Extensive reading and L2 development:
a study of Hong Kong secondary learners of English

Aileen Irvine

Ph. D.
University of Edinburgh
2006

Abstract

Although extensive reading is regarded by many practitioners as a potentially very useful means of assisting L2 development, experimental enquiry into its effectiveness has so far produced little more than a collection of somewhat disparate findings. Nor has any attempt been made to categorically link any such research findings with second language acquisition theory. Consequently, we have no coherent, research-based theory of L2 extensive reading.

Using data from a large-scale project implemented in Hong Kong secondary schools, the L2 English writing of students participating in an extensive reading scheme as part of the school curriculum was compared to that of non-participant students. Samples of timed narrative writing from 392 students in Secondaries 2 and 3 were rated holistically on a scale of 1 - 6 for overall quality, grammatical complexity, grammatical accuracy, vocabulary range, coherence, spelling and conventions of presentation. A subset of 150 compositions from two control and two experimental classes were further evaluated on a range of objective measures.

Results from the two evaluation procedures were cross-referenced, and indicate that extensive reading in an L2 may benefit language development in quite specific ways. Findings are discussed within the context of current psycholinguistic and neurolinguistic theory and an explanation consistent with such theory is proposed. It is argued that, because it is likely to be subserved by a different memory system from that which subserves formal classroom instruction, extensive reading may enhance levels of automaticity, thus favouring the development of fluency, and, concomitantly, complexity and coherence. At low levels of L2 competence, extensive reading may also accelerate the acquisition of basic grammar through frequency effects.

DECLARATION

In accordance with Regulation 3.8.7 of the programme of postgraduate study, I declare that this thesis has been composed entirely by myself. The work it contains is my own and has not been submitted for any other degree or professional qualification.

Aileen Irvine

Edinburgh,
August 2006

Acknowledgements

This PhD thesis might not have been written but for the following four people: Dr. Tony Lynch and Professor Alan Davies from the University of Edinburgh, Mr. David Hill of the Edinburgh Project on Extensive Reading and Dr. Vivienne Yu from the Hong Kong Institute of Language in Education.

Dr. Yu tirelessly oversaw the Hong Kong Extensive Reading Scheme in English for six years, and was no doubt largely responsible for its success. She also, amongst many other things, managed the collection of data from schools for the large-scale programme evaluation whence my own data originated, and organized my many school visits. I am extremely grateful to her.

Mr. Hill has been my friend and colleague for 15 years. His energy and enthusiasm for extensive reading and his practical and moral support for my own projects, including, latterly, my PhD research, have remained constant throughout.

My second supervisor, Professor Davies, advised on the original research design for the large-scale programme evaluation in Hong Kong. His wisdom and guidance have been invaluable, and his patience and wit have always made meetings with him a pleasure.

My gratitude must be greatest, however, to Dr. Lynch, my principal supervisor, who has had the thankless task of keeping me on the path to PhD completion. His help and support have always been unfailing, and his professionalism no less than inspirational. I consider it a privilege to have been able to work with him.

I would also like to thank the teachers and students from Hong Kong who have unknowingly helped me in my research, and the many friends and colleagues at the University of Edinburgh who have offered unstinting moral support, particularly Dr. Joan Cutting, Dr. Cathy Benson, Mrs. Joan MacLean and Mrs. Yvonne Foley. Mr. Eric Glendinning, Director of the University of Edinburgh's Institute for Applied Language Studies, has given much practical support. Mr. John McEwan and Mr. David Burzala have provided expert — and consistently good humoured — computing assistance.

Finally, I must thank Mrs. Sarah Irvine, who has never for a moment believed that her daughter would not achieve anything she set out to do, and to whom it would not, in any case, have mattered if she hadn't.

TABLE OF CONTENTS

APPENDICES

TABLES

ABBREVIATIONS USED

EPER	Edinburgh Project on Extensive Reading
ERS	Extensive Reading Scheme
GSL	General Service List (West, 1953)
HKERS	Hong Kong Extensive Reading Scheme
HKRA	Hong Kong Reading Association
ILE	Institute of Language in Education
LFP	Lexical Frequency Profile (Laufer and Nation, 1995)
REAP	Reading and English Acquisition Program
SRA	Science Research Associates

1. INTRODUCTION

Extensive reading comes with a long list of practical advantages. A programme of extensive reading can provide a rich and varied input where such might otherwise be missing; it allows for individual tastes and learning speeds; reading materials may be carried about by individuals who may thus engage in reading anywhere and at any time. Additionally, books are relatively cheap, constitute a fairly easily managed classroom resource and are not beyond the personal and professional experience of teachers, wherever they may happen to be teaching.

Rather disquietingly however, extensive reading lacks any sound, research-based theory of language learning. Instead, its many supporters must rely on truisms such as "more is better", "simpler is easier" and "interesting is motivating". These may justify extensive reading *methods*, but do nothing to explain how input derived from extensive reading may engage cognitive mechanisms in any way different from those activated by input from other teaching methods. Even Krashen's highly influential *comprehensible input* theory (1985), cited by many as a justification for extensive reading methods, does not, in fact, provide any research-based insight into this question.

There are many problems attendant on any long-term, classroom-based research project, not the least of which may be, quite simply, managerial. The particular problems faced by researchers investigating experimentally the language learning benefits of extensive reading are discussed in Chapter 2 of this thesis. In addition to field studies, certain component parts of extensive reading have been found by researchers to lend themselves to short-term experimental investigation. This type of research has focused on such narrower questions as the effects on learners' comprehension of different types of text modification and percentages of unknown vocabulary in a text. Whilst potentially informative, research which focuses exclusively on properties of the reading input, without regard for the practical and psychological realities of extensive reading, does not, however, carry the same *ecological validity* (*cf.* N.C. Ellis and Schmidt, 1997) as field studies.

In 1991, a large-scale English extensive reading scheme was set up in secondary schools in Hong Kong. This project was carefully managed, locally and with the help of the University

of Edinburgh, so as to be sustainable and long-term. As a result, the project provided a very rich source of classroom-based, ecologically valid data on extensive reading. The aim of this present study has been to analyse, more comprehensively and rigorously than has sometimes been the case with extensive reading studies, the free written language production of students from participant schools and, hence, to identify in exactly what ways the written English of students who engaged in extensive reading over a period of nearly three years might differ from that of peer students who did not. Specific findings have led to the formulation of what I believe to be the first research-based, SLA theory referenced explanation of how extensive reading in an L2 differs from other kinds of text-based L2 language instruction.

The study begins with a review of the accumulated findings from field studies over the past 25 years, since the publication in 1981 of Elley and Mangubhai's landmark report on the impact of the Fiji Book Flood. Collectively, these studies do not present a particularly coherent picture. Chapter 2 also discusses the findings from those investigations into effects of specific properties of reading input which are of relevance to extensive reading.

Chapter 3 outlines the particular context of the study, describing the Hong Kong Extensive Reading Scheme (HKERS) in some detail. Chapter 4 deals with the methods which were used to obtain the data and analyse it. Two stages of analysis were undertaken. The first stage consisted of evaluation by raters, and draws from the discipline of language testing, addressing, in particular, issues of validity and reliability. Results from this first stage of analysis, for the whole data set of 392 compositions, are given in Chapter 5.

The second stage of data analysis, deriving from traditions of text-analysis, comprised an investigation into objectively identifiable and countable surface text-features. A subset of 150 compositions from a single school was used for this part of the study. Chapter 6 presents the results. This chapter is considerably longer than the other chapters as it also includes the results from the first stage of evaluation for these particular 150 compositions and triangulates the two sets of findings. It is the conclusions drawn from the two evaluation procedures, in the light of this cross-referencing, which form the basis for the discussion in Chapter 7. This last chapter also integrates the findings into current psycholinguistic and neurolinguistic theory. From this we may see that extensive reading embodies not just a practical input device, but also a language acquisition method in its own right, which may engender cognitive processes quite distinct from those activated by other teaching methods.

