Expression of meaning was still given priority, and word retrieval was also prioritised for conveying meaning in the process of L2 writing, as Ellis and Yuan (2004) and Yang et al. (2014) have evidenced. However, these participants, who would have had a high level of awareness of lexical diversity due to the training they had received, as long as they caught words that sufficed to express certain intended meanings, would use them, without further thinking over what other more complex words would represent the same meaning and could be used instead. In reference to Kellogg’s (1996) model, this would mean that the participants’ translating processes in charge of word choice they would have utilized under a silent condition were significantly inhibited in terms of their degree of thoughtfulness under the TA condition. This finding offered evidence for Kellogg’s (1996) prediction that verbal reporting may negatively affect the quality of translating due to the possible
competition for WM resources. This finding, however, contradicted Yanguas and Lado’s (2012) finding of increased lexical complexity. While I admit that writing in a heritage language, which was the case in Yanguas and Lado’s study, might be different from writing in a foreign language, I doubt if there was sufficient evidence that, when thinking aloud, their participants could write with no constraint within 25 minutes and expend greater-than-usual attention to lexical choice (and accuracy as well), so that lexical complexity (and accuracy) could be raised (see also Yang et al.’s 2014 questioning). Design issues may explain the difference in the contrastive results. I will explain the different effects of TA on lexical diversity found in this study and Yang et al.’s study later when I discuss task effects. In this experiment, lexical diversity was impaired because this aspect of performance was unstable for the lexically active L2 learner writers. The measure was perceived to be peripheral to meaning conveyance, and it could be cognitively demanding to keep this aspect of performance up due to the extra effort needed to find nonprototypical forms (e.g., Hu, 2002), so that they sacrificed this aspect of performance in face of the additional load of TA.

TA did not affect accuracy, even when accuracy was measured multi-dimensionally, and with the fine measure of the number of errors. This confirmed Yang et al.’s (2014) finding of nonreactivity on L2 writing accuracy, where EFCs and CUVs were used as accuracy measures. Both studies indicate that TA did not benefit translating or activate effective monitoring/editing processes, based on Kellogg’s (1996) model of writing. The finding of nonreactivity also supported previous findings of nonreactivity for TA on similar accuracy or correctness measures of reading comprehension (e.g., Bowles & Leow, 2005; Leow & Morgan-short, 2004; Rossonondo, 2007) or recall of propositions in the reading passage (i.e., Yoshida, 2008), rule learning in reading a passage (Bowles & Leow, 2005; Goo, 2010; Leow & Morgan-short, 2004; Polio & Wang, 2005), L2
grammar learning after a computerized lesson was conducted (i.e., Sanz et al., 2009, all tests in Experiment 1 and most tests in Experiment 2), L2 grammar learning through an L2 problem-solving task providing opportunities to learn (Bowles, 2008), and the effects of recasts on interactive production of verb forms in story-retelling (Sachs & Suh, 2007). It also corroborated Ericsson and Simon’s (1993) and Fox et al.’s (2011) conclusion of nonreactivity on accuracy of performance in completing problem solving tasks. The assumed facilitative effects on accuracy due to “motivational shift”, which derived from an intention to perform better in anticipation of CVRs being known (Russo et al., 1989, p. 765) were therefore not corroborated. In reference to Kellogg’s model, stable accuracy may indicate that while TA demanded concurrent processing, the correctness of form in L2 writing was still duly attended to; that is, formal accuracy was still given due care in the processes of translating, execution, and monitoring. This was because accuracy was given priority usually, especially in a testing-like writing environment (Wigglesworth, 1997; Ellis, 2009), and considering the somewhat exam-oriented English learning setting in China and the general emphasis on grammar accuracy in China’s English classrooms (Yang et al., 2014; Zhang, 2010). It should be noted that TA caused L2 writing to slow down insignificantly. This slight time delay may also have helped L2 writing to keep up its formal accuracy.

The absence of facilitative effects on accuracy found in this study and in Yang et al.’s (2014) study using L2 writing tasks and in the majority of SLA studies that used tasks other than writing outweighs Sanz et al.’s (2009) finding of facilitative effects that was only found in grammaticality judgment of new items and production of old items in their Experiment 2. It also contrasts with Yanguas and Lado’s (2012) finding of improved accuracy (and increased lexical complexity) in heritage language writing. Besides the experimental design issues I have pointed out, the participants in Yanguas
and Lado’s study would have needed to be rather familiar with TA and have written and verbalized under no constraint to benefit from it. This aspect of evidence, however, was not provided or discussed. Raising the accuracy of L2 writing could be cognitively demanding and may be somewhat difficult when the dual task of writing and reporting are to be juggled. Whether accuracy could be raised was also constrained by participants’ existing pool of knowledge (Yang et al., 2014).

TA did not affect the quality of the content. This would offer some support to Yang et al.’s (2014) finding of no effect on the overall quality of L2 writing and to Ransdell’s (1995) report of no impact of TA on the overall quality of L1 writing. No reactivity in the content aspect indicated that the participants’ planning of what he meant to write, namely, meaning conveyance, was not affected significantly. That is, even when they had to report, the participants still wrote what they intended to write and did not give up any planned propositions simply because they undertook the additional task of simultaneous reporting. This was because content was the core of meaning and meaning conveyance took priority even over formal concerns.

However, TA affected the quality of organization significantly in this experiment while it only tended to affect it in Yang et al.’s (2014) study ($F = 3.137, p = .048, \eta^2 = .066$, NTA vs. NMTA through the Bonferroni procedure: $p = .072, d = 0.582$). This showed that under the pressure of reporting their thoughts concurrently, the participants failed to or forgot to use sufficient connective words or expressions or other organizational means to make their writing flow smoothly. In reference to Kellogg’s model, this finding indicated that TA significantly constrained the participants’ organization-related planning, translating, and/or monitoring processes by significantly lessening their executive control over the coherence and cohesion of their writing. The reason was that organization, to them as L2 writers, may not appear crucial to meaning
They just wrote down what they had planned, without further considering connecting their ideas using cohesive ties and organizing these ideas in a coherent manner. The organizational aspects, to many of these Chinese L2 writers, were presumably their unstable knowledge. They would need to pay particular attention to these aspects to keep their performance in organization consistent so that, with the extra load of verbalization, the participants sacrificed organization for its particular demand due to the keener contention for attentional control. The sacrifice of organization might also relate to the participants’ lack of awareness of the importance of organization. They may not have established the idea of organization being crucial but may think it was only secondary to other concerns they perceived to be central.

8.2 Reactivity and Task Type

The different effects of TA found in this study and in Yang et al.’s study could be explained in relation to task differences and group differences as well. That TA appeared to have caused greater effects in Yang et al.’s (2014) study than it did in this study was in accordance with previous assumptions that the reactivity could be greater in cognitively demanding tasks (Ransdell, 1995) and corroborated Janssen et al.’s (1996) similar findings of greater reactivity for the more complex task of business lettering writing than for the less demanding explanatory writing task. The argumentative writing task in Yang et al.’s study was more challenging, which required more words to be written and entailed more complicated planning because no outlines were given and because more critical thinking and reasoning (i.e., presenting problems, analyzing reasons, and providing solutions) were involved. Given that in completing an argumentative writing task the demand for WM resources was already keen, the crowding-out effect of the additional task of TA was more likely to be greater than it was in performing a narrative writing task, where the contention was not as keen. It
should be noted, however, that the effects of task difference may have been reduced by the significantly more time taken for completing the argumentative writing task (62.183 minutes, 11.233 syllables per minute, for the TA condition in Yang et al.’s study vs. 39.649 minutes, 12.215 syllables per minute for the TA condition in this study) so that many of the differences in the reactive effects were minor.

It should be noted that, in both studies, the reactive effects or lack thereof were displayed in somewhat similar aspects of L2 writing performance except that TA significantly decreased Lexical Diversity in this study while only causing the same measure to decrease slightly in Yang et al.’s study. This difference in reactivity could be attributed to participant differences. The participants in this study were different from those in Yang et al.’s study in that they had developed an active level of awareness of diction due to the class lecture they received which familiarized them with online and WORD lexical resources and to the textbook tips which encouraged them to use “wonderful words”. As proof, many of the participants who thought aloud also mentioned their inability to think over “换词” (changing words) in their reflections. The participants could have paid or maintained much attention to diction in silent writing, affected or encouraged by the instructions they had recently received; however, under the think-aloud condition, they did not have the leisure to think about using nonrepetitive and diversified words, a concern that was not central to meaning conveyance to them. This finding suggests that TA could repress the effects of short-term training, and may be harmful to creativity in writing.

The aspects of reactivity found in this study, when compared with those found in previous reactivity studies in L1 writing using performance measures (Levy & Ransdell, 1995; Ransdell, 1995), tend to support the assumption that reactivity was more serious with an L2 (particularly in this study, an English as a foreign language, EFL) writing
task. This was because many aspects of writing performance, for example, dysfluencies, organization, or lexical choice, which were affected significantly for the Chinese EFL writers, may have become stabilized and may not be easily subject to external influence for L1 writers. However, given that the reactivity studies within L1 writing did not measure writing performance as multi-dimensionally as this study did, any conclusion drawn based on a comparison of the findings available so far needs to be seen as tentative.

The reactive effects found in this study also indicate the differences between a cognitive problem-solving task and a writing task. Performance of a writing task can be measured multi-dimensionally, unlike that of a cognitive problem-solving task, which is usually measured in terms of accuracy. This advantage of comprehensive measurement makes finer examination of reactive effects in writing possible. More importantly, a writing task, as Stratman and Hamp-Lyons (1994) pointed out, is an ill-defined task, where task completion is achieved via multiple routes that are unclear. I would add that in writing, especially L2 (EFL) writing, alternative routes may differ greatly in terms of their degree of centrality to the key task of expressing meaning and may involve the use of different states of knowledge (e.g., implicit or explicit, prototypical or aprototypical) and therefore differ in the amount of cognitive effort spent, so that those “best” but demanding routes that are possible under the silent condition may be abandoned under the TA condition due to competition for processing capacity.

8.3 Reactivity and L2 Writing Proficiency

Using baseline performance scores and one of the WMC scores as two covariates, statistically significant interaction between baseline performance and Condition was found on Speed Two, General Complexity, and CVU and, additionally, on Length Two only when WMC Two was used as one of the covariates. High General Complexity L2
learner writers benefited from TA in General Complexity, low General Complexity L2 learner writers suffered from TA in General Complexity, and low CVU L2 learner writers did even worse in terms of CVU. In a different direction, high Length Two L2 writers were significantly affected in Length Two, in the case of WMC Two used as a covariate, and low Speed Two L2 writers got a lift in Speed Two. The differential effects of TA on different L2 learners provide empirical evidence for the previous concern or observation that different writers may demonstrate different levels of adjustment to TA (e.g., Faigley & Witte, 1981; Hayes et al., 1987; Stratman & Hamp-Lyons, 1994). The finding largely supports Manchón et al.’s (2009) assumption that better learner writers were more able to deal with parallel activities cognitively during the writing process, since the low achievers suffered (in terms of general complexity or CVU) and the high achievers benefited (in terms of general complexity) from TA, except that low Speed Two L2 writers increased their Speed Two and that high Length Two L2 writers displayed less Length Two. It should be noted that an increase in Speed Two may not necessarily be a good thing, just as a decrease in Length Two may not necessarily be bad, because both measures count words involved in dysfluencies. Another new finding revealed by the differential effects found in this study is that the differential effects were limited since they happened only on certain aspects of L2 writing performance and in the majority of the other aspects, TA affected the participants equally.

That TA tended to make high Speed One L2 writers slow down was understandable, given the general conclusion that TA increased time on task or latency, which was based on a synthesis or meta-analysis of reactivity studies conducted both inside and outside SLA (e.g., Bowles, 2010a; Ericsson & Simon, 1993; Fox et al., 2011). Fast L2 writers might be easily affected temporally because they would have to slow down to capture and verbally encode their compact, efficient planning, translating, execution, and
monitoring processes in order to report them. In contrast, slow L2 writers, whose writing processes were relatively easy to track and then encode, would not have to sacrifice their speed for verbalization; instead, compared with fast L2 writers, they might have benefited more from the verbalization activity, which could enhance their memory of planning and remind them to keep writing. That Speed Two was significantly increased for previously low Speed Two L2 writers may also be attributed to the (overall) significant increase in dysfluencies, which may have made the otherwise insignificant differential effects on speed more salient.