Insights gained from these findings point in a useful direction for future investigation and may ultimately contribute to a much needed research-based theory of L2 extensive reading. They may also help classroom teachers to make more informed choices as to the types of reading activity which can be useful in language learning, and what kinds of benefit might be expected from extensive reading in particular. This, in turn, could lead to a better informed evaluation of the impact of this kind of reading practice. Finally, results from this study underscore the pedagogical value of independent silent reading, uninterrupted by form-focused, teacher-led intervention, and must be seen as lending considerable support to Krashen's *comprehensible input* hypothesis.

2. PREVIOUS RESEARCH

2.1 Overview

Over the last 10 to 15 years, extensive reading as an EFL pedagogical practice has received considerable support from a number of well-known names in the profession. It is endorsed by teacher educators in recent major methodology handbooks (*e.g.* Hedge, 2000; Richards and Renandya, 2002); it is supported by high-profile educational theorists such as Krashen, who believes it to be a valuable source of *comprehensible input* (1985, 1989), and Nation, who sees it as a vehicle for *incidental* vocabulary learning (2001); there is a respected body of extensive reading programme management literature, spearheaded by Day and Bamford (1998) and Hill (1992, 1997, 2001). In addition to this, there is no shortage of enthusiastic testimonies and recommendations from classroom teachers, such as Shlayer who writes: "For some pupils, the extensive reading program comes as an awakening" (1996: 32) and Shelton, who claims:

> There seems to be no shortage of experts and literature extolling the inherent power and purpose of extensive reading: there seems to be no lack of reasons why we, as teachers, should not [sic] be encouraging our students to do this. (*n.d.*: *unpaginated*)

Notwithstanding this relatively high level of support at all levels of the organizational structure of the EFL profession, there is, in fact, very little reliable, documented experimental evidence for the L2 language learning benefits of extensive reading. Intuitively, the notion is very appealing. The counterclaim — that reading extensively in a target L2 will make no difference to the learner — seems irrational. Many practitioners *believe in* extensive reading. There may even be said to exist an extensive reading lobby. Why then do we as yet have no strong body of coherent experimental evidence?

Day and Bamford (1998) have been criticised for claiming as an "approach" what is, in fact, no more than simply "reading extensively" (Bruton, 2002). Although no time-scale is actually mentioned by Day and Bamford in their taxonomy of what they consider the primary characteristics of extensive reading, it must be assumed that *extensive reading*, like "reading extensively", applies to longer rather than shorter time periods. (Day and Bamford's taxonomy is reproduced in Appendix 1.) This requirement may, in turn, raise practical difficulties, as the longer a programme runs, the greater the chance of outside factors

impinging on the project. This has been found to be particularly the case in Africa, for instance (Davidson and Williams, 2005). Outside factors may also impinge on any research design which relies on long-term implementation. What may be the most widely-cited study to investigate the language learning effects of extensive reading, that of Elley and Mangubhai (1981, 1983), who reported on the Fiji Book Flood project, evaluated a large-scale reading scheme over a period of nearly two years. In perhaps the next most commonly cited study, Hafiz and Tudor (1990) reported on a project in Pakistan which lasted 23 weeks. Both studies reported research-project management problems. In the case of the Hafiz and Tudor study, which used no more than three groups of students, data from one of the groups, an intended control group, had to be discarded because of "a lack of commitment to the project on behalf of the learners involved" (Hafiz and Tudor, 1990: 33). Elley and Mangubhai reported that, in a study involving eight rural Fijian schools, although overall results showed a number of significant effects in favour of two experimental groups, results school by school varied quite markedly. During the first phase of the study, one experimental teacher was known to "have used the method rarely, as he felt it was unsuitable for him and his pupils" (Elley and Mangubhai, 1983: 59); conversely, one control teacher liked the reading scheme methodology so much that she unofficially borrowed it for her own English lessons, using her personal collection of reading materials. In the second year of the project, whilst eleven of the 16 Book Flood classes maintained the advantages they had shown in the first round of post-tests, implemented after eight months of the project, five others showed less than average progress.

Clearly, it is very difficult to control conditions in the normal classroom for the length of time required to be able to lay any creditable claim to having investigated experimentally the effects of *extensive* reading, and this may be one reason why so few long-term studies have been reported. An alternative to depending on required high levels of cooperation from others is, of course, for the experimenters to teach control and experimental groups themselves (*e.g.* Yang, 2001; Bell, 2001) or to teach the experimental group and use a previously evaluated comparable group as a pseudo control group (*e.g.* Lai, 1993). Such a context may often, however, carry the risk of an *experimenter effect*. In the Hafiz and Tudor study mentioned above, Hafiz taught the experimental group. Prior to the experiment, he had not been a member of the school's staff, and Hafiz and Tudor themselves admit that "the novel effect of the programme itself, combined with the personal contribution of the experimenter, may have constituted significant causal factors in the overall success of the

project" (1990: 41). Even in cases where there may have been no experimenter effect, it is difficult to disprove the possibility, and claims for a programme effect under these kinds of circumstances are open to counterclaims of a possible "programme plus enthusiastic experimenter" effect.

Another type of study has used self-reporting questionnaires to establish amounts of reading, either prior to a language proficiency evaluation (*e.g.* Janopoulos, 1986; Lee, Krashen and Gribbons, 1995; Constantino, Lee, Cho and Krashen, 1997; Renandya, Rajan and Jacobs, 1999), or during the period of time between a pre-test and a post-test (*e.g.* Renandya *et al.*, 1999; Hayashi, 1999). Variables are operationalised (*e.g.* test scores, gain scores, number of pages read, number of books read) and correlations between these are reported. While these studies may offer much interesting material for speculation, a major weakness is that it is not possible to attribute to amount of reading any causal effect as regards higher or lower test scores. In the case of the amount of reading done *before* a language test, it is just as likely that a proficient language learner will have read more in the target L2 than one who is less proficient, simply because he *can* do so, as it is that reading in the L2 has caused this proficiency. In other words, the proficiency may have preceded the reading rather than the other way around. In the case of the amount of reading done over a specific period during which time a proficiency gain was observed, this may have been the result of any one of several factors. An increasing proficiency, perhaps caused by a concurrent language programme, may have facilitated an increase in amount of text read. (Renandya *et al.* 's study ran alongside a concurrent teaching programme; Hayashi does not make clear what actually went on in the language classroom.) A sudden increase in motivation may have boosted enthusiasm for both reading and making better use of other learning facilities, including classroom instruction, leading to gains in both amount of reading and measured language proficiency, brought about by this, or indeed some other, intervening variable.

In addition to these difficulties, it is very problematic to disengage what may be effects of extensive reading in particular from what may be effects of additional exposure to the target L2 — and herein lies a large part of the difficulty of differentiating between extensive reading as a methodology, and simply reading extensively. In Hafiz and Tudor's 1990 study, for example, experimental students from a rural Pakistani school had twice as much classroom exposure to English as control students, with four extra hours of reading-class time weekly, resulting in 90 more hours than control students of in-class L2 contact over the period of the study. In the companion study of a year earlier, in Leeds (Tudor and Hafiz,

1989), experimental students benefited from approximately 60 additional hours of reading-class time over a period of 12 weeks. This is not to deny any language learning benefits of extensive reading which are reported by studies where experimental students have had additional periods of L2 exposure, but to say that these must be taken, not as resulting from extensive reading as a particular method, but as the results of a whole approach — that is to say a package that we could call "the benefits of additional exposure through the use of extensive reading". There is still much value in such findings, as these kinds of extra reading classes may provide an easily organised, low-stress, "language-booster" option in situations where, although students need more than their current instructional situation appears to be giving them, it is not possible to provide extra classes. Supervised extensive reading groups require little or no preparation from the teacher, and have the added advantage of being able to accommodate weak and strong students alike, so that the weakest are not in a despair of discouragement whilst the strongest are bored and resentful.