TA facilitated the performance in general complexity of those whose writing had already displayed a high level of general complexity but it impaired that of the low achievers. This finding showed that TA might have enhanced the high achievers’ awareness of writing subordinate clauses and using modifiers and made them engage in more elaborate translating (or even monitoring) processes. But it may have crowded out a part of the low achievers’ attention to build such constructions and inhibited some of their translating (or even monitoring) processes. The high achievers could benefit possibly due to the “auditory feedback” (Stratman & Hamp-Lyons, 1994, p. 98) and memorization effects (on a single strategy) generated by TA (Penney, 1975; Russo et al., 1989). They could make better use of their already firm awareness to write long and complex sentences, as they may have long been instructed to do, and so allocated an even higher level of attention to this aspect of performance. The low general complexity achievers, in contrast, could not benefit from the above facilitative effects because their awareness of writing long and complex sentences was weak, and they failed to treat general complexity as an aspect that they needed to pay particular attention to in translating and other processes. Therefore, when juggling the dual task of L2 writing and TA, due to competition for processing, they disregarded this aspect of performance.
by turning to easier alternatives, namely, shorter or less complex sentences. ANCOVA analyses using Condition and (General Complexity) Grouping as two independent variables, the total number of words and the total number of T units, respectively, as a dependent variable, and the corresponding baseline scores and WMC Average as two covariates indicated that previously high general complexity L2 writers wrote a decreased number of T units under the TA condition (adjusted $M = 29.261$) compared with their counterparts in the NTA condition (adjusted $M = 28.100$), but an increased number of words (adjusted $M = 364.8$ versus adjusted $M = 352.9$), resulting in their significantly increased general complexity. In contrast, previously low general complexity L2 writers showed an increase in the number of T units they produced (adjusted $M = 26.324$ versus 27.607) but a drop in the number of words (from 339.2 to 336.1), resulting in their significantly decreased general complexity. The results of these analyses on the two components of general complexity supported the above explanation for the differential effects. The differential effects found on general complexity in this study provide some empirical evidence for the learning effect that TA might give rise to (e.g., Ahlum-Heath & Di Vesta, 1986; Russo et al., 1989), but the finding also suggests that the facilitation effects would be learner- and measure-specific.

The interaction on CVU was shown as a significant deterioration of low CVU achievers’ performance. This finding indicated that the L2 writers, who had not done well in the levels of translating and monitoring processes that were in charge of assigning what were thought to be correct forms to verbs in accordance with time, aspect, modality, voice, etc., or deciding whether verbs should be used anyway, did not allocate their normal share of attention for these undertakings, under the TA condition. This negative effect was possible for these low CVU L2 writers who usually did not pay much attention to CVU as an aspect of performance and did not have a stable
performance in this aspect. Therefore, when TA was added, they would sacrifice this aspect of performance for what they thought were more crucial concerns than correct use of verb forms, for example, expression of meaning, or aspects of accuracy other than CVU, for which they needed to retain attention.

The interaction between baseline Length Two and Condition (found when WMC Two was used as one covariate) was shown in the significant constraining effects of TA on the length measure for those previously high Length Two achievers. ANCOVA analyses using Condition and (Length Two) Grouping as two independent variables, Length One and the number of words deleted, respectively, as a dependent variable, and the corresponding baseline scores as a covariate and WMC Two as another covariate found that the high Length Two performers showed a significant drop in Length One (from adjusted $M = 383.5$ to $347.9$, $p = .014$) and also a drop in the number of words deleted (from adjusted $M = 22.853$ to $22.568$, $p = .941$) under the TA condition. The results indicated that the differential effects of TA on high Length Two performers were primarily attributable to their significantly fewer words written under the TA condition. This finding indicated that while, generally, the length of L2 writing remained unaffected, the length could be affected when the L2 writers had previously written particularly long texts silently. The finding showed that TA was able to constrain the overall planning of particularly “informative” L2 writers who had a lot to write. That length could be affected should be understandable because length was in itself an elastic measure in task completion because any length beyond the minimum word count (300 words, in this case) was acceptable. The high Length Two performers had more room to adjust in response to the addition load of TA so that they could write much fewer words while still meeting the requirement of task completion.
8.4 Reactivity and Working Memory Capacity

Statistically significant interaction between WMC Average/WMC One and Condition was found on organization, and statistically significant interaction between WMC Two and Condition was found on Length Three, Length Four, and Lexical Diversity. The interactions indicated that the effects of TA varied with WMC, but the differential effects were also WMC-specific and measure-specific. WMC Average/WMC One interacted with Condition in that the ability of the low WMC Average/WMC One group to organize texts was significantly impaired whilst that of the high WMC Average/WMC One group remained stable. This finding provides evidence for the positive correlation between available WMC and writing performance and/or writing processes that has been suggested in the literature (e.g., Bergsleithner, 2010; Ellis & Yuan, 2004; Kellogg, 1988, 2001; Ransdell, Arecco & Levy, 2001; see also Olive, 2004 for a review of studies using the dual/triple-task technique). The interaction of WMC Two with Condition, however, showed a different trend. Those with a high level of WMC Two were significantly negatively affected whilst those with a low level of WMC Two were exempted from influence, on Length Three, Length Four, and Lexical Diversity. This finding reveals that the WMC-reactivity issue could be more complicated than being simply unilateral, given that in Goo’s (2010) study, high WMC (Average) L2 readers were also found to be the significantly impaired (in embedded rule learning). Just as Goo suggested, the issue may not be fully understood by referring only to the quantity of WMC, without considering the quality of mental activities.

The low WMC Average (One) L2 writers’ significant sacrifice of organization showed that while TA significantly affected organization for the whole body of participants, they were the most severely affected and the greatest contributor to the overall effects on organization. While organization was an easily ignored aspect of L2
writing performance for these L2 learners in general and their performance on organization was generally not stable, the extent of this negligence and instability appeared particular acute for the low WMC learners. A question that remains to be answered and explored is why WMC Average and WMC One (i.e., the operation span), rather than WMC Two (i.e., the reading span), distinguished the participants in the reactivity on organization. Süß, Oberauer, Wittmann, Wilhelm, and Schulze (2002) revealed that WM explains reasoning. It may be surmised that under the TA condition, the participants with low WMC Average (One) were most likely to ignore the logic between/within sentences and wrote more incoherently.

The significant negative impact of TA on high WMC Two L2 writers’ lexical diversity indicated that while TA significantly impaired lexical diversity for the whole body of participants, the high WMC Two L2 writers were the most severely affected and the greatest contributor to the overall effects on lexical diversity. While lexical diversity was an easily unattended aspect of L2 writing performance due to its perceived peripheral relationship to meaning by these L2 learners in general and, while the L2 learners’ performance on lexical diversity was generally not stable, the likelihood of this negligence under the cognitive pressure of TA and the extent of the instability of performance in lexical diversity appeared particularly acute for the high WMC Two learners. The particular effect of TA on the high WMC Two group was compatible with Goo’s (2010) similar finding that TA significantly hampered high-WMC learners’ secondary rule learning while comprehending a passage. Goo attributed the high-WMC group’s vulnerability to TA to their particular processing mechanism that was able to manoeuvre WM resources in favour of comprehension rather than concomitant rule learning. Similarly, in this study, the high WMC Two L2 writers, in response to the additional task of CVR, may have concentrated on conveying meaning and employed
less central executive thinking over using diverse words, a plus that was considered secondary because repeated words still fulfilled the task of meaning expression. Additionally, the differential effect was related to the particularly unstable state of the high WMC Two L2 writers’ performance in lexical diversity. When the correlations between WMC Two and Lexical Diversity were examined, the highest correlation was found for the high WMC Two group, $R^2 = .055$ ($R^2 = .001$, for the mid and low WMC Two groups), for the baseline $D$ data, and $R^2 = .051$ ($R^2 = .026$, for the mid WMC Two group, and $R^2 = .008$, for the low WMC Two group), for the main task $D$ data. The correlations showed that the high WMC Two L2 writers were the group who had the highest performance in lexical diversity. They could be the group that best accepted the idea of varying words in writing from the lecturing they received, hence the group that exerted the greatest extra effort to care for this aspect of performance. In other words, their performance in this aspect could be the most unstable and adjustable to TA. Admittedly, how WMC Two, the reading span, is related to learning efficiency is an issue worthy of further study.

The WMC Two and Condition interaction regarding Length Three and Length Four was related to that on Lexical Diversity, given the correlation between Lexical Diversity and the number of syllables written (baseline $D$ could explain 6.8% of the variance in baseline Length Three and 5.4% of the variance in baseline Length Four, and main task $D$ could explain 4.1% of the variance in main task Length Three, and 5.0% of the variance in main task Length Four). The decline in Length Three or Length Four caused by the decline in lexical diversity made the effect of TA on Length One for the high WMC Two group (the high WMC Two group wrote 25.367 words less under the TA condition than their counterparts in the NTA condition did, 308.5 versus 333.9, $p = .066$) more salient.
8.5 Completeness of Concurrent Verbal Reports in Comparison with Retrospective Verbal Reports

It should be noted that the CVRs, when examined, were all eligible CVRs, because the participants all kept reporting and none of them intentionally paused unreasonably (e.g., for longer than 9 seconds). Yet, while going blank mentally might occur at certain moments, some thoughts did happen within the reasonable pauses, prior to execution of revisions, but were not reported. By comparison with that reported in RVRs, the information elicited in the CVRs was found not able to contain fully, among others, the information involved in much metacognitive thinking that underlay quite a number of revisions and was supposed to have been heeded. The finding supports previous arguments concerning the incompleteness of CVRs (e.g., Cooper & Holzman, 1983; Dechert, 1987; Dobrin, 1986; Ericsson & Simon, 1979, 1993; Nisbett & Wilson, 1977; Smagorinsky, 1989; Waern, 1988), its seriousness being another matter. It also echoes Williams and Davids’s (1997) finding that soccer players’ verbal reports failed to reflect swift alteration of attention sufficiently while they watched simulations and Barkaoui’s (2011) finding of incomplete CVRs given by essay raters, which did not catch automatic, intuitive, and covert processes and reactions during rating.

The reason why the thoughts were not reported was that they were transitory and hard to verbalize online due to their being higher-order (Nisbett & Wilson, 1977) and to spare much effort to verbalize them would have affected the realization of such primary goals of writing as meaning expression, time and speed, and accuracy, which the participants may have perceived. The nonveridicality could therefore be understood as a choice to preserve nonreactivity (e.g., Ericsson & Simon, 1984). The incomplete reporting of thinking underlying revisions could also be related to the high processing pressure (Durkin, 1937) the participants faced in doing the revisions. There is evidence that significantly more dysfluencies were incurred with a little more time spent in the
TA task, indicating that the revisions may have happened all in a very swift and hasty way.

8.6 Participants’ Perceptions on Reactivity

A majority of the participants answered that, in different degrees, TA affected their thinking negatively. Their answers supported the reactive effects found (on dysfluencies, organization, lexical diversity, etc.) in the quantitative analyses of the performance data. Especially, a considerable portion of the participants mentioned that they failed to change the simple/repetitive words they used. These answers pointed to the significant decline in lexical diversity found in the quantitative analyses. The different degrees of reactivity they perceived also corroborated the differential effects of TA found in the quantitative analyses. Their perceived reactivity confirms Faigley and Witte’s (1981) early observation that many writers found their writing processes disturbed when thinking aloud. Some of their perceptions also resembled the struggles Donald M. Murray acutely experienced when writing with think-alouds (Berkenkotter, 1983). It should be noted that the participants may have overstated the actual effects of TA on their writing, given that their accounts were subjective and sometimes emotional, and that they might tend to “scapegoat” think-alouds for their unsatisfactory writing. However, while their answers should be taken with some caution, their answers do raise the issue of how participants will experience and accept think-alouds affectively.

8.7 Participants’ Perceptions on Completeness and Accuracy

Overall, in answering the completeness question, some of the participants admitted that they were not able (completely) to report hard-to-grasp/verbalize thoughts that were abstract, elusive, and fast-flashing, that flooded into their mind, and that were privacy concerned or irrelevant. Their perceptions on the incompleteness of their CVRs supported the findings obtained through the comparison of their CVRs and RVRs. Their
perceptions also enriched the analogous accounts concerning the incompleteness of CVRs made by essay raters in Barkaoui’s (2011) study and confirmed Williams and Davids’s (1997) finding of the inability of CVRs to capture swift attentional shift.

In reference to Ericsson and Simon’s (1993) framework, both the findings from the comparative study of the participants’ CVRs and RVRs and their perceptions elicited by the reflection questionnaire indicate that CVRs were indeed not able to contain all the information that was heeded, in addition to the information that was not heeded. The types of information that were heeded but not reported, and were identified in the participants’ CVRs were similar to those reported by the participants themselves, both including difficult-to-verbalize thoughts as well as irrelevant thoughts. The types of missing thoughts derived from both sources in this study corroborate and enrich previous predictions in this regard (Durkin, 1937, 1945; Ericsson & Simon, 1979, 1993; Smagorinsky, 1989; Wilson, 1994). The lack of reporting of information that was heeded and could have been verbalized with effort may represent a sacrifice of complete reporting for goals that were considered more important (i.e., time and speed), and could be seen as a choice the participants made to lesson reactivity when faced with difficult-to-verbalize situations (e.g., Ericsson & Simon, 1984). The incomplete reporting of the types of thoughts found in this study was not serious considering that the CVRs did contain a fairly complete record of the sequences of thought and considering that the missing thoughts could be inferred to a certain extent and may be elicited via complementary means such as RVRs, if necessary.

In answering the accuracy question, the majority of the participants replied that their CVRs were accurate and those who thought their CVRs were not did not provide sound reasons. The participants reported what they thought because they followed the
instructions earnestly and verbalized their thoughts that arose in implementing the narrative writing task as these were not of an embarrassing nature.

It should also be noted that although thoughts were somewhat repressed under the TA condition, the CVRs embodied the actual discounted thoughts mostly completely and accurately, except for the few missing episodes. Therefore, the extent of nonveridicality found in this study should not affect the validity of CVRs and their usefulness much.

8.8 Summary

The quantitative and qualitative findings of this study, taken together, indicate that the primary validity issue for CVRs may lie in their reactivity. The findings of this study, together with those of Yang et al.’s (2014) study on reactivity, suggest that previous studies on writing that used the TA methodology to establish writing models (e.g., Flower & Hayes, 1980a; Hayes & Flower, 1986; Zellermayer & Cohen, 1996), to identify the distinctions between skilled and unskilled writers (e.g., Bereiter & Scardamalia, 1987), to differentiate texts of different cognitive demands (e.g., Durst, 1987), to investigate the relationship between cognitive processes and text quality (e.g., Breetvelt et al., 1994), and to study the distributions of cognitive activities or (sub)processes (e.g., Breetvelt et al., 1994; Roca de Larios et al., 2001) may have been based on a somewhat constrained and distorted version of the information otherwise heeded under the silent condition, although the gross procedures or sequences about writing inferred from CVRs in these studies might remain largely credible. The findings of this study, in particular, cautions against the common practice in previous studies that treated participants non-differentially, by asking participants to think aloud as a whole, regardless of their proficiency, working memory, and/or creativity or active state of mind.
CHAPTER 9 CONCLUSIONS, LIMITATIONS, RECOMMENDATIONS FOR FURTHER RESEARCH, AND SUMMARY OF THESIS

Chapter 9 ends this thesis by drawing important conclusions from the discussions I have presented in Chapter 8. It also points out the limitations of this study, and makes recommendations for further research. Chapter 9 finally provides an overall summary of this study.

9.1 Conclusions

Several conclusions can be drawn from this study. First, the reactivity of CVRs for “ill-defined” verbal tasks such as L2 writing may be greater than that for “well-defined” problem-solving tasks. For well-defined tasks, as Ericsson and Simon (1993) and Fox et al. (2011) have concluded, there was no reactivity (on accuracy) except for more time on task or latency. For ill-defined tasks, the reactivity issue becomes more complicated. As Stratman and Hamp-Lyons (1994) have pointed out, in ill-defined tasks, there are multiple routes towards the same goal, which are often unpredictable. My study provides specific evidence for the existence of multiple (almost countless, in fact), unpredictable routes in L2 writing. It further suggests that the routes may differ significantly in terms of the extent to which they are peripheral or central to the expression of meaning and in terms of their stability, resulting in their significantly different requirements of cognitive effort for processing. Regarding the difference in reactivity between a problem-solving task and an L2 writing task, my study indicates that, when multiple routes existed for the ultimate goal of expressing a definite plan, that is, when there were multiple forms to express the same meaning, the participants would give up harder or more cognitively demanding forms and opt for simpler ones or those that minimise the use of cognitive resources, as a short cut, while undertaking the additional task of reporting their thoughts. Specifically, in this experiment, the participants resorted to simpler words or repeated
words that seemed more prototypical to them, to save effort for what they viewed to be more important concerns. That they gave up means of organization, or even that they sacrificed the fluency of their writing (in terms of more dysfluencies), may be understood in the same way. Based on my findings (especially the reactivity on lexical diversity in general and for the high WMC Two group in particular, and the effect on the length for high Length Two L2 writers) and the ill-defined nature of L2 writing, an inference can be made that TA may constrain L2 writers’ creativity, given that creative thoughts represent the most cognitively demanding and are most likely to be sensitive to interference. This may also be true with L1 writers. In fact, an expert L1 writer, Murray, reported being able to write only 17 words within an hour when thinking aloud (Berkenkotter, 1983; see also Murray, 1983). Russo et al.’s (1989) assumption that the learning effects of TA existed especially in tasks that contained multiple options was therefore not corroborated.

Second, my study basically confirms the accuracy and temporal aspects of effects relating to Level 1 and Level 2 verbalizations maintained by Ericsson and Simon (1993) and Fox et al. (2011) but, more importantly, it provides a reinterpretation of the (lack of) effects and suggests exceptions. My finding of no significant effects on time and speed and accuracy in L2 writing, which Yang et al. (2014) also reported, may indicate the workings of an attention allocation system characteristic of L2 writers that prioritise writing speed and accuracy among the multiple dimensions of performance. Besides, the differential effects on the speed of L2 writing and on a certain aspect of L2 writing accuracy (CVU) found in this study suggest that Ericsson and his colleagues’ conclusion of no effects on accuracy and an increase in time on task may not apply to all L2 writers and all aspects of accuracy.
Third, my study suggests the importance of measurement of multiple dimensions of L2 as well as L1 writing performance. As is shown, when performance is measured multi-dimensionally, it is more likely that reactive effects can be found, than when it is measured only in terms of accuracy (sometimes together with some temporal measures), as has usually been done in previous research using problem-solving tasks. That L2 or L1 writing performance can be measured in multi-dimensional ways is a special trait that makes possible a finer examination of the reactive effects and a closer look at the trade-off between different dimensions of L2 or L1 writing performance. In fact, in studies that examined the effects of additional cognitive load on writing performance or processes (see e.g., Olive, 2004 for a review), a multitude of writing measures could have been used to provide more comprehensive insights due to the complications of writing performance or proficiency as a construct.

Fourth, the reactivity of TA in L2 writing may be greater than that in L1 writing, given that many aspects of L2 writing performance are still not in a stable state due to L2 writers’ level of proficiency in the L2 language (e.g., spelling, word choice, organization, or even verb use, as this study has shown), and therefore are liable to external interference. L1 writers may have a firmer mastery of these aspects of writing and they may take many of their forms or decisions as their single or definite options, hence less, if any, reactivity.

Fifth, the overall effects of TA on L2 writing processes did not seem to be serious. It only significantly affected the aspects of performance which were thought (perhaps inappropriately) by the L2 writers to be not essential and were easily discounted/ignored without special attention (i.e., dysfluencies, lexical variety, and organization), and which were linked to peripheral or extra value-adding processes. TA did not affect what L2 writers apparently prioritized, namely, what they perceived as more important concerns
such as content, accuracy, the writing speed, and complexity. It could also be seen that TA did seem to prevent L2 writers from bringing their writing potential into full play, but this withholding effect was only temporary and their writing ability was recoverable. In this experiment, the participants were asked to submit their compositions as soon as they finished them. This arrangement actually deprived them of the opportunity to revise them. That their after-writing reflections became an outlet for their regret about lack of revision was the evidence.

Sixth, the overall effects of TA were negative except that it facilitated the performance of a certain number of high achievers in terms of general complexity and that of a certain number of low achievers in terms of speed. There might be some minor facilitative effects when the L2 writers were considered as a whole (e.g., the insignificant increase in General Complexity and Syntactic Complexity), but these effects may have been overwhelmed by the negative effects. The negative results, as Yang et al. (2014) have pointed out, indicate that among the causes for the reactive effects, negative factors, which include the processing load of verbalization primarily, may have prevailed over positive factors, such as critical attention and reflection (e.g., Jourdenais, 2001; Russo et al., 1989; Stratman & Hamp-Lyons, 1994), or monitoring (e.g., Yanguas & Lado, 2012), so that the effects of TA were shown primarily in terms of constraining writing processes and information processing.

Seventh, TA did affect translating processes at least, as Kellogg (1996) has predicted, by constraining thoughtful translating processes concerning diction, interfering with pre/while-executing monitoring processes to result in more dysfluencies, and inhibiting planning, translating, and monitoring processes to weaken organization.

Then, the effects of TA can be different among L2 writers, but the differential effects may be WMC specific and measure specific. My study indicates that L2 writers with
different dimensions of proficiency or with different WMC Average, WMC One or WMC Two reacted to TA differently in some particular measures. The participants’ answers as to the reactive effects of TA also confirm the distinct effects of TA on them. This conclusion necessitates a re-examination of previous studies or accounts distinguishing skilled from unskilled writers based on the think-aloud data.

Additionally, the finding of this study provides support to Yang et al.’s (2014) proposal of a mechanism in attention allocation among multiple dimensions of L2 writing performance in face of TA or, presumably, in face of any additional cognitive load, a mechanism that operates following the principle of “sacrifice of the most easily ignored and the most cognitively demanding” to maintain the most valued and the most fundamental. This proposal is in fact compatible with the Limited Attentional Capacity Model on task complexity (Skehan, 1998; Skehan & Foster, 1999, 2001), which generally posits that an increase in task complexity causes changes among different aspects of performance.

Finally, the veridicality of CVRs can be quite acceptable, except for some transitory or verbalization-unfriendly thinking processes which may be missing from CVRs.

The findings of this study suggest that previous studies on L2 or even L1 writing models, (sub)processes, writer differences, and/or cognitive demands that relied on CVRs as data may have been based on somewhat constrained, distorted, and incomplete representations of writers’ ongoing thoughts, and may have glossed over the particular effects of TA on writers with particular traits.

Despite their possible reactivity and nonveridicality, CVRs will still remain an important means of data collection in writing research and beyond. As Russo et al. (1989) aptly remarked, “Given the unique benefits, the challenge is to identify and reduce
causes of their invalidity” because “nothing can match the processing insights provided by a verbal protocol” (p. 767).

It is recommended that strict procedures should be followed to ensure that CVRs are administered with the least reactivity and nonveridicality incurred (Cohen, 2000; Ericsson, 2002; Ericsson & Simon, 1993; Green, 1998; Wade, 1990), as has been meticulously done in this study. Green’s (1998) key steps for implementing TA, for example, could be referred to, which highlighted a procedure that included preparation in terms of giving good instructions, briefing, practicing, and offering feedback (p. 41-p. 42). Gu, Hu, and Zhang’s (2005) study could also be consulted. In that study, a one-hour think-aloud training session was administered to primary 1 and 3 children, in which the researchers played games with the young participants to establish rapport and showed them recording devices and played recordings to get them prepared.