One other factor which may have contributed to the shortage of sound studies is the lack of face-validity of "just reading" as an instructional practice. It may be difficult to justify what can be seen as hardly fulfilling the proper role of a teacher in the eyes of senior colleagues or fee-paying students. Robb and Susser (1989) found themselves unable to implement a full-blown extensive reading methodology in their teaching institution, as they would have wished, as they felt it would "not be acceptable to the students or our colleagues for the experimenters to let students merely read in class" (1989: 243). Consequently, they had to adopt a compromise method involving only some of the aspects of extensive reading. Lai (1993) set up a summer reading course with the express intention of being able "to isolate and investigate the effect of extensive reading" (1993: 89), hence to provide good evidence of its effectiveness, following on from a disappointing result in a previous project which she felt had not been conducted properly, with too many uncontrolled intervening variables. In this second project, only half of the class time was used for extensive reading and the other half for whole-class activities quite unconnected to extensive reading, such as songs and games. It seems very probable that she felt she could not ask students who had paid for a 4-week summer course to simply spend the time reading silently. Indeed, teachers themselves often feel uncomfortable at taking on such a passive role in the classroom. In the major secondary-school extensive reading scheme implemented in Hong Kong classrooms in the 1990s (from which the data for this present study came), whilst visiting schools in order to observe reading classes and interview teachers, I found this to be a common complaint, and

that the removal of the teacher from centre stage in the classroom even made the scheme unpopular in some schools. Rather tellingly, the experimental methodology which was borrowed by a control teacher in the Fiji study was the shared-book method, *not* the individual silent-reading method (Elley and Mangubhai, 1983).

In the rest of this chapter, 12 studies will be discussed in terms of their specific findings and their specific weaknesses. These represent the main body of published experimental research into the effects of L2 extensive reading programmes to date. Some are well-known, such as the work of Elley and Mangubhai (1981, 1983) and Hafiz and Tudor (1989, 1990; Tudor and Hafiz, 1989), and are cited in most articles concerned with the potential benefits of extensive reading. Others are as yet less well-known, such as Yu (2000) and Bell (2001), but deserve better recognition. To afford a more coherent picture, these studies are grouped under two main headings:

- evaluation of long-term extensive reading schemes implemented at national level

- shorter-term individual extensive reading projects which have been set up expressly as objects of investigation, or studies where existing teaching situations have been deliberately managed so as to supply useful data on extensive reading.

This latter group is further organized into three categories:

- studies which do not attempt to separate extensive reading from additional input, but which investigate the benefits of reading which is *supplementary* to the normal syllabus

- studies which attempt to compare methodologies, investigating the use of extensive reading practices *instead of* a competing methodology; experimenters in these cases have made efforts to ensure that control groups received as much exposure to the L2 as experimental groups (although this may often be difficult, given the nature of *extensive* reading)

- studies which fall somewhere between these first two categories, and where in-class extensive reading has encroached on at least some of the time which would normally be used for other teaching practices, but where additional reading is also routinely undertaken outside class time, as part of the teaching method.

Studies correlating reading-habit questionnaire answers and proficiency test scores will not be discussed because of the impossibility of attributing cause or effect. Good readers may become good students, good students may become good readers, or the two characteristics may evolve concurrently. However, a second type of research which is potentially informative is that which takes the form of tightly controlled, relatively brief experiments into several of the deconstructed parts of extensive reading. Whether it is true that extensive reading in an L2 — as a *style* of reading, rather than as a whole methodology — can lead to the *incidental* acquisition of new vocabulary as a by-product of engaging with text for *meaning,* with no deliberate focus on form, for example, is a question which has raised much interest amongst researchers (*e.g.* Pitts, White and Krashen, 1989; Day, Omura and Hiramatsu, 1991; Dupuy and Krashen, 1993; Horst, Cobb and Meara, 1998). Some researchers have taken this further, trying to establish exactly how many times a new word needs to be met for learning to occur (*e.g.* Rott, 1999; Waring and Takaki, 2003). Researchers have also investigated what percentage of words in a text needs to be known before satisfactory text comprehension will occur (*e.g.* Laufer, 1992; Hu and Nation, 2000) or, more explicitly, before previously unknown words can be understood and acquired (*e.g.* Liu and Nation, 1985). Another strand of research has focused, with varying degrees of success, on the possible effects on the reading process, and hence on the learning outcome, of using graded or simplified texts (*e.g.* Blau, 1982; Yano, Long and Ross, 1994; Tweissi, 1998). These studies will be discussed in the second part of this chapter.

2.2 Large-scale extensive reading programmes

The implementation of large-scale L2 reading programmes requires not only the long-term cooperation of teachers and administrators, but the power to change a syllabus. Hence, these must have support in very high places. The widely-cited report of Elley and Mangubhai (1983) on the Fiji Book Flood project of the early 80s was not, in fact, the first of its kind. In 1979, a book-based English teaching programme was launched in primary schools on the small South Pacific island of Niue (Elley, 1991). It was the success of this programme which prompted the implementation of the Fiji Book Flood. Other book-based projects followed on from the success of the Fiji project, most notably the Singapore Reading and English Acquisition Program (REAP), between 1985 and 1989. These projects were prompted by a pressing need to raise the English literacy levels in education systems which

used English as the medium of instruction in primary schools — from Primary 1 onwards in the case of Singapore, and from the end of Primary 3 onwards in Nuie and Fiji — and which had until then been using a very highly controlled and input-impoverished method of L2 English instruction.

All of these programmes reported significant improvements on a variety of English language measures for the students following the book-based methodology. However, only the study reported by Elley and Mangubhai (1983), on the Fiji project, attempted to separate out the effects of silent, individually controlled extensive reading. Whilst the REAP programme did provide books for independent silent reading, at an unspecified later stage in the programme, this was not investigated as a separate feature. In fact, these book-based programmes could not really be said to be reading programmes so much as integrated or "whole-language" approaches, using books as input. At pre-service training sessions, teachers were shown how to optimally exploit the texts for listening, speaking, writing and language-focus activities. *Shared reading* constituted the larger portion of class time, with students, working from the same story book, reading aloud, listening to the teacher reading aloud, acting, re-writing, discussing the story, discussing vocabulary and playing follow-up vocabulary games.

The Fiji Project was alone in establishing two experimental groups. One group used only the shared-book method outlined above, and one group used only, to use Krashen's (1993) terminology, *sustained silent reading*, with books of the students' own choice, which they read for 20-30 minutes every day. These Primary 4 and 5 children had been educated in the vernacular for the first three years of their schooling but had begun, in Primary 4, to use English as a medium of instruction in all school subjects. It is not clear from the various reports (Elley and Mangubhai, 1981, 1983; Elley, 1991; Elley, Cutting, Mangubhai and Hugo, 1996) what other instruction in English, if any, the children concurrently received. The 1981 report notes only that "All teachers retained their usual English timetable" (Elley and Mangubhai, 1981: 8). Books used were L1 children's story books, often highly illustrated.

The control group continued to use the method which all students, and, in fact, all schools in the South Pacific territories, had used until then: the Tate audio-lingual programme. This method taught selected structures and words orally, and in strictly prescribed sequence, which were drilled until mastered. Only then might the words and structures be seen in

print. Two 15-minute periods of drilling took place every day. Reading tuition was conducted separately using the accompanying series of graded readers, the linguistic content of which strictly matched the previously-drilled aural input, and students were denied access to any other text than that which had already been drilled orally.

Results from the first round of post-tests, eight months into the programme, showed a significant difference in scores on a 32-item, multiple-choice reading comprehension test, in favour of the two reading groups combined, in both Forms 4 and 5. There was no significant difference *between* the two reading groups in Form 4, but a significant difference in favour of the shared-book group in Form 5. A multiple-choice test of structures revealed a significant difference in favour of the reading groups in Form 4, but not in Form 5. There was no difference between the reading groups and the control group on an *oral sentences* test given to Form 4, in which students had to repeat increasingly complex sentences after these were read aloud to them. However, when compared to the silent-reading group, the shared-book group on its own performed significantly better on this. In Form 5, the reading groups significantly outperformed the control group on a listening test, and the shared-book group surpassed the silent-reading group. A short written composition, based on a sequence of pictures depicting a story and marked holistically on a scale of 0 to 2 points for each of three categories of *content*, *sentence sense* and *mechanics*, showed no significant differences.

Follow-up tests one year later showed significant differences in favour of the book-based groups in *reading*, *word knowledge*, *English structures* and *written composition*, but no significant differences between the shared-book and silent-reading groups.