It is also recommended that when asking L2 learners, especially those that are new to verbal reporting, to provide CVRs, measures should be taken to ensure that they are trained sufficiently in warm-up trials. However, as this experiment has shown, although it was established during the TA procedure that each and every participant did write and think aloud as well as possible, during the 30-minute think-aloud training, in which they wrote for 15 minutes approximately as a trial, several of them still reported that they felt somewhat nervous when they completed the longer TA task in the main writing task. It is true that training should be individualized, but participants’ reported competency in giving CVRs may not be very reliable. This poses a challenge to think-aloud training: to what extent should participants be trained? As Eysenck and Keane (2000) put it, “the demands that a task makes on attentional or other central resources may be reduced with practice” (p. 137) (see also Levy & Ransdell, 1995). Practicing thinking aloud for a
longer time and several times may help, but such training makes think-aloud a rather costly tool.

Even if CVRs are implemented with great care, the systematic effects found in this study point to the intrinsic cognitive pressure characteristic of TA that may not be avoided by enforcing stringent procedures alone (Russo et al., 1989; Stratman & Hamp-Lyons, 1994). This indicates the importance of measuring the range and the extent of the effects, and also the importance of complementary data.

The general effects found on lexical diversity and the differential effects found on several aspects of performance in this study might suggest differential treatment in training potential participants. However, more importantly, these effects reveal the limitations of TA as a tool to elicit thinking during L2 or even L1 writing processes. It may not be appropriate for certain types of participants, including those whose proficiency (e.g., in terms of CVU) is rather low, whose WMC (e.g., the operation span) is rather low, and especially, those who are very creative, or very enthusiastic about a certain aspect of writing performance. It may strangle the creativity and dampen or extinguish the enthusiasm for writing.

In research design, given that CVRs may not reveal transitory and hard-to-verbalize thoughts, researchers need to think over whether such thoughts are necessary as part of the data they intend to collect. If they are, researchers are advised to consider if there are other measures that can be employed to get the elusive data more accurately and completely. Otherwise, they should adopt complementary measures.

To minimize reactivity and nonveridicality, other real-time or post hoc means could be used, if necessary. In writing research, different types of data, which comprise several or all of the means of CVRs, video-recording, immediate retrospections, interviews, notes, drafts, written texts, stimulated recall, keystroke logging data, and eye-tracking
data may provide the most complete basis for developing insights into writing processes (e.g., Leijten, & Van Waes, 2013; Levy, Marek & Lea, 1996; Ronowicz et al., 2005; Zimmermann, 2000). In fact, CVRs have been used together with RVRs and/or other means of data collection in a number of studies on writing (e.g., Armengol & Cots, 2009; Cohen, 1991; Durst, 1987; Langer, 1986; Manchón et al., 2009; Selfe, 1984; Upton & Thompson, 2001; Uzawa, 1996; Zimmerman, 2000).

9.2 Limitations of this Study and Recommendations for Further Research

This study has several limitations. The reactivity of CVRs in this experiment was operationalized as significant changes in performance. I inferred, based on Kellogg’s (1996) cognitive model of writing, from what changed or did not change in performance indexes, what happened or did not happen in thinking processes. The inferences I made, admittedly, were still indirect.

My study on the completeness of CVRs was also indirect, where I used RVRs as a reference. The RVRs might have suffered from some inaccuracy. Although I tried to treat the RVRs with discretion, I could not tell if I could totally avoid the subjectivity involved in my judgment.

One key assumption of this study concerns the construct of WM, especially, the tripartite distinction of its components. Given that it is still unclear what types of WM resources (the central executive, the visuospatial sketchpad, the phonological loop, or a combination of them) the various writing subprocesses and TA tap, I used, very sparingly, the specific names of the WM components but preferred more general terms such as WM resources, WMC, or attention. The theories concerning WM, its components, and their relationships with L1 and L2 writing processes and TA need to be further researched, cognitively. Besides, I used an operation span and a reading span task to gauge the participants’ WMC. It should be noted that they might not be able to
represent the WMC of a participant sufficiently, given the complexity of this construct. Other tests might yield somewhat different results.

Another limitation as regards the research design could be the small number of participants in each WMC subgroup or in each proficiency subgroup when it came to studying the mediation of WMC or proficiency in reactive effects.

Still another limitation is that I should have managed to measure the participants’ heightened lexical awareness, despite the difficulty of such a measurement in the actual implementation of the experiment, and regardless of the complementary evidence in their reflections that they reported their word choice was affected.

Also, the interaction between reactivity and task type could have been better explored had I included an additional argumentative writing task in this study.

Further research on the validity of CVRs may be conducted in a computer-based writing environment, given that a large amount of writing has been taking place using computers. An advantage of validity research in such a context is that screen capture devices can be used to record writing behaviours more accurately so that the veridicality of CVRs could be better examined in reference to what actually happens with keystrokes. Further research could also include primary or secondary school students to see if they respond to TA differently from university young adults. Since most of the research on reactivity so far has been conducted with novice verbalizers who received training for around half an hour or less, research might be needed that investigates the effects of TA on experienced verbalizers. In that case, the negative reactivity might vanish or positive reactivity might even occur. Validity studies in L1 writing are also needed, especially those using multiple dimensions of measurement. The validity of other types of verbal reports, for example, concurrent metacognitive verbal reports, may also be studied.
9.3 Summary of the Thesis

This study addressed the validity of CVRs in L2 writing research. By employing quantitative and qualitative methods, the study found that CVRs had overall and differential reactive effects while they were basically complete and were accurate. These findings have important ramifications in that they necessitate reflections on previous studies as regards their implementation of the TA technique, the limitations of their verbal data, and especially their use of participants. The findings also raise awareness for the future judicious use of TA in terms of TA training, participant selection, and use of complementary techniques. Empirical studies on the validity of TA in both L2 and L1 writing research have just been initiated, and more endeavors are clearly expected to explore this meaningful area.
REFERENCES


APPENDIXES

Appendix 1 Information sheet and consent form for participants

INFORMATION SHEET
(STUDENT PARTICIPANTS)

Project Title: The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

Researcher: Yang Chengsong

Researcher introduction

My name is Yang Chengsong, a PhD candidate at the School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland. I am conducting research on the reactivity and veridicality of concurrent verbal reports in second language writing.

Project description and invitation

The purpose of the research is to examine whether concurrent verbal reporting (think-aloud) is a valid tool in eliciting second language writers’ process data. The research will involve around 100 undergraduates randomly assigned to different groups. Some of you will write compositions silently, while others may need to write and at the same time report your thoughts. Those of you who are asked to report your thoughts will also be asked to recall some of your thoughts immediately after you complete the writing and think-aloud task and will need to fill in a questionnaire.

I would like to invite you to participate in this research. Your voluntary participation is greatly appreciated.
Your involvement

Once you decide to participate, you will need to fill in a personal information form. You will be asked to take a series of memory tests, and complete a baseline writing task silently. These procedures will be conducted in several language labs simultaneously, or in a single language lab with participants taking the tasks in several batches. Then, you will be assigned to different groups to write another composition. Some of you will conduct the second writing task silently in a large language lab or a classroom. Others of you will be asked to conduct the second writing task in a different language lab or several different language labs while reporting your thinking. Those of you who report your thoughts while writing will also need to report some of your thoughts retrospectively immediately after the writing activity and fill in a questionnaire afterwards. All of you will receive sufficient training before you do the main writing task.

Those of you who are asked to verbalize your thoughts will wear microphone headsets and have your reports recorded by a digital recorder pre-installed into the computers in front of you or by any recording software installed in your own notebooks. You will be instructed so that every one of you is able to operate on the recorder.

The research will take about 3 hours to 3.5 hours. The series of memory tests will take about 1 hour; the baseline writing task will take about 1 hour. The second writing task will take around 1 hour to 1.5 hours. Your recordings will be transcribed.

Compensation

Your benefits for participation will include provision of detailed feedback on essays upon request. They will also be compensated according to the tasks they complete. Participants in the silent group will be paid 75 yuan RMB for their participation, which will include 15
yuan RMB for the working memory tests, 25 yuan RMB for the baseline writing task, 5 yuan RMB for the think-aloud training activity which they will also receive, 25 yuan RMB for the main writing task, and 5 yuan RMB as a completion bonus. Participants in the think-aloud group will be paid 105 yuan RMB for their participation, which will include 15 yuan RMB for the working memory tests, 25 yuan RMB for the baseline writing task, 5 yuan for the think-aloud training activity, 25 yuan RMB for the main writing task, 10 yuan RMB for their additional task of verbalization, 5 yuan RMB for the recording effort, 5 yuan RMB for retrospection, 5 yuan RMB for filling in the questionnaire, and another 10 yuan RMB as a completion bonus.

Due to unpredictable contingencies, this research design and scheme for payment might be subject to minor changes when the experiment is field implemented.

Data storage/retention/destruction/future use

Hard copy data (i.e., your essays, your personal information forms, your consent forms, etc.) will be securely stored in a locked cabinet at the University of Auckland and electronic data will be stored on the researcher’s computer, which is password protected. After six years, all hard copy data will be shredded and the digital information will be deleted. The data collected from the research will be used for the researcher’s PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations. If any one of you would like to have a copy of the research findings at the end, you can indicate this on the consent form, and I will send a summary to you.

Right to withdraw from participation

You are entitled to withdrawal at any time of the study or demand for returning or deletion or exclusion of any data provided by you at any time before 1st January, 2015 without having to give a reason. You have the right not to answer any specific question in the
questionnaire and have the recorder turned off at any stage of the think-aloud experiment. You have the assurance that participation or withdrawal will not affect the normal courses, grades, or relationship with the faculty.

**Anonymity and confidentiality**

You will be randomly given a digital code (a number ranging from 001 to 120) which you will use to mark your compositions, your questionnaires, your recordings, and so on. The researcher himself does not know you have that digital code. The anonymity of non-participants in group situations will also be preserved and students who decline to join may choose to return a blank questionnaire. Confidentiality is assured. Information of the university and the faculty will be disguised. If the information you provide is reported/published, pseudonyms or your unique identification code will be used to protect your identity. No identifying information and data collected from the research will be disclosed to a third party.

Thank you for taking the time to read this information sheet. If you have any inquiries or questions, please do not hesitate to contact me or my supervisors.

Yours sincerely,

Yang Chengsong
### Contact details

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Main supervisor</th>
<th>Co-supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang Chengsong</td>
<td>Associate Professor Lawrence Jun Zhang</td>
<td>Professor Judy Parr</td>
</tr>
</tbody>
</table>

**School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland**, Gate 3, 74 Epsom Ave, Auckland, yche9631@aucklanduni.ac.nz

**Ph:** +64 09 623 8899 ext. 48463

Local contact in China: Ph: +86 13649265604

---

You may also contact the head of the School of Curriculum and Pedagogy, Professor Judy Parr by jm.parr@auckland.ac.nz or +64 09 623 8899 ext. 88998.

For any queries regarding ethical concerns, you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland, 1142. Telephone: 09 373-7599 ext. 83711.

**APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046**
CONSENT FORM (STUDENT PARTICIPANTS)  
(THESE FORM WILL BE HELD FOR A PERIOD OF 6 YEARS)

Project title: The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

Researcher: Yang Chengsong

I have read the Participant Information Sheet, and/or been given a full briefing of the study, including its nature and procedures, and understood what is involved in the research and why I have been asked for a consent. I have had the opportunity to ask questions and have them answered satisfactorily.

- I understand the purpose and process of the research project and my involvement in it.
- I agree to participate in this research.
- I understand that participation is voluntary.
- I understand that I am allowed to withdraw at any time of the experiment, and that after the experiment I may still request my data to be withdrawn until 1 Jan 2015.
- I may be audiotaped/vidiotaped during the study.
- I understand that participants' recordings may be transcribed by the researcher, but will only be accessible to the researcher and will not be disclosed to any third party.
- I understand that if participants' recordings are transcribed by a research assistant, she/he will not be given any personal information of participants, and she/he will
keep the contents of recordings confidential and upon completion of transcription, will delete all the relevant materials from her/his work station.

- I understand that I may require a copy of the transcript of my recording, or a digital file of my recording.
- I understand that if I request a digital file of my recording, it may take time before my request can be met because of the researcher spending most of his time in New Zealand and because of difficulty in transmitting large digital files over the internet.
- I understand that participation, non-participation or withdrawal will not affect my grade, academic performance, and relationship with the teacher and the faculty.
- I assure that there is no grade evaluation on the writing.
- I understand that I will be asked to use a randomly drawn digital code to identify my compositions, recordings, questionnaire, and so on, and that only I know the code corresponds to me myself.
- I understand that hard copies and digital data will be stored separately and securely for a period of six years and then destroyed.
- I understand that the data collected from the research will be used for the researcher’s PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations.
- I understand that if the information provided by participants is reported/published, anonymity is assured and pseudonyms will be used to protect their identity.
- I understand that no identifying information will be disclosed to a third party or the public. I understand that the ethical aspects of the project have been approved by the ethics committee of the University of Auckland.
- I hope to have feedback of my compositions sent to my email address: ____________ (If you don’t want the researcher’s feedback to be sent, keep this blank.)
I hope to have the transcript of my recording sent to my email address if it is available: ______________ (If you don't want it, leave this blank).
I wish to receive a copy of the research findings sent to my email address: ______________ (If not, keep it blank).

I therefore give my informed consent to participation of the research project (“The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing”).

Name __________________________
Signature _______________________
Date ___________________________
Email address ____________________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046
Appendix 2 The University of Auckland Human Participants Ethics Committee approval letter

Office of the Vice-Chancellor
Research Integrity Unit

UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE

19-Sep-2013

MEMORANDUM TO:

Dr Lawrence Zhang
Curriculum & Pedagogy

Re: Application for Ethics Approval (Our Ref. 10046)

The Committee considered your application for ethics approval for your project entitled The Reactivity, Veridicality, and Utility of Concurrent Verbal Reports in Second Language Writing.

Ethics approval was given for a period of three years.

The expiry date for this approval is 19-Sep-2016.

If the project changes significantly, you are required to submit a new application to UAHPEC for further consideration.

In order that an up-to-date record can be maintained, you are requested to notify UAHPEC once your project is completed.

The Chair and the members of UAHPEC would be happy to discuss general matters relating to ethics approvals if you wish to do so. Contact should be made through the UAHPEC Ethics Administrators at humanethics@auckland.ac.nz in the first instance.

All communication with the UAHPEC regarding this application should include this reference number: 10046.

(This is a computer generated letter. No signature required.)

UAHPEC Administrators
University of Auckland Human Participants Ethics Committee

c.c. Head of Department / School, Curriculum & Pedagogy
Mr Chenguang Yang
Prof Judith Parr
Appendix 3 Information sheet and consent form for dean and teachers

**INFORMATION SHEET (DEAN)**

**Project Title:** The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

**Researcher:** Yang Chengsong

**Researcher introduction**

My name is Yang Chengsong, a PhD candidate at the School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland. I am conducting research on the reactivity and veridicality of concurrent verbal reports in second language writing.

**Project description and invitation**

The purpose of the research is to examine whether concurrent verbal reporting (or think-aloud, TA) is a valid tool in eliciting second language learners’ process data. The validity of concurrent verbal reporting concerns two facets of the instrument, that is, its reactivity and veridicality. The reactivity concerns whether the acts of simultaneous reporting might serve as an additional task altering the very thinking processes they are supposed to represent and keep intact. The veridicality relates to whether think-aloud protocols (TAPs) represent participants’ true and complete thinking processes. The research will involve around 100 undergraduates randomly assigned to two groups, control (no think-aloud, NTA), and nonmetacognitive think-aloud (TA, i.e., think-aloud without having to justify). All participants will conduct a series of working memory tests and a baseline writing task first, before they complete a similar writing task under different conditions. Their essays will be analysed, their verbal reports will be listened to, and a portion of the reports will be transcribed and coded. The reactive effects of TA will be operationalized as between-group
differences in L2 writing performance measured in terms of fluency, complexity, accuracy, and overall quality of writing, and will be examined in general and in relation to individual differences, including L2 writing proficiency and working memory capacity. The veridicality of TA will be operationalized as inconsistencies between the information elicited by TA and that elicited by a recall task immediately following the TA task, which asks participants in this particular group (TA) to state what they thought when they made revisions. The reactivity and veridicality of TA will also be explored qualitatively by administering a follow-up questionnaire survey asking participants questions concerning the effects, accuracy and completeness of their reports.

I would like to invite undergraduate students to participate in this research. Your permission for me to contact them and for the relevant information to be distributed is being sought.

**Teacher involvement**

The project has adopted a pretest-intervention-posttest design. For recruitment of participants, teachers or administration staff might be involved in circulating research information, that is, Participant Information Sheets (PISs), advertisements, and Consent Forms (CFs), to students.

**Student involvement**

All participants will be asked to fill in a personal information form, then take the series of working memory tests, and then the baseline writing task, which will take two hours approximately. These procedures will be conducted in several language labs simultaneously, or in a single language lab with the participants taking the tasks in several batches, if more labs are not available. Then, all the participants will be assigned to one of three WMC groups (high, mid, low) based on their performance on the working memory
span tasks. After that, the participants in each WMC group will be randomly assigned to one of two experimental conditions: NTA or TA. For the main experiment, those in the NTA group will conduct the main writing task silently in a large language lab or a classroom. Those in the TA group will be asked to conduct the writing task while reporting their thinking in a different language lab. Before they start to think aloud, they will receive instructions as to how to conduct thinking aloud, and then will be asked to practice writing a small paragraph until all of them think that they are able to think aloud competently. All of them wear microphone headsets and their reports will be recorded by a digital recorder pre-installed into the computers in front of them (or their own notebooks). They will be instructed so that every one of them is able to operate on the recorder. After the think-aloud writing task, the TA group will be asked to reflect on what they thought when they made corrections. They will also be required to fill in a questionnaire asking them to answer a series of questions primarily concerning the reactivity and veridicality of their think-alouds. The participants' recordings will be transcribed.

Students' benefits for participation will include provision of detailed feedback on essays upon request. They will also be compensated according to the tasks they complete. Participants in the silent group will be paid 75 yuan RMB for their participation, which will include 15 yuan RMB for the working memory tests, 25 yuan RMB for the baseline writing task, 5 yuan RMB for the think-aloud training activity which they will also receive, 25 yuan RMB for the main writing task, and 5 yuan RMB as a completion bonus. Participants in the think-aloud group will be paid 105 yuan RMB for their participation, which will include 15 yuan RMB for the working memory tests, 25 yuan RMB for the baseline writing task, 5 yuan for the think-aloud training activity, 25 yuan RMB for the main writing task, 10 yuan RMB for their additional task of verbalization, 5 yuan RMB for the recording effort, 5 yuan RMB for retrospection, 5 yuan RMB for filling in the questionnaire, and another 10 yuan RMB as a completion bonus.
Due to unpredictable contingencies, this research design and scheme for payment may be subject to minor changes when the experiment is field implemented.

Faculty involvement

Your permission and cooperation are the prerequisite for conducting this project. I am expecting your permission to get access to EFL teachers and students. Initially, I am seeking your consent to approach EFL teachers or staff to explain the research to them and get help from them to spread the research information to students and deliver the Participant Information Sheets (PISs), the advertisement, and the Consent Forms (CFs). Later I will return to collect these forms from your site. Secondly, I would like to request the use of classrooms and a/several language lab(s) for the intervention. All the procedures will be discussed in detail with potential participants. I would also like to get your assurance that participation or non-participation will not affect teachers' or students' career, future employment, grade, and academic performance in any way. Additionally, I would like to give my assurance that those who do not participate will not be disadvantaged in any way.

Data storage/retention/destruction/future use

Hard copy data (i.e., student essays, students’ personal information forms, their consent forms, etc.) will be securely stored in a locked cabinet at the University of Auckland and electronic data will be stored on the researcher’s computer, which is password protected. After six years, all hard copy data will be shredded and the digital information will be deleted. The data collected will be used for the researcher’s PhD thesis at the University of Auckland, for academic publications, and for conference presentations. If any participant would like to have a copy of the research findings at the end, he/she can indicate this on the consent form, and I will send a summary to him/her.
Right to withdraw from participation

Once involved, participants are entitled to withdrawal at any time of the study or demand for returning or deletion or exclusion of any data provided by them at any time before 1\textsuperscript{st} January, 2015 without having to give a reason. They have the right not to answer any specific question in the questionnaire and have the recorder turned off at any stage of the think-aloud experiment. Their withdrawal will not affect their normal course taking, grades, or relationship with the teacher and the faculty.

Anonymity and confidentiality

Each participant will be given a digital code ranging from 001 to 120 as his/her identification. The participants will use this code throughout the experiment to mark their essays, questionnaires, recordings, and so on. All participants' information will be kept separately from the data and will only be known to the researcher. The anonymity of non-participants in group situations will be preserved and students who decline to join may choose to return a blank questionnaire. Confidentiality is assured. Information of the university and the faculty will be disguised. If the information participants provide is reported/published, pseudonyms or the unique identification name will be used to protect their identity. No identifying information and data collected from the research will be disclosed to a third party.

Thank you for taking the time to read this information sheet. If you have any inquiries or questions, please do not hesitate to contact me or my supervisors.

Yours sincerely,

Yang Chensong
Contact details

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Main supervisor</th>
<th>Co-supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang Chengsong</td>
<td>Associate Professor Lawrence Jun Zhang</td>
<td>Professor Judy Parr head of the School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland</td>
</tr>
<tr>
<td>School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland</td>
<td>Associate Dean International Strategic Engagement, School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland</td>
<td></td>
</tr>
<tr>
<td>Gate 3, 74 Epsom Ave, Auckland</td>
<td>Gate 3, 74 Epsom Ave, Auckland</td>
<td></td>
</tr>
<tr>
<td><a href="mailto:ychen63@audlantic.unl.ac.nz">ychen63@audlantic.unl.ac.nz</a> Ph: +64 09 623 8899 ext. 48463</td>
<td><a href="mailto:lj.zhang@audlantic.unl.ac.nz">lj.zhang@audlantic.unl.ac.nz</a> Ph: +64 09 623 8899 ext. 48750</td>
<td></td>
</tr>
<tr>
<td>Local contact in China Ph:+86 13645926904</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

You may also contact the head of the School of Curriculum and Pedagogy, Professor Judy Parr by jm.parr@audlantic.unl.ac.nz or +64 09 623 8899 ext. 88998.

For any queries regarding ethical concerns, you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland, 1142. Telephone: 09 373-7599 ext. 83711.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046
CONSENT FORM (DEAN)
(THIS FORM WILL BE HELD FOR A PERIOD OF 6 YEARS)

Project title: The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

Researcher: Yang Chengsong

I have read the Participant Information Sheet, and understood the nature of the research and why I have been asked for consent. I have had the opportunity to ask questions and have them answered satisfactorily.

- I agree to provide research sites.
- I agree to allow the researcher to approach teachers or administration staff to explain the research and ask for their help.
- I agree to allow teachers or administration staff to help with this research.
- I agree to allow undergraduate students to join this research.
- I understand that participation is voluntary.
- I understand that participants are allowed to withdraw at any time of the experiment, and that after the experiment they may still request their data to be withdrawn until 1 Jan 2015.
- I assure that participation, non-participation or withdrawal will not affect career, employment, grades, academic performance, and relationship with the faculty.
- I assure that there is no grade evaluation on the writing intervention.
• I understand that each student participant will be asked to write a unique identification code on his/her essay, questionnaire and all other materials they are required to provide or complete, which can only be recognized by themselves.
• I understand that participants’ verbal reports will be recorded, but only with their consent.
• I understand that participants’ recordings may be transcribed by the researcher, but the recordings and the transcriptions will only be accessible to the researcher and will not be disclosed to any third party.
• I understand that if participants’ recordings are transcribed by a research assistant, she/he will not be given any personal information of participants, and she/he will keep the contents of recordings confidential and upon completion of transcription, will delete all the relevant materials from her/his work station.
• I understand that hard copies and digital data will be stored separately and securely for a period of six years and then destroyed.
• I understand that the data collected from the research will be used for the researcher’s PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations.
• I understand that if the information provided by participants is reported/published, anonymity is assured and pseudonyms or digital codes will be used to protect their identity.
• I understand that the information about the university and faculty will be disguised.
• I understand that no identifying information will be disclosed to a third party or the public.
• I wish to receive a copy of the research findings by email __________________ (If not, keep it blank).
I therefore give my informed consent for the research project ("The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing") to be carried out in our faculty and give permission for you to approach the EFL teachers in our faculty to request their assistance in organizing this research.