Whilst these results are striking, they are not very surprising. The control group functioned within the most controlled L2 teaching environment imaginable. The story books effectively lifted the lid off this L2 teaching capsule and allowed hitherto undreamt-of amounts of input to come streaming into the classrooms. Perhaps just as importantly, they introduced the notions of choice, individual responsibility and enjoyment into the L2 learning process. Further, as Elley and Mangubhai note, "Theoretically.... new learning takes place at the point of interest, rather than in accordance with a carefully graded linguistic pattern" (1983: 58).

What *is* perhaps less expected, and worthy of comment, is the evolving pattern of differences between the shared-book and silent-reading groups. Initially, the shared-book group

performed better than the silent-reading group on a number of measures, namely *reading comprehension* and *listening* in Form 5, and *oral sentences* in Form 4. Over the longer term, such differences disappeared (although the *oral sentences* test itself was not repeated in the second round of tests). In the case of reading comprehension, this, and the fact that the Form 4 silent-reading group had, in any case, already obtained significantly better reading comprehension scores than the control group in the first round of post-tests, showed that silent reading did not merely result in improved reading comprehension, independently of teacher intervention, but, over the longer term, in as much improvement as resulted from the teacher-led, intervention-heavy shared-book approach. If language improvement derives from *intake*, then students, reading independently, were able to intake all by themselves, without needing to have possible candidates for intake pointed out to them by their educators.

It is, however, rather surprising that the difference in listening skills between the two groups did not persist, given the much greater opportunity afforded to the shared-book students, in their English classes at least, to listen to and interact verbally with their teachers and classmates. Elley and Mangubhai do not comment on this point. Although there may be some leakage between inputs obtained via different modalities (Harley, 1995), it is also quite possible that using English as a medium of instruction had evened out the initial difference. One other point of note is that over the long term both reading groups evolved into better composition writers than the control group, suggesting that writing skills may take longer to benefit from reading input than reading skills.

One other large-scale reading programme evaluation which has isolated and experimentally investigated individual silent reading as part of a national L2 syllabus is the Hong Kong Extensive Reading Scheme (ERS) evaluation carried out by the government-run Hong Kong Institute of Language in Education, with the help of the University of Edinburgh, between 1993 and 1994. The impetus for implementation of the reading scheme had been similar to that for the Fiji Book Flood, which is to say that poor standards of English in schools had led to concerns at the highest levels for a Ministry of Education which permitted English as a medium of instruction in any secondary school which wished to use it as such. (Normal *primary* schools used Chinese as the medium of instruction.) The materials and mechanics of this reading scheme are described in more detail in Chapter 3, since the data used in this present study originated from the government-funded evaluation of the scheme.

There are a number of major differences between the Hong Kong and the Fiji studies. The Hong Kong ERS operated in Secondaries 1, 2 and 3. Students taking part in the evaluation were aged 13 to 14, and had received approximately five hours a week of L2 English instruction throughout six years of primary school, prior to entering the ERS. Consequently, in theory at least, they were already able to read much more fluently than the Fiji primary schoolchildren, which may have had some bearing on amount and type of intake. Whilst the book-based approaches of the Fiji project provided an alternative to a very tightly controlled, low-input audio-lingual method, Hong Kong secondary schools offered a more balanced programme of textbook-based language study, dictation, composition writing and listening practice. This methodology was formally underwritten by communicative principles and aims, although in practice much classroom activity was teacher-orientated and examination-driven. Nonetheless, the Hong Kong reading programme was compared against a methodology which was very different to the Tate syllabus. Thirdly, the socio-economic environment was also very different. The rural primary schools used in the Fiji study had been deliberately chosen in preference to schools from a more sophisticated socio-economic environment, in order to optimize the impact of a sudden availability of reading materials, and to reduce as far as possible any experimental noise caused by outside factors.

The Hong Kong ERS was evaluated in Secondaries 1, 2 and 3, using a reading test for Secondary 1, and an extensive reading test, a multiple-choice vocabulary test, and a timed composition for Secondaries 2 and 3. (The research design and evaluation tools are described more fully in Chapter 4.) The reading test results were conclusive, showing significant differences in favour of the experimental group at all three form levels (Yu, 2000). Control and experimental groups did not, however, perform differently on the multiple-choice vocabulary test in either Secondary 2 or Secondary 3. Two reasons are suggested for this: the vocabulary test had been written specifically to accompany the reading scheme at the request of the Hong Kong Ministry of Education, who had stipulated that a single one-hour test must accommodate the complete range of reading abilities encompassed by the reading scheme materials. In other words, one test must range in item difficulty from near-beginner to high upper-intermediate. This is rather a lot to ask of one short test, which thus had to make narrow discriminations amongst a very wide range of levels, and the test, in fact, did not do this so well at the middle levels (Davies and Irvine, 1992c). Nonetheless, the test did discriminate *between* the two form levels, and also between

higher and lower ability schools, and this would imply either that any effects on vocabulary knowledge brought about by the reading scheme were not susceptible to the type of measurement offered by a multiple-choice test, or that the reading scheme did not have any significant effects on vocabulary knowledge.

Results obtained for the timed compositions were not very illuminating, showing an erratic mixture of effects, with no clear pattern or obvious explanation. Two teacher-raters, using a Scoring Guide with sets of descriptors, but working independently, each provided a score, ranging from 1 to 5, for the three constructs of *content*, *narrative structure* and *language and style*. Results, reported in Yu (2000), showed no overall difference between control and experimental groups, with the exception of one of the raters having awarded significantly higher scores for *language and style* to the experimental students. Inter-rater reliability was, in fact, not reported.

When the data was subsequently investigated by level, and split into four ability groups, no significant differences were found between control and experimental for the highest group. However, the lowest group showed a significant difference in *language and style*, in favour of the experimental students, the second lowest showed a significant difference in *narrative structure,* again in favour of the experimental students, and the second highest group showed a significant difference in *content*, this time in favour of the control students. A superficial reading of these results might be that the ERS improved *language and style* at a low level of proficiency, *narrative structure* at a slightly higher level, led to a highly significant deterioration of *content* ($p < .005$) at a yet higher level, and ultimately to no long-term advantages for participant students at the highest proficiency level (although *content* then recovered, despite the ERS).

It is, of course, possible, indeed likely, that different skills benefit differently from a particular methodology or type of input at different proficiency levels. Nonetheless, these results were puzzling, particularly in the light of teachers' responses to the questionnaire also administered as part of the ERS evaluation. Of teachers who had used the ERS at these levels, 50% (N = 64) thought the scheme had helped the highest level students "very much" and 46.9% (N = 60) thought it had helped them "moderately". However, only 5.1% (N = 7) thought it had helped the weakest classes "very much", whereas 60.6% (N = 83) thought it had helped only "a little" or "not at all". These responses, across a wide range of schools,

"showed clearly that the teachers perceived the scheme to be most effective for the brighter students in the best class and least effective for the weaker students" (Yu, 2000: 196) — the complete antithesis of the writing test findings.

The huge number of students involved (over 3,300) and the dissociation of the researchers from the day-to-day teaching of the ERS constitute strong evidence for the reading scheme's enhancement of reading skills, after three years, two years, and even only one year. Since only some of the reading was undertaken as a replacement for normal classroom instruction, and students were allowed, and even encouraged, to continue their reading outside class, this may have resulted from a combination of the different methodology *and* an increase in exposure to the L2. It is impossible, given the programme design, to separate the two. Some students did little or no reading outside the allotted two ERS classes; other students became, quite literally, prize readers. That the vocabulary test could reveal no effect either overall or at different form levels suggests that an increase in L2 vocabulary knowledge was *not* one of the definite benefits of the scheme. The question of whether the ERS benefited writing skills was not satisfactorily settled by the evaluation procedures used. It is this data which is re-evaluated in the present study.