Name __________________________
Signature _________________________
Date _____________________________
Email address _______________________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046
INFORMATION SHEET
(TEACHERS)

Project Title: The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

Researcher: Yang Chengsong

Researcher introduction
My name is Yang Chengsong, a PhD candidate in the School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland. I am conducting research on the reactivity and veridicality of concurrent verbal reports in second language writing.

Project description and invitation
The purpose of the research is to examine whether concurrent verbal reporting (or think-aloud, TA) is a valid tool in eliciting second language learners’ process data. The validity of concurrent verbal reporting concerns two facets of the instrument, that is, its reactivity and veridicality. The reactivity concerns whether the acts of simultaneous reporting might serve as an additional task altering the very thinking processes they are supposed to represent and keep intact. The veridicality relates to whether think-aloud protocols (TAPs) represent participants’ true and complete thinking processes. The research will involve around 100 undergraduates randomly assigned to two groups, control (no think-aloud, NTA), and nonmetacognitive think-aloud (TA, i.e., think-aloud without having to justify). All participants will conduct a series of working memory tests and a baseline writing task first, before they complete a similar writing task under different conditions. Their essays will be analysed, their verbal reports will be listened to, and a portion of the reports will be transcribed and coded. The reactive effects of TA will be operationalized as between-group differences in L2 writing performance measured in terms of fluency, complexity, accuracy, and overall quality of writing, and will be examined in general and in relation to individual...
differences, including L2 writing proficiency and working memory capacity. The veridicality of TA will be operationalized as inconsistencies between the information elicited by TA and that elicited by a recall task immediately following the TA task, which asks participants in this particular group (TA) to state what they thought when they made revisions. The reactivity and veridicality of TA will also be explored qualitatively by administering a follow-up questionnaire survey asking participants questions concerning the effects, accuracy, and completeness of their reports.

I would like to invite undergraduate students to participate in this research. Your permission for me to contact them and for the relevant information to be distributed is being sought.

Teacher involvement
The project has adopted a pretest-intervention-posttest design that will be administered during a weekend. For recruitment of participants, teachers are involved in helping circulate research information, the Participant Information Sheets (PISs), the advertisement, and the Consent Forms (CFs) to their students.

Student involvement
All participants will be asked to fill in a personal information form, then take the series of working memory tests, and then the baseline writing task, which will take two hours approximately. These procedures will be conducted in several language labs simultaneously, or in a single language lab with the participants taking the tasks in several batches, if more labs are not available. Then, all the participants will be assigned to one of three WMC groups (high, mid, low) based on their performance on the working memory span tasks. After that, the participants in each WMC group will be randomly assigned to one of two experimental conditions: NTA or TA. For the main experiment, those in the NTA group will conduct the main writing task silently in a large language lab or a classroom. Those in the TA group will be asked to conduct the writing task while reporting
their thinking in a different language lab. Before they start to think aloud, they will receive instructions as to how to conduct thinking aloud, and then will be asked to practice writing a small paragraph until all of them think that they are able to think aloud competently. All of them wear microphone headsets and their reports will be recorded by a digital recorder pre-installed into the computers in front of them (or their own notebooks). They will be instructed so that every one of them is able to operate on the recorder. After the think-aloud writing task, the TA group will be asked to reflect on what they thought when they made corrections. They will also be required to fill in a questionnaire asking them to answer a series of questions primarily concerning the reactivity and veridicality of their think-alouds. The participants’ recordings will be transcribed.

Students’ benefits for participation will include provision of detailed feedback on essays upon request. They will also be compensated according to the tasks they complete. Participants in the silent group will be paid 75 yuan RMB for their participation, which will include 15 yuan RMB for the working memory tasks, 25 yuan RMB for the baseline writing task, 5 yuan RMB for the think-aloud training activity which they will also receive, 25 yuan RMB for the main writing task, and 5 yuan RMB as a completion bonus. Participants in the think-aloud group will be paid 105 yuan RMB for their participation, which will include 15 yuan RMB for the working memory tasks, 25 yuan RMB for the baseline writing task, 5 yuan for the think-aloud training activity, 25 yuan RMB for the main writing task, 10 yuan RMB for their additional task of verbalization, 5 yuan RMB for the recording effort, 5 yuan RMB for retrospection, 5 yuan RMB for filling in the questionnaire, and another 10 yuan RMB as a completion bonus.

Due to unpredictable contingencies, this research design and scheme for payment may be subject to minor changes when the experiment is field implemented.
Data storage/retention/destruction/future use

Hard copy data (i.e., student essays, students’ personal information forms, their consent forms, etc.) will be securely stored in a locked cabinet at the University of Auckland and electronic data will be stored on the researcher’s computer, which is password protected. After six years, all hard copy data will be shredded and the digital information will be deleted. The data collected from the research will be used for the researcher’s PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations. If any one of the participants would like to have a copy of the research findings at the end, he/she can indicate this on the consent form, and I will send a summary to him/her.

Right to withdraw from participation

Once involved, participants are entitled to withdrawal at any time of the study or demand for returning or deletion of any data provided by them at any time before 1st January, 2015 without having to give a reason. They have the right not to answer any specific question in the questionnaire and have the recorder turned off at any stage of the think-aloud experiment. The Dean and the teachers will be required to give an assurance that participation or withdrawal will not affect the normal courses, grade, or relationship with the faculty.

Anonymity and confidentiality

Participants will be asked to write a unique identification code, which can only be recognized by themselves. Each participant will be given a code with numbers ranging from 001 to 120 and a list will be maintained to link participants, via unique identification name with the questionnaires and writing samples. This information will be kept separately from the data and will only be known to the researcher. The anonymity of non-participants in group situations will be preserved and students who decline to join may
choose to return a blank questionnaire. Confidentiality is assured. Information of the university and the faculty will be disguised. If the information participants provide is reported/published, pseudonyms or the unique identification name will be used to protect their identity. No identifying information and data collected from the research will be disclosed to a third party.

Thank you for taking the time to read this information sheet. If you have any inquiries or questions, please do not hesitate to contact me or my supervisors.

Yours sincerely,

Yang Chengsong

Contact details

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Main supervisor</th>
<th>Co-supervisor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang Chengsong</td>
<td>Associate Professor Lawrence Jun Zhang</td>
<td>Professor Judy Parr</td>
</tr>
<tr>
<td>School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland, Gate 3, 74 Epsom Ave, Auckland. <a href="mailto:ychent63@aucklanduni.ac.nz">ychent63@aucklanduni.ac.nz</a> Ph: +64 9 623 8888 ext. 48463</td>
<td>head of the School of Curriculum and Pedagogy, Faculty of Education, The University of Auckland, Gate 3, 74 Epsom Ave, Auckland. <a href="mailto:jm.parr@auckland.ac.nz">jm.parr@auckland.ac.nz</a> ph: +64 9 623 8888 ext. 88988.</td>
<td></td>
</tr>
</tbody>
</table>
You may also contact the head of the School of Curriculum and Pedagogy, Professor Judy Parr by jm.parr@auckland.ac.nz or +64 09 623 8899 ext. 88998.

For any queries regarding ethical concerns, you may contact the Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Office of the Vice Chancellor, Private Bag 92019, Auckland, 1142. Telephone: 09 373-7599 ext. 83711.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046
CONSENT FORM (TEACHERS)
(THIS FORM WILL BE HELD FOR A PERIOD OF 6 YEARS)

Project title: The Reactivity and Veridicality of Concurrent Verbal Reports in Second Language Writing

Researcher: Yang Chengsong

I have read the Participant Information Sheet, and understood the nature of the research and why I have been asked for a consent. I have had the opportunity to ask questions and have them answered satisfactorily.

- I agree to help with this research, by circulating materials, and by helping with other matters relating to recruitment.
- I agree to allow undergraduate students to join this research.
- I understand that participation is voluntary.
- I understand that participants are allowed to withdraw at any time of the experiment, and that after the experiment they may still request their data be withdrawn until 1 Jan 2015.
- I assure that participation, non-participation or withdrawal will not affect career, employment, grade, academic performance, and relationship with the faculty.
- I assure that there is no grade evaluation on the writing intervention.
- I understand that student participants will be asked to write a unique identification code, which can only be recognized by themselves on questionnaire and a pseudonym on all items involved.
- I understand that hard copies and digital data will be stored separately and securely for a period of six years and then destroyed.
• I understand that the data collected from the research will be used for the researcher’s PhD thesis at the University of Auckland, and may be used for academic publications, and conference presentations.
• I understand that if the information provided by participants is reported/published, anonymity is assured and pseudonyms will be used to protect their identity.
• I understand that no identifying information will be disclosed to a third party or the public.

I therefore agree to help with the research to be carried out in my faculty and with my students by providing assistance in organizing this research.

Name ________________________________
Signature ______________________________
Date ________________________________
Email address __________________________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 19 SEPTEMBER, 2013 FOR 3 YEARS, PROTOCOL NUMBER 10046
### Appendix 4 A list of Chinese four-character sentences used in the Reading Span Task

<table>
<thead>
<tr>
<th>组</th>
<th>练习组</th>
<th>按题目顺序判断（打√或X），有几个句填几格</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>狗逗耗子√</td>
<td>猫有九命√</td>
</tr>
<tr>
<td>2</td>
<td>螃蟹有壳√</td>
<td>蝴蝶振翅√</td>
</tr>
<tr>
<td>3</td>
<td>蟹有双钳√</td>
<td>羊生四蹄√</td>
</tr>
<tr>
<td>4</td>
<td>练习组</td>
<td>猫咩咩叫X</td>
</tr>
<tr>
<td>5</td>
<td>练习组</td>
<td>虾头长角X</td>
</tr>
<tr>
<td>6</td>
<td>练习组</td>
<td>马长鬃毛√</td>
</tr>
<tr>
<td>7</td>
<td>练习组</td>
<td>苍蝇能飞√</td>
</tr>
<tr>
<td>8</td>
<td>练习组</td>
<td>鱼能耕田√</td>
</tr>
<tr>
<td>9</td>
<td>练习组</td>
<td>猪啃白菜√</td>
</tr>
<tr>
<td>10</td>
<td>练习组</td>
<td>狮子喵喵X</td>
</tr>
<tr>
<td>11</td>
<td>练习组</td>
<td>蜂蜜长啸X</td>
</tr>
<tr>
<td>12</td>
<td>练习组</td>
<td>鸟飞枝头√</td>
</tr>
<tr>
<td>组 11</td>
<td>老虎飞翔 (\times)</td>
<td>麻雀吞象</td>
</tr>
<tr>
<td>-------</td>
<td>------------------</td>
<td>----------</td>
</tr>
<tr>
<td>组 12</td>
<td>害群之狗 (\times)</td>
<td>鸭子呱呱  (\sqrt)</td>
</tr>
<tr>
<td>组 13</td>
<td>驴声唧喳 (\times)</td>
<td>蛇有四肢  (\times)</td>
</tr>
<tr>
<td>组 14</td>
<td>虾能吼叫 (\times)</td>
<td>蛙子鸣鸣  (\sqrt)</td>
</tr>
<tr>
<td>组 15</td>
<td>虎假狐威 (\times)</td>
<td>狗兔三窟  (\sqrt)</td>
</tr>
<tr>
<td>组 16</td>
<td>麻雀食木 (\times)</td>
<td>鼠有百耳  (\times)</td>
</tr>
<tr>
<td>组 17</td>
<td>鼠目寸光  (\sqrt)</td>
<td>狮子威猛  (\sqrt)</td>
</tr>
<tr>
<td>组 18</td>
<td>猪爱哼哼  (\sqrt)</td>
<td>蟹飞于天  (\times)</td>
</tr>
</tbody>
</table>
## Appendix 5 A list of Chinese characters used in the Reading Span Task

<table>
<thead>
<tr>
<th>Group</th>
<th>Characters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Practice Group 1</td>
<td>请回忆</td>
</tr>
<tr>
<td>Practice Group 2</td>
<td>利成</td>
</tr>
<tr>
<td>Practice Group 3</td>
<td>安如经没</td>
</tr>
<tr>
<td>Group 1</td>
<td>机知的所而此</td>
</tr>
<tr>
<td>Group 2</td>
<td>这明</td>
</tr>
<tr>
<td>Group 3</td>
<td>会物国</td>
</tr>
<tr>
<td>Group 4</td>
<td>在相全使同</td>
</tr>
<tr>
<td>Group 5</td>
<td>别者现每受色把事</td>
</tr>
<tr>
<td>Group 6</td>
<td>作再</td>
</tr>
<tr>
<td>Group 7</td>
<td>何快金任男妈师变吃</td>
</tr>
<tr>
<td>Group 8</td>
<td>多和花许时</td>
</tr>
<tr>
<td>Group 9</td>
<td>交终林医制决传画运则</td>
</tr>
<tr>
<td>Group 10</td>
<td>先名问当</td>
</tr>
<tr>
<td>Group 11</td>
<td>光法年自些回</td>
</tr>
<tr>
<td>Group 12</td>
<td>远条呢始达完空求非</td>
</tr>
<tr>
<td>Group 13</td>
<td>定次后</td>
</tr>
<tr>
<td>Group 14</td>
<td>其并表但间员却好</td>
</tr>
<tr>
<td>Group 15</td>
<td>找罗欢吗约各即合论该</td>
</tr>
<tr>
<td>Group 16</td>
<td>西更拉应直字</td>
</tr>
<tr>
<td>Group 17</td>
<td>场报告至放关告</td>
</tr>
<tr>
<td>Group 18</td>
<td>英岁军往</td>
</tr>
</tbody>
</table>
Appendix 6 The instructions for the reading span task

In this part, you will read 108 4-character sentences in Chinese on the screen. All of them are about animals (e.g., their features, abilities, living habits, etc.). Every sentence is followed by a Chinese character so there are 108 Chinese characters as well. The sentences and characters are presented in sets of two, three, four, five, six, seven, eight, nine, and ten. There are two sentences for each set length and there are altogether 18 sets. All sets will be presented randomly. Whenever you read a sentence in a set on each slide, please immediately judge its semantic appropriateness according to your common sense knowledge and put either a tick (✓) on your answer sheet if you think it is a true statement or put a cross (×) if you think it is wrong. After each sentence, you will see a Chinese character. Please try to keep the Chinese character in memory (Don’t write it down! Otherwise you would miss the next sentence judgment task!). Then you will immediately see another sentence on the screen. Judge its appropriateness, and remember the Chinese character that follows it. Repeat this process with any sentence and Chinese character that may appear. At the end of each set, you will see a slide asking you to recall and write down all the Chinese characters that have been presented in this set (请回忆, meaning Please recall). Then when you see this slide, please write down all the Chinese characters in the set in the order they have been presented to you on the back side of the answer sheet. You should come back to the front side when you finish recalling since a new set of sentences will appear again. Please follow the above procedure until you finish all the sets. The end of this task will be marked by a slide indicate the completion of the test with a thank-you note (i.e., 本部分测试已经结束，谢谢, meaning the test has ended; thank you).
Appendix 7 Coding Schemes for Syllables

1. Code syllables according to British pronunciations.
2. Count misspellings basically as they are, e.g., toliet (toilet misspelt, 2 syllables, in reference to believe, achieve), and opinion (opinion misspelt, 4 syllables, in reference to onion).
3. Count clusters of consonants that likely form syllables as syllables in misspellings with missing vowels, e.g., themselves (themselves misspelt, still 2), hardly (hardly misspelt, still 2).
4. Count abbreviations as they are. Count USA or VCD as one syllable, count can’t and don’t as one, but count couldn’t, didn’t, or doesn’t, as having 2 syllables.
5. Count numbers as one, e.g., 1980 as one, CET4 or C102 as two, one for the letter(s) and the other for the digit(s), and 2010.5.7. as three.
6. Count Chinese pinyin separately, according to the Chinese pronunciation rules, e.g., Rujiawenhua (4 syllables), and Jiangxi (2 syllables).
7. Count slash-divided words separately.
Appendix 8 Coding Schemes for T-Units

Note: The numbers in the brackets indicate the total number of T units coded.
1. A sentence fragment that could have belonged to a previous sentence as a component was counted together with the sentence as a whole.
   a. Time problem is one reason, and there are some other reasons. For example, the gap between the old and the young. (2)
   b. First, strict and practical measures should be taken to change the situation. For example, to encourage Chinese students to be active and give them some appealing awards. (1)
   c. I hate being here. Because it’s so hot. (1)
   d. Today, I have many thoughts want to talk to you, about my past, present, and future. Because I realize that if we want to make your dreams come true, the most important thing we should do is think about our past, live in today, and make plan to future. (1)
   e. Although the life is hard, he got into college by his hard work. After which, his life became richer and richer. (1)
   f. In the future, I think he will focus on the healthy condition of his body. Because, he used to drink wine and eat friend food too much. (1)
   g. I know in those years a students may be faced with inadequate foods or instruments. While lack of mental fortune never came up. (1)
   h. I feel a little uncomfortable getting alone with them. Some times nervous. (1)
   i. We try all kinds of living patterns, in some words, our lives begin. Learning subjects I like, exercising, playing computer games and so on. (1)
   j. I want one, too. A girl that is so easy to be taken out, to hook around, to have hands together, to play some games which is not available when alone. (1)
   k. I still wake up at 6 o’clock, just like before, and go to college. Learn, practice, do research in the lab or stay with my girl. (1)
   l. He will like to show others the process of cooking. To cook for them and amuse them at the same time. (1)
2. Although ... but constructions or similar attempts were coded as one T-unit. Even used instead of even if or even when to lead a clause was still treated as not making an extra T-unit.
   a. For those, although they have time to be together, but they won’t enjoy the time together. (1)
   b. Even they come back home, they just stay to have a dinner. (1)
   c. For instance, I usually playing basketball with my friends for a whole afternoon even the weather as very hot which was impossible for me now. (1)
   d. Although there are also a lot of exams and classes even homework just makes me think I am is another middle school. But to be honest, the pressure here is much less than both of my two middle schools. (1)
   e. My mom is a simple woman, no matter it is in the past, now, or in the future. (1)
3. Coordination relating to the main clause or the subordinate clause was coded contextually.
   a. If you call your mother, she will be happy and she will forgive you. (1)
   b. If we don’t take measure now, I think the relationship will get worst and worst and we will be even indifferent to friends, brother, parents. (1)
   c. Although the condition is not good enough at the moment, we can make great changes if we all try our best and the students will make progress on their behavior. (2)
   e. If it didn’t rain, we would go, and how I wish we could enjoy a good weekend! (2)
f. Yes, it’s a bit hard, but if you know more about what others have in mind, and you stand in their shoes, it might be much easier. (2)
g. Nevertheless, it gives the family members some kind of information that if you have some problems, just call me and I will tell you what to do on the phone. (1)

4. Sentences in direct quotations acting as a clause were counted as separate T-units except the first one; when direct quotations were made as the objects of a preposition, they were counted separately.
   a. Asked the question that “Do you love your families? Will you live with them when they are old” almost everyone answered “Yes”, (2)
   b. he might only say some words like “Don’t be this next” and “You are wrong, correct it next time” at most. (4)
   c. Some young couples refuse drop in on parents’, with a simple word of “I am busy.” (2)

5. Run-on sentences were coded as two T-units, even when there was not a comma in between.
   a. Do what you can think to make it a better family, you will have a better life. (2)
   b. Sometimes, the quality of their teachers are not desirable, such as many teachers in Shaanxi Province give lessons with dialect. (2)

6. When run-on sentences followed a subordinate clause of object, they were usually coded separately from the previous clause/sentence.
   a. He always told me, life is full of fun, Sometime when we pursue something that we think it is important, don’t forget to enjoy life. (2)
   b. When I compare my present with past, I feel that I am richer in knowledge that the past, I have more friends and a more aboard field. (2)

7. An elliptical sentence or its (wrong/unsuccessful) resemblance was coded as one T-unit (e.g., Why not? But why? Why come here? How about now? No pains, no gains. Just a joke. Last but not least, my future life.).
   a. The life of a man is made up of three parts as follows. First, the past that is only your memory. Second, the present you have to meet and should make every effort to work. Finally, the future which you always dream. (4 T-units as well as finite clauses)

For the above construction groups, an alternative coding method might be to take the listings as fragments and code them together with the first sentence. In that case there would be one T-unit only, which could perhaps unduly increase the complexity indicators involved greatly. This was why a cautious approach was adopted in coding noun phrases in listings.

8. When Dash (or any similar punctuation mistaken for it) separated two sentences, they were counted separately.

9. In an independent construction where a nonfinite verb was used instead of a finite one, a T-unit was counted.
   a. I sleeping, watching TV and playing games everyday. (1)
Appendix 9 Coding Schemes for finite clauses

Note: The numbers in the brackets indicate the total number of finite clauses coded. Finite verb forms in a compound predicate, in tag questions, or led by more than, other than, rather than, or instead of, correctly or wrongly, were not coded as additional finite clauses.

a. He went to a music school part time at eight and learned to sing. (1)
b. Only in sometimes, he can understand something I did not only insist on his decision. (2)
c. And people around appreciate this behavior more than criticize. (1)
d. It looks like quite ordinary and dull, isn’t it? (1)
e. He would rather have a cup of tea than go out to do some exciting things. (1)
f. But our friends can’t understand our feeling totally because they don’t stand near us and can’t make them in our mood. (2)
g. Schools can talk to them, persuade them and encourage them when they make even a little progress. (2)
h. It is a marvelous dream, isn’t it? (1)
i. He told me that he would go back to our country to live after my college ending, planting some flowers and reading some books everyday, sometimes go travelling and go hiking as well. (2, go travelling and go hiking considered coordinating with go back. There was no definite evidence that the student writer intended to write two nonfinite clauses, leaving two errors there)
j. In a word, I become lazier and lazier, this not only showed in life styles but also showed in studies. (2)

2. Finite verb forms wrongly missing in attempts for subordinate clauses were counted, as long as the antecedents or guide words (what, which, whether, that, who, etc.) were written together with other components; those wrongly missing in attempts for main/independent clauses were also counted, as long as other sentences components were written and there were signs of clear clause boundaries. In these cases, verb form errors and erroneous clauses were also counted for not having finite verbs.

a. While the children have their own choice different from what their parents’ choice, generation gap comes into being, resulting in the unharmonious family full of fight between the two gap today. (3)
b. We have to think of what the causes of this and what solutions you can suggest. (3)
c. He was only seventeen when he went out of home to join the army. (2)
d. So, my father is a people strong, stubborn and self-confident. (2)
e. when we old, we would go to a quiet countryside to spend our rest life. (2)
f. My father doesn’t talk much with me when we at home. (2)
g. As my father older, he didn’t try to do something that was very difficult for him, (3)
h. And they still my friends now, although we are not in the same college. (2)
i. Time in Junior High chronologically but I can call back a sense of innocence and pure happiness. (2)
j. but now, we have more choice to learn what we interested in. (2)
k. And I think what I leaned in this year more than the past two years. (3)
l. I think I can got a lot when I in college. (3)
m. I was in college when I just a 15-year-old high school girl. (2)
n. That because I’m not a child anymore. (2)
o. and thus they have no chance to get access to whatever outside the school. (2)
there is no such phenomenon that teachers ask the students to pay attention to what is written on the blackboard as in high school. (4)

3. Set phrases such as what’s more, what’s worse, or what’s important were not coded as finite clauses, but as far as I am concerned was.