2.3. Extensive reading experiments

2.3.1 Extensive reading as supplementary input

The two studies by Hafiz and Tudor (1989, 1990; Tudor and Hafiz, 1989) are frequently cited as strong evidence in favour of extensive reading. In the Leeds study (Hafiz and Tudor, 1989; Tudor and Hafiz, 1989), an experimental group made significant pre- to post-test gains on a set of three reading tests and four writing tests, over a period of 12 weeks. One control group made significant gains over the same period in only two of these tests, and a second control group made no significant gains. In the Pakistan study (Hafiz and Tudor, 1990), pre- and post- compositions were analyzed. Over a period of 23 weeks, the experimental group had made significant gains in *writing readiness* (number of words produced), *vocabulary base*, percentages of syntactically acceptable and semantically acceptable T-units and *spelling correctness*. The control group made no significant gains. As has already been noted, both these projects afforded relatively considerable amounts of additional classroom-

based L2 exposure, as they involved extra-curricular reading classes set up by the researchers outwith the normal school syllabus of the participant students. This, however, must simply be taken as a condition of the extensive reading packages which were provided. This apart, the evidence from these two studies, on closer examination, is not particularly strong.

A major weakness in the first study, in Leeds, is the complete lack of information concerning the two control groups. We do not even know if these students had the same L1 as the experimental students. One group is from the same school as the experimental group, and one group is from another school, but we are not told which is Control Group 1 and which is Control Group 2. Results are presented only as pre- and post- means, with a t-value for each paired set of scores. There is no information concerning the spreads of scores, as standard deviations and ranges are not given. However, scrutiny of the figures which *are* provided points rather strongly to three conclusions. Firstly, both control groups were higher ability than the experimental group at the outset. This need not necessarily threaten an experiment, but it is generally true that it is easier to make appreciable, measurable gains at lower levels. In Tsang's (1996) experiment, for example, in which Hong Kong students from four Form levels followed their normal school English syllabus plus an additional after-school enrichment programme over a period of 24 weeks, the mean gain score for Form 1 was 15.79, whilst it was only 4.43, 4.13, and 2.20 for Forms 2, 3 and 4 respectively, who had nevertheless followed exactly the same programme. Secondly, one likely reason that Control Group 1 did not make any significant gains is that this group seems to have encompassed a wider spread of abilities, and the bigger overlap between pre- and post- test scores would have made it technically much harder to obtain a significant result. (This may be deduced from the fact that certain differences between pre- and post-test means for this group did not produce significant t-values, although comparable, or even smaller, differences between pre- and post-test means for Control Group 2, which had the same number of students, did.) Thirdly, and perhaps most seriously, Control Group 2, which was the highest ability of the three groups, may have suffered from a ceiling effect across the whole range of tests. For example, on a 28-item sentence-to-picture matching test, Control Group 2 achieved a mean score of 26.4 on the pre-test and of 26.8 on the post-test. The experimental group, on the other hand, made a significant gain on this test. On a 15-item test of synonyms, Control Group 2 achieved mean scores of 12.6 and 13 on pre- and post-tests respectively. This pattern was observed for each test, with Control Group 2 already achieving very high

mean scores in pre-tests. Even in the only open-ended task, a short composition, it seems quite possible that Control Group 2 was constrained in this way, going from a mean pre-test score of 26.3 to a mean post-test score of 28.9, whilst the experimental group went from 19.7 to 29.8. It is very unfortunate that Hafiz and Tudor do not report the range of possible scores for this test.

These pre- and post-compositions were analyzed for *writing readiness*, *vocabulary range*, *syntactic maturity* and *accuracy of expression* — the same variables as were later used in the Pakistan study. However, only compositions produced by the experimental group were analyzed. Whilst this may show in what ways the writing of this group changed over the experimental period, we cannot make the assumption that these changes were brought about by the reading programme rather than normal classroom instruction. To be able to make any claims regarding this it would have been necessary to compare changes in the writing of the extensive reading group against changes in the writing of the non reading groups, but these were not investigated.

The Pakistan study demonstrates an even more transparent mismatch between control and experimental groups. Whilst experimental students attended school in a rural area, and were the children of agricultural and factory workers, control students attended a school in the centre of Islamabad, and were the children of white-collar professionals "with a generally higher level of education" (Hafiz and Tudor, 1990: 31). Pre-tests (compositions) point to a substantial difference between the groups at the outset. However, the most unsatisfactory feature of this study is the fact that, from pre- to post-test, a period of 23 weeks, during which time normal classroom L2 English instruction took place for — we must suppose, since we are not told — four hours a week (we are told that this is the normal amount of English instruction for the *experimental* group), the control group performed significantly worse than they had done at the outset. If a group of students perform, not just worse, but significantly worse, on a test than they did six months before, and there has not been a six-month period of national holiday, then something must be seriously wrong with, or at least extremely unusual about, the students, the instruction methods or the test and/or test-administering procedures. One really cannot use such a group as a control group without sound theoretical reasons. Since there was no test other than the composition, and all the test variables, which is to say text features, were derived from that one source, which, in the case of the control students must *all* be affected by the significant deterioration in performance,

(pre-read H+T, 1990)

we cannot validly compare these as between-groups variables within the theoretical framework of the experiment. It is true that the experimental students made considerable progress over the 23-week period, notably in number of words produced whilst writing a 30-minute timed composition, which rose from a mean of 111 to a mean of 264, and in percentages of syntactically and semantically acceptable T-units, which rose from means of 26.6% to 53% and 44.9% to 66.5% respectively. However, what may be the effects of the reading programme and what may be the effects of normal classroom instruction cannot be validly hypothesised.

Lai (1993) also used for her experiment purpose-created reading classes which were independent of any institutional syllabus. A series of special 4-week summer courses was set up, accepting Hong Kong secondary school students from Forms 1, 2 and 3, on a voluntary and fee-paying basis. One course ran each summer over three consecutive years. Classes were held in the mornings only, for 2½ hours each morning, giving a total of 50 hours, of which half were spent reading graded readers silently, whilst the other half were spent on whole-class or group activities such as reading poetry aloud and working on riddles or puzzles.

Lai reports results for each summer course as a whole, such that students from Forms 1, 2 and 3 are included in a single group. Overall, students on the first and second courses showed gains on a reading comprehension pre- and post-test which were statistically significant and similar to the gains made on the same test by a previously-tested group of comparable students within the Hong Kong secondary school system over a period of 24 weeks. The third experimental group, however, made little progress, which Lai explains as the result of their being "less motivated towards reading according to their teachers" (1993: 94). The first two groups also made significant gains in reading speed, which, again, the third group did not.

Students from the first two courses were grouped into four bands according to how many books they had read, and an ANOVA was performed using comprehension scores as the dependent variable and reading-bands as four independent categorical variables. (Lai does not make clear whether "comprehension score" here refers to the gain score from pre- to post-test, or simply to the second comprehension score.) For the first group (N = 126) there was no significant main effect for number of books read and comprehension score, whilst for

the second group (N = 88) there was. However, as with correlational effects, this does not mean that one condition is necessarily predictive of the other, simply that, for whatever reason, the two conditions co-occurred.

The third evaluation instrument was a short descriptive composition, written only by students on the third course. Significant gains were made in total number of words, number of error-free T-units and a holistic evaluation of *style* (for which no details of the procedure or of measurement reliability are reported).

Probably the main conclusion to be drawn from Lai's study is that students who are motivated to read *can* achieve language improvement just from silent reading. This conclusion is possible, where it is not possible from Hafiz and Tudor's studies, because there was no other concurrent instruction. Although songs and riddles may have provided *some* additional input, it is unlikely that this type of input would improve reading comprehension, reading speed or facility in writing. It is, however, probably not valid to compare gains in reading comprehension achieved after four weeks of concerted extensive reading practice with gains in reading comprehension after 24 weeks of schoolroom English instruction. Given the elite nature of the course, which had classes of no more than 20 students (compared to a normal Hong Kong class of 40), following a special, voluntary programme for which they had had to pay, and with fewer distractions and competing demands, such as history and geography lessons, and less forgetting-time in between English sessions, it is only to be expected that English language gains would be accelerated.

2.3.2. Extensive reading as a methodology

As with the Hafiz and Tudor studies, Robb and Susser's (1989) experiment is also frequently cited in the cause of extensive reading. This study was more strictly a comparison of teaching methods, with only limited extra L2 exposure for the experimental group. Control students, in a Japanese University, followed a then popular "skills-building" EFL reading course, using a coursebook containing short reading passages and "exercises designed to teach the skills of efficient reading" (1989: 243). Experimental students used class time to read modules from the SRA Reading Laboratory (Science Research Associates, 1959), a self-access reading system which allows students to choose individually what they wish to read, before answering accompanying questions and checking their own answers against

separate answer cards. Both groups were given reading as homework, although questionnaires revealed that the experimental students, who were required to read teenage L1 fiction as homework, actually spent more time reading at home than the other group.

In fact, what Robb and Susser call the "extensive" group only followed *some* of the principles behind extensive reading, as the SRA Reading Laboratory may itself be called a skills-based approach. There is considerable intervention into the autonomous reading process, in the form of the workcards. We are not told how teacher-led the "skills-building" class was, but we must assume a normal amount of teacher activity, even though we are told that, in class, the students individually read passages and did the accompanying exercises. Reading between the lines, as regards the reading done in class, it does not seem so much that two wholly distinct types of reading were compared, but rather that reading comprehension classes with immediate, live teacher feedback were compared against autonomous reading comprehension classes with choice of reading matter and no live teacher feedback. As Robb and Susser themselves put it, the experimental group "were not taught any skills overtly" (1989: 243).

This may still be an interesting comparison to make. The report does not inspire confidence, however, as there are a number of serious anomalies in reported figures. Scores were "analysed using Analysis of Co-variance (ANCOVA)" (*ibid*: 244). ANCOVA, in this kind of context, requires post- *and* pre-scores. It cannot be performed without such a set of paired scores. Two variables, "j. Comprehension" and "k. Vocabulary skills", did not *have* pre-scores, therefore ANCOVA could not have been performed on these variables, and the reported F values for these, significant at $p < .05$, must have derived from some other procedure. Both these variables appear suddenly, with no information as to whence they came, and each is included as a contributor to one of the three significant differences found between the two groups. To take the example of *guessing vocabulary from context*, which is one of the three constructs where a significant difference was found: this construct is a compound of variables "d. Vocabulary skills" and "k. Vocabulary skills", presumably added together in some way. Variable d., which *did* have pre- and post-test scores, did not produce a significant F value. In fact, it produced a somewhat *in*significant F value. Only when it was combined with variable k. was significance found. However, variable k., as we have just seen, was not one of the original variables.

Even more strangely, improvement in reading speed, the third of the three significant differences between groups, is reported as being significant at $p < .001$, in favour of the experimental group. The mean pre-test score for this group was 238.5, which is to say that it took, on average, 238.5 seconds for a student to read the first text. The post-test score was 336.3. Thus it took the average experimental student 97.8 seconds longer to read the second text. (We assume the second text to have been longer, or more difficult for the students.) The mean pre-test control score was 366.1, and the mean post-test score was 411.9. The average control student took only 45.8 seconds longer to read the second text. Any improvement in reading speed is clearly in favour of the control group. Although increasing scores *normally* represent improvement, in the case of total reading time it is *de*creasing scores which represent improvement, and it seems that this change in direction has not been taken account of. There is some possibility that ANCOVA was *not* performed, but perhaps a two-group ANOVA, which would give the same results as a t-test, and that the reported significance derives only from comparison of the two sets of post-scores, with no reference to the pre-scores. In this case, the direction of the significance would be correct, and we might say that the experimental students significantly outperformed the control students, since they took less time to read the passage, but, given the great disparity between the groups at the outset, this does not show very much. Moreover, working against this interpretation of ANOVA having been used instead of the reported ANCOVA, it seems highly implausible that a post-test control-experimental difference of 75.51 seconds should produce a significant result, when a pre-test control-experimental difference of 127.62 seconds, with the same two groups and therefore most likely roughly comparable within-group variance, is reported on the previous page as *not* having done so. Thus we arrive back at the ANCOVA, and a significant result in favour of the *control* group.

Lastly, as regards the variable "b. Getting the facts", it seems extremely unlikely that a pre-test score of 14.45 coupled with a post-test score of 18.84 (*i.e.* an improvement of 4.39) for the control group and a pre-test score of 15.14 coupled with a post-test score of 19.45 (*i.e.* an improvement of 4.31) for the experimental group, could result in an F value showing a significant difference in improvement between the two groups, at $p < .02$, in favour of the experimental group. This variable is also used as a contributor to one of the constructs which is claimed to show a significant difference between groups, combined — just as variables "d. Vocabulary skills" and "k. Vocabulary skills" were combined to produce the only other construct to show a significant difference apart from reading speed — with the second of the

two newly-introduced variables, "j. Comprehension". In short, something appears to have gone seriously wrong, either in the statistics, or in their reporting, and two of the significant findings are highly questionable for more than one reason, whilst the third is actually significant in the wrong direction, *against* the hypothesis.

Tsang's (1996) study compares the effects of three after-school programmes on the composition writing of Hong Kong secondary school students. A total of 144 students from Forms 1, 2, 3 and 4, within the same secondary school, were assigned randomly to a control group, a reading group or a writing group. Students attended their normal English lessons during school time. Since students from a number of Form classes were assigned to each group, there was no intact-class effect, minimizing the possible impact of variables such as different teachers, class dynamics or class sizes. After-school programmes were carried out with a minimum of teacher intervention, with all activities undertaken in the students' own time, outside the classroom.

The aim of Tsang's experiment was to compare an input-only (reading) programme with an output-only (writing) programme, within the paradigms of Krashen's *Input Hypothesis* (1982, 1985, 1989) and Swain's *Output Hypothesis* (1985, 1993). The experiment is very well designed. Students in the reading and writing groups were each given eight tasks, to be performed over a period of 24 weeks, which it was estimated would take approximately the same amount of time to perform. Reading-group students read eight graded readers, and writing-group students wrote eight compositions. A third group of students, acting as a control group, were required to hand in eight mathematics assignments. Teacher feedback was minimal, the intention being to maintain student motivation without providing explicit, contaminating, additional instruction. The writing group's compositions were given an impressionistic holistic grade and returned to the students with no more than a brief positive comment. Reading students completed a short review form for each book they read, with only minimal writing required, primarily to monitor their reading, but also to provide a physical outcome which might be treated in parallel fashion to the writing students' physical outcome — which is to say that these short reviews were also assigned letter grades. Since feedback on the mathematics assignments was not related to language use, these were corrected (thus implicitly graded) before being returned.

Students wrote 30-minute pre- and post-compositions in class, entitled "My favourite

person". Compositions were graded using the ESL Composition Profile (Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey, 1981) and gain scores were compared by Form level, across all 4 forms together, and on the range of constructs represented by the ESL Composition Profile scoring method. Unfortunately, it is at this stage that a very promising research design is spoiled by over-complex, unmanageable statistics and a poor choice of rating instrument.

Univariate statistics tests (we do not know which) showed a significant difference amongst the three treatment groups in performance on *content*, *language use* and *impression total*. However, the variable labelled *"impression total"* was not, in fact, a separate variable, but a composite of scores on the five ESL Composition Profile constructs of *content*, *organization*, *vocabulary*, *language use* and *mechanics*. Since *content* and *language use* are heavily weighted by the ESL Composition Profile, accounting, jointly, for a possible 55 points, compared to the maximum possible 45 points accorded to the three other constructs between them, it is not surprising that the composite score merely repeats the findings for *content* and *language use*. It is not valid to consider this as a third finding.

In any case, the ESL Composition Profile may not be very appropriate for the writing task which was set. It is hard to imagine what an improvement in *content* might, more precisely, consist of, in a composition entitled "My favourite person", given that Jacobs *et al.* provide a set of descriptors for this ranging from "knowledgeable" to "does not show knowledge of subject" (1981: 90). It seems rather harsh to penalise a student, in an L2 language exercise, for not having much information about someone they nonetheless rather like. The ESL Composition Profile was originally developed for use with ESL students attending Texas A&M (Agricultural and Mechanical) University, and was field-tested in universities throughout the United States, presumably with students writing to an academic programme within which *"content"* requirements were quite specific and almost as important as language requirements. EFL, which generally involves the teaching of a language for its own sake, is *not* the same as ESL, in which subject classes are taught through the medium of L2 English.