4. Imperative sentences were counted as finite clauses.
   a. Do what we can to keep our families close, and I’m sure we all can benefit from it. (4)

5. Generally a take-it-as-it-is approach was adopted when finite verbs and nonfinite verbs were used in confusion. In particular, there be + subject + a finite verb form was coded as two finite clauses, with an error, a clausal error, and a verb form error if the form was not appropriate as a finite verb form for an assumed subordinate clause. Similar constructions (a finite verb + a noun + another finite verb) were treated similarly.
   a. But when the spring festival approach, there are always some of his subordenates come to visit my father. (3)
   b. On my way to school, there was many friends walked with me. (2)
   c. Today, I have many thoughts want to talk to you, about my past, present, and future. (2)
   d. I could always see the happiness inside her eyes when she told me the interesting stories happened in her childhood. (3)
   e. Going back to his hometown and living with his brothers and sisters, may be the best thing can do for him. (2)
   f. Having a grateful heart, there’s nothing can be in your way. (2)
   g. Ofcurse, I have more time to play what I liked never means that I forget to study. (4)

6. In several cases, nonfinite verb forms used where finite verb forms should have been used were coded as constituting finite clauses, and finite verb forms used where nonfinite verb forms should have been used were coded as constituting nonfinite clauses, when there was contextual or observational evidence to believe that they were better treated as intended unsuccessful attempts. When a finite clause was coded in this case, an error, a clausal error, and a verb form error were coded. When a nonfinite clause was coded in this case, an error and a clausal error were coded but no verb form error was coded. Therefore, another reason for such a coding method was the possibility to minimize the number of the variety of errors coded.
   a. I having been dreaming of a wonderful life with a harmony family and successful job. (1 finite clause; the nonfinite verb form was taken as an unsuccessful attempt for making the complex present perfect continuous verb form)
   b. Because many people make me believed that study is everything. (2, believed 1 nonfinite clause)
   c. I can make friends with someone share the same habits and interests with me. (1 finite and 1 nonfinite)
   d. Days were unbearable when I stepped back into my city of memories, ’cause no time left for homework. (3)
   e. When talking about my past, I will always feel a little regret and ashamed, but thought about it carefully. I believed it worthwhile. (3)
   f. I did learn a lot whilst lost a lot. (1)
   g. God won’t give you something unless take your something away. (1)

7. Independent elided clauses, rightly or wrongly used, were counted as finite clauses (e.g., Why not? But why? But so what? Why come here? How about now? No communication, no understanding. No pains, no gains. Just a joke. I a student. Not only me, but also other new classmates. Last but not least, my future life. Maybe it totally due
to my study).

8. Finite verbs were distinguished from nouns based on the linguistic contexts.
   a. As the time goes, the view towards some people and some normal things changed.
      For the reason that the facts of the world influence. (3, influence was taken as a finite
      verb, for that’s sake)

9. Some rare cases where prepositions or adjectives may have been misused as verbs
   were also coded, in addition to misspellings of finite verbs.
   a. Personily speaking, I glad that she can change. (2, 1 verb form error)
   b. In a word, his childhood is so though that it beyonds my imagination. (2, no verb
      form error)
   c. That it the study. (1, it was taken as is misspelt, no verb form error)

10. In a sentence a subject plus an auxiliary verb in coordination was not counted as an
    additional finite clause, neither as an additional T-unit, but that in subordination was
    counted as an additional finite clause.
    a. Not only learning is important for me, but also the campus events are. (1)
    b. I try to serve others and learn to do things the best I can. (2)
    c. I am not as dynamic as I was. (2)
    d. We keep the relationship as our parents did. (2)
    e. Hence I want to lighten his burden as much as I can. (2)

In the following example, however, when does was used as a content verb, with so as its
object, an additional finite clause was coded.
   f. As far as I’m concerned, all mothers think highly of their children and so my mother
      does. (3, and 2 T-units)
Appendix 10 Coding Schemes for Errors

1. Errors were counted separately, even though they concerned a single word. For example, spelling errors were counted together with other types of errors concerning the misspelt words.
   a. I hardly read literatures ** and only write* essays on* writing class. (literature, wrote, in)
   b. But from that time to now, my math grades has* never became* lower. (have, become)
   c. The teachers are all modest and have a good knowledge of their major* field* (field of expertise, one error concerning the choice of the word major and the other concerning the modification).
   d. My father negotiated** with my mother these days that he was* willing to travel around the world when they retired* (diction, tense, is, retire).
   e. You said* “Working hard would delate** my disadvanta
      |  ^                ^
      |  


2. The missing comma before a coordinating conjunction (i.e., and, but, or, nor, for, so, yet) joining two independent clauses and that between a dependent clause placed at the beginning of a sentence and the independent clause was not counted as an error unless confusion may be caused, neither was a comma that existed.
   a. My past friends start* their own life* but we always get together in the holiday. (have started, lives)
   b. I really like the knowledge I study** now and I hope they* can help me find jobs* in the future. (am learning, it, a job)
   c. And I will try my best to help him/her for I treat his/her dream as my own dream.
   d. Now, I * conscious about* that only if we do* hard in* our goals, can we success* in the future. (am, conscious, work/study, towards/for, succeed)

3. Both commas and missing commas after sentence-initial prepositional phrases, adverbs such as at present, surprisingly, thus, therefore, moreover, so, sometimes, now, then, today, evidently, of course, occasionally, soon, on the other hand, instead, and furthermore, and the conjunction but were accepted, unless confusion was caused.
   a. At that time I just realized that what he has* now* should 6be attributed to his endevour* and perseverance but not * luck or talent. (had, then, spelling, ^his)
   b. Moreover she will be still happy toward* her life and work, (preposition)

4. The missing comma following however, For example, What's more, and first, that in 4-digit or longer numbers, and that before a direct quotation was counted considering that confusion would be caused and that the use of a comma in such cases has been repeatedly emphasized in China’s English classrooms.
   a. For example* my English writing is very terrible. (comma missing)
   b. What’s more* *He will never be ambitious. (comma missing, capitalization)
   c. Now I am* in Xian* XXX University, 1000* miles away from my hometown. (spelling, Xi’an, 1,000)

5. Comma errors related to restrictive/non-restrictive relative clauses should be counted.

6. Extraneous commas should also be considered errors. For example, Because followed by a comma should be counted as an error (but not but).
   a. *Because,* he used to drink wine and eat friend* food too much*. (fragment, extraneous comma, fried, order of the noun modifier)
   b. But, she doesn’t* like me. (didn’t)
   c. So I think,* the father’s love is differ* from *mother’s love, (extraneous comma, different, ^article)
d. Maybe,* I can get acquaintance* of* more person*, (extraneous comma, spelling, with, persons/people)
e. Maybe, * I’ll be a millioniari*, but I’ll never lost* mine* away*. (extraneous comma, spelling, lose, myself/my goal, redundant preposition)
f. I hope*, I can be a better person in the future. (extraneous comma)

7. A full stop that was used where a comma should have been used was counted as an error, with great discretion, so was a missing full stop at the end of a sentence.
8. A comma that was used where a full stop should have been used was counted as an error, with great discretion.
a. My father was a polite person, and he seldom got angry,* He always educated me in a more soft* way, so I liked to stay with him at the time. (full stop, soft)
b. That * all of mine,* It’s for you, also * to me. (no finite verb, full stop, for)
c. My father was born in He Nan* province,* He can say* * * * He Nan* language* until* now. (Henan, full stop, speak, *the, Henan, dialect, even)

9. Tense/reference errors were coded based on contexts. The seemingly inconsistent present tense usages in between or at the end of a series of past events were not considered errors if they were used to indicate how one feels about his/her past at present.
a. Every day time* he tried his best to raise* more money, but at night, we whole family would get together to watch TV, which thing* is* his favorite. Ohe*, I forget to say,* he also likes* to play chess since he was very young,* And he can play it very well. When he was free, I mean, relative* free,* He would play it with his old friends. (the next paragraph) Now my family are living better and better. (Every day/In the daytime, earn, which, was, Oh, extraneous comma/other bigger corrections, tense/other bigger corrections, full stop, relatively, comma)
10. Accept both British and American usages (e.g., colourful, any more & anymore).
11. Oral/Informal English, including words, abbreviations, and idioms (’cause, wanna, etc.), was not counted.
a. We always shotted* during class and shotted* louder after class.
b. And I think I will take good care of him, just like he looked after me when I was young.
c. I hung out with a girl but our love ended in failure.
d. Therefore, though unconsciously, I became kind of proud.
e. Days were unbearable when I stepped back into my city of memories, ’cause no time left* for homework. (was left)
12. Article errors were counted.
13. Errors in proper nouns, especially errors in proper nouns or others nouns in the Chinese pinyinized forms, were counted (e.g., Shan Dong should be Shandong).
14. Chinese characters were counted as errors; blanks (left for words) were counted.
a. I didn’t let him down and was accepted by a famous university, which made him more *
15. Errors in capitalization were counted.
a. Playing *chinese chess with old friends and reading books will be good choices to kill time.
16. Only the last of a series of run-on sentences was considered right.
a. At class, the teachers give me lots of knowledge,* however, I can’t understand what they said*. (say)
b. I went to school*, he told me the news.
c. I listning** to the* rock music everyday**, isn’t it a happy college life? (listen, article, every day, run-on)
d. Apart from studying, I hardly go to play basketball, it accounts to that now I have relaxing ways. I could go to a cinema or a KTV and so on. (run-on, misspelling, is due to, ^the fact, ^other, run-on)

But the following was not considered a run-on sentence:
a. We don’t have a teacher, we don’t have experience, but we all like acapella than anything. (more than, ^else)

17. Several cases where it was hard to judge full stops from commas in students’ writings when such punctuations were followed by lower-case words were tolerated.

a. But things changed. since my father left us to support my family.

18. A sentence fragment was coded as an error (the first capitalized letter of the fragment or the body sentence was not counted as an additional error).

a. We don’t have a teacher, we don’t have experience, but we all like acapella than anything. (more than, ^else)

b. Our holiday of past was truly have fun. (more than, ^else)

c. Now I can see my future: reading books, enjoying a common job, having an ordinary wife and children, writing articles and books, * enjoying the thinking world I fabricate for myself. (and, diction)

d. I will make future to repay my parents and have a fortunate family. (verb use, determiner, fortunate was tolerated)

e. In class, I also confront some strange students. (verb use)

f. In my eyes, my father aways showed up as an soldier. (spelling, verb phrase use, article)

22. As may have been shown, some expressions that might not be idiomatic but were grammatical and intelligible were tolerated.

a. There are some persons in the world thinking too much*. (persons tolerated, so profoundly)

b. The changes are also in our bodies.

23. Errors related to collocations were counted once, and language points related to one error that should be correct in reference to the error were exempted.

a. lay more attention on*

b. He will become more wisdom*, more happy, and healthy. (wiser should have been used, but only one error was counted)

c. Another different area* between college and high schools is the course* you study *which is more difficult. (aspect, courses, non-restrictive, is in the subordinate clause
was no longer coded as an error)

24. Some redundancies (the italicized, in the following) were tolerated.
   a. In the past time, he had * opposite opinion. (an)
   b. He always keep* a good attidute* towards the things that are bad in his life. (keeps, spelling)
   c. It makes me feel very sad that *mother’s parents died when she was 20 years old. (*my)

25. Some China English forms (the italicized) were also tolerated.
   a. A fly without a head can present* my near* situation. (represent/describe, recent/present)

26. Though rarely seen, repeated errors were counted repeatedly.
   a. my love for him is becoming deeply** and deeply** day by day. (deeper, deeper, two errors counted for each, one for the misused adverb and the other for not using the comparative degree)

27. Misspellings were counted.
   e.g., And it’s not hard for me to guess what his future life will be life*. (like)