Disregarding this weakness, the other constructs of the ESL Composition Profile may still provide useful information. However, some of the findings in Tsang's study are not very explicitly reported. A table in the report shows pre- and post- composite scores for the three treatment groups at the four Form levels, showing clearly that by far the greatest gains were

made by Form 1, with group-mean gain scores ranging from 13.2 to 17.4, across the three treatment groups, compared to group-mean gain scores ranging from 0.5 to 7.2, 2 to 7.1 and 0.4 to 4.1 for Forms 2, 3 and 4, respectively. A second table shows gains in the five individual ESL Composition Profile constructs, by treatment group, taking all four forms together, and a third table shows gains in these five constructs, by Form level, but taking all three treatment groups together. Frustratingly, we are not told the gains made in each construct, by each treatment group, at each of the four Form levels, despite the fact that an initial MANCOVA, for the set of five constructs, found significant main effects for treatment and for Form level. Although the MANCOVA showed no significant interaction between Form and Treatment, this means only that the three treatments did not produce differing effects at different Form levels, not that gains were similar in magnitude at the four levels. It is not particularly enlightening to know the gains made in each of the five constructs, at each Form level, for all three treatment groups added together, when the point of the research is to identify the effects of the treatments. In addition, Tsang makes claims which are hard to unravel, such as "both Programs A (Mathematics) and B (Reading) brought about a significant effect in language use, with Program B (Reading) showing a statistically greater effect than Program A (Mathematics), the latter in fact performed significantly worse in the post-test" (1996: 225). The means for the mathematics group for *language use* were 13.97 and 15.20 for the pre- and post-test respectively, whilst these were 14.21 and 16.41 for the reading group.

In spite of these shortcomings, we may deduce from Tsang's study that the reading group outperformed the writing group and the mathematics group, probably on *language use*, and probably largely due to gains at the lowest level, Form 1. One other interesting point is that no differences were found amongst the three treatment groups for *vocabulary*.

By far the clearest study to compare extensive reading as a methodology against a rival methodology is that of Bell (2001). Two groups of elementary learners (total N = 26), in Yemen, were exposed to different types of reading programmes over a period of two semesters. One group followed an intensive reading programme consisting of short reading passages of approximately 300 words, each accompanied by a battery of exploitation activities such as comprehension questions, discussion questions, gap-fill, word-building and guided composition exercises. The second group followed a normal extensive reading programme, using graded readers, with only simple record-keeping as an additional activity.

The reading programmes comprised one quarter of the English class time, totalling 36 hours over the period of the experiment. It must be assumed that, apart from the different reading programmes, students were subject to the same instruction methods and materials. Although a lot of reading was carried out in class, both groups were also given considerable amounts of homework. Attempts were made to ensure that one group did not spend significantly longer contact time with their particular type of reading materials than the other group. Students were asked to report how much time they spent with these, and a t-test found no significant difference in contact time between the two groups.

Pre- and post-tests for reading speed and reading comprehension showed conclusively that the extensive reading group made considerably more progress than the intensive reading group. For reading speed, the intensive reading group achieved mean pre- and post- speeds of 78.45 and 92.54 words per minute, whilst the extensive reading group achieved 68.10 and 127.53 words per minute. ANCOVA was not performed, but a t-test on the pre-scores showed no significant difference between the groups, whereas a t-test on the post-scores showed a between-groups difference which was significant at $p < .001$. On a gap-fill reading comprehension test, the intensive reading group achieved mean pre- and post- scores of 50.88% and 60.72%, compared to mean pre- and post- scores of 50% and 81.12% for the extensive reading group. On a different set of reading comprehension scores, derived from multiple-choice and true/false items, mean pre- and post- scores were 45.45% and 59.90%, and 41.86% and 80.81% for the intensive and extensive reading groups respectively. Independent-samples t-tests showed, in both cases, no significant difference between the groups on their pre-test scores, but a significant difference on the post-test scores.

These are very encouraging findings, and more studies of this type are needed to confirm the consistency of these kinds of effects. The two groups were very small, consisting only of 12 intensive readers and 14 extensive readers. The reading comprehension tests may not have been optimally suited to measuring reading comprehension. Moreover, as Bell notes, the students in the extensive reading group were very aware of being involved in a special, and, in fact, high-profile, reading programme. Facilities were those of the British Council Language Centre in Sana'a, Yemen, and the extensive reading treatment included regular visits to a probably rather nice library. Motivation may have played a very important role in the resulting improvement in performance. Nonetheless, the difference in improvement between intensive and extensive reading groups was indisputably very large indeed.

2.3.3 Extensive reading as part of a syllabus

In a series of three rather briefly-reported experiments, Mason and Krashen (1997) investigated the effects of an extensive reading programme involving the replacement of a pre-existing methodology *in* class, combined with large doses of extensive reading *out of* class. In other words, the experiments compared methodology effects and extra-input effects in such a way that these could not be separated.

The first experiment compared the pre- and post- performance on a 10th-word deletion cloze test of two intact EFL reading classes at a women's university in Japan. Students were exposed to the same "traditional" methodology for one semester, but during the second semester, whilst one group continued with this, the method was replaced by an extensive reading programme for the second group. Students in this group were required to read approximately 50 graded readers over the semester, to write a short synopsis in English of each book they had read and to keep a reading diary in Japanese. It is not clear from the report whether teachers gave feedback on the written synopses, nor what other concurrent L2 instruction was given to students in their other classes. (One must suppose that L2 English instruction at the university did not consist solely of one reading class a week.)

Statistics tests performed were slightly unorthodox, consisting of t-tests on gain scores, in preference to what would be the more usual ANCOVA procedure. (Although the authors of the article point out that the use of ANCOVA is controversial for intact groups, the use of t-tests on gain scores can be no less controversial since this is merely approaching the problem technically differently, but coming from exactly the same theoretical direction.) Although it may be tempting to believe that the reported result of a significant difference between the groups in favour of the extensive reading students provides reliable evidence of the superior effects of extensive reading, a number of small details are unsettling to the reader. It is not made clear, for example, why results from only 20 subjects were (randomly, we are told) selected for the study out of a class of 30, when 30 is a more acceptable number for statistical analysis; equally, it is not clear how a 10th-word deletion ratio applied to a 1,600-word reading passage resulted in a 100-item cloze test.

The second study was a repetition of the first, but on a slightly larger scale, involving four

classes, one control and one experimental from each of two tertiary institutions in Japan, over one year. In addition to the same 100-item cloze-test, experimental students wrote a book summary at the beginning of the course and a second book summary, on a different book, at the end of the course. These were judged by three native speakers (who may or may not have been teachers — we are not told) as Good, Average or Not Good.

T-tests on gain scores in the cloze-test showed significant results in favour of the extensive reading groups at both institutions. Writing (book summary) judgements were compared, for each judge separately (no inter-rater reliability, or consistency, was even attempted), using a 2 x 2 Chi-square design. This showed significant differences in numbers of pre- and post-programme summaries assigned by the raters to each category, with all three raters assigning significantly higher numbers of post-programme summaries to the Good category, and of pre-programme summaries to a collapsed Average + Not Good category. We are not told why the two categories of Average and Not Good were collapsed, except to say that this was "to allow statistical analysis" (1997: 94). Since Chi-square may equally well use a 2 x 3 design, which would have allowed the three categories to be retained, this is not sufficient explanation. Moreover, many writers feel that, in any case, a 2 x 2 design tends to produce an overestimate of the Chi-square value (Pallant, 2001).

Claims that the improvement in summary writing in the second experiment was caused by extensive reading were just as amenable to the hypothesis that this occurred because of the additional practice in summary writing experienced by the reading group. (Experimental students may have written between 30 and 100 summaries over one year; we are not told how many.) For this reason, the third experiment included two reading groups and one control group: the first reading group wrote summaries in English and the second wrote summaries in Japanese. Testing procedures were the same. Although the English-summaries group performed significantly better on the Cloze than the control group, and the Japanese-summaries group did not, there was no significant difference between the two reading groups. On the writing test, one of the two raters judged the Japanese-summaries group to have made the most progress, followed by the English-summaries group, and, lastly, the control group. The second judge, however, rated the Japanese-summaries group as having made most progress, followed by the control group, and then the English-summaries group. It should be noted, however, that the Japanese-summaries group was the lowest-rated group at the outset, and so could show progress more easily.

It is difficult to know exactly what to make of these reports. There is no between-rater reliability, and one has to be rather wary of some of the statistics. For example, it is not possible for a t-test comparing one group of 40 students and one group of 36 students (in the third experiment) to produce results displaying a reported 106 degrees of freedom. However, since similar results were reported for the respective gains made on the cloze by reading and control groups across the three experiments, there is probably some degree of reliability to these results. The greater gains made by the extensive reading groups may have been caused by additional exposure, the methodology, the special treatment (one of the report authors taught the extensive reading classes), or by a combination of these. Results on the writing test, however, must be considered as no more than suggestive, given the lack of any inter- or intra-rater reliability checks and the inconsistency of the findings.

An interesting study is that of Yang (2001), who deployed extensive reading as part of a programme of L2 English evening classes for non-academic adult learners in Hong Kong, using two Miss Marple novels by Agatha Christie. The reading was done outside class, at a rate of about 40 pages per week, over a period of 15 weeks. Students used one hour of their once-a-week three-hour class, which would normally have been given over to the usual evening-class textbook instruction, in order to service and exploit the reading scheme. This took the form of discussion groups, oral reports and some text-based language discussion. Of four evening-class groups of approximately 30 students each, two used the extensive reading method, nonetheless retaining two hours of textbook-based instruction, and two acted as controls, using only the textbook.

A one-hour multiple-choice sentence completion test with 100 items designed to test grammar, sentence structure and usage — in other words a general proficiency test — was taken at the beginning of the course and again, without prior knowledge, during the last class. Test results were impressively clear and consistent. In the pre-test, the four groups achieved mean scores of 62.3, 62.1 (experimental groups) 61.9 and 61.3 (control groups). This homogeneity of ability at the outset may be accounted for by the fact that all course participants had had to achieve a certain grade in English in the Hong Kong Certificate in Education Examination, equivalent to 450 on the old TOEFL system, before being admitted to the course. Post-test means were 75.3 and 73.7 for the two experimental groups, and 66.7 and 67.0 for the control groups. ANOVA showed a significant difference amongst groups

on these post-scores, and, comparing the scores of all 60 experimental students with those of all 60 control students, the experimental students were found to have performed significantly better.

In an extension of the study, 50 experimental students responded to a questionnaire on reading-related activities. Of 20 questions, the majority addressed affective issues and reading habits. A few, however, addressed language learning issues, namely whether the students thought the reading programme had helped their English language learning in general, and their vocabulary, grammar and writing in particular. Forty of the 50 felt the scheme had benefited their English in general. What is perhaps more interesting is that what students felt had most benefited was their writing (38 respondents), followed by their grammar (33 respondents), and, lastly, their vocabulary (26 respondents). In fact, of the whole range of questions, only two attracted lower agreement rates than the question of whether reading had helped with vocabulary learning — "I have read at least one full-length English novel before" (10 agreements) and "When I read these two novels I always consulted the dictionary when I came across a new word" (9 agreements). When a smaller group of students were interviewed and asked in what way, specifically, they felt their writing to have benefited, they could not say, but simply felt "writing came easier than before" (2001: 459).

2.4 Summary of findings from field studies

In conclusion, what strong evidence for the beneficial effects of L2 extensive reading can we draw from the above studies? The best study, in terms of breadth of investigation, and validity and reliability of research procedures, remains that of Elley and Mangubhai (1981, 1983), reporting on the Fiji Book Flood. However, we should not forget that every significant difference between the reading groups and the control group was the result of comparison with a particularly weak, input-deficient methodology which had already been identified as being in urgent need of replacement. Effects of the book-based methodologies were striking, and much credit must go to the innovators, but comparison with a more communicative, learner-centred, motivating, or even just less input-deprived methodology probably would not have given rise to such striking results.

Lai (1993) provides evidence that students *can* derive intake from their own autonomous

reading processes. Results from the Elley and Mangubhai study support this. Yang (2001) provides evidence that general proficiency may be improved through a reading programme which is integrated into a course design. Mason and Krashen's (1997) studies also point in this direction, although they may be less reliable. Yu (2000) and Bell (2001) give good evidence for an improvement in reading comprehension when extensive reading is used concurrently with other classroom instruction. Bell's study shows, too, that extensive reading can be superior to intensive reading in promoting the development of reading speed. Lai's study also shows that reading speed may improve as a result of extensive reading, but her results cannot be taken to show that extensive reading is better in this than any other method of instruction, since there was no comparison control group.

As regards writing, apart from Elley and Mangubhai's results, we have little more than Lai's (1993) report and Tsang's (1996) somewhat unclear findings to support the theory that extensive reading may lead to overall improvement in writing quality and facility. Although Hafiz and Tudor make that claim, the absence of any reliable control group and the presence of concurrent classroom instruction in both their studies means credit cannot be conclusively accorded to extensive reading. Mason and Krashen's writing evaluation methodology was not reliable enough for the claim to be creditably made from evidence taken from *their* studies. The voices of Yang's (2001) interviewees might, however, be added to the mix, in a non-experimental supporting role. Again, other than Elley and Mangubhai's findings (for *word knowledge*), there is a noticeable lack of sound experimental evidence for the acquisition of vocabulary simply through L2 reading. This is the experimental evidence which we have in favour of L2 extensive reading, more than 20 years on from the Fiji Book Flood.

2.5 Incidental learning of vocabulary through extensive reading

Can a previously unknown word be learned simply from encountering it in the course of one's reading? In L1, the matter is clear. Since we do not learn the vast majority of the words we use in our native language through direct, focused instruction, it follows that we learn these incidentally, through hearing them or seeing them written. Since many of these words are not used in everyday conversations, it also follows that written text is an important vehicle for incidental L1 vocabulary acquisition. In fact, in their seminal 1985 study, Nagy,

Herman and Anderson estimated that "the number of words the typical middle grade child learns in a year from context during reading is between 750 and 5,500; the point-value estimate is 3,125" (1985: 250). Even taking into account that, by this, Nagy *et al.* do not mean word families (for example, "moodily" is regarded as a word in its own right), this is still an impressive tally. More recently, Nation and Waring have estimated that, until the age of approximately 20, native speakers of English "will add roughly 1,000 word families a year to their vocabulary size" (1997: 7). Many of these will be acquired through reading.

In L2, however, the situation is somewhat less unequivocal, and the evidence from L2 reading studies, as we have just seen, has been less than overwhelming. Nonetheless, there is continued support amongst researchers and practitioners for the theory that the incidental vocabulary learning hypothesis (Nagy and Herman, 1985) has at least some relevance within the L2 context (*e.g.* Nuttall, 1996; Nation, 2001). The first step has been to show, firstly, that such learning *can* occur. More recently, researchers have attempted to uncover those factors which influence the chances of a word being successfully acquired through incidental learning whilst reading in an L2.

In a much-cited study, Pitts, White and Krashen (1989) tested whether incidental vocabulary learning can occur in an L2 by using as an experimental text part of *A Clockwork Orange*, in which the protagonists use an invented slang, or *nadsat*. In the original L1 *Clockwork Orange* study (Saragi, Nation and Meister, 1978) adult native speakers of English who read the whole novel thinking only they would later be given a test of comprehension and literary criticism, achieved a mean score of 76 % correct meaning identification a few days later on a surprise multiple-choice test covering 90 nadsat words. In the L2 study, one experimental group of intermediate ESL learners who read only the first two chapters of *A Clockwork Orange* and were tested 10 minutes after completion of the reading task achieved a mean score of 1.81 on a multiple-choice test of 28 nadsat words, and a second experimental group achieved a mean score of 2.42 on a test of 30 nadsat words. (Both these scores were adjusted using a correction-for-guessing formula and so these represent estimated *true* scores which have not been inflated by the one-in-four chance a student had of guessing correctly on the words he did not know.) This represents an approximate average of 7 % learning of nadsat words from meaning-focused reading — a rather small figure by comparison to the learning of the L1 group in the original study.

Other research has reported similar small gains. Day, Omura and Hiramatsu (1991) found that experimental students in a Japanese high school and undergraduate EFL students who read an adapted version of a 1,032-word story achieved mean scores on a multiple-choice test of just one and three correct answers (out of 17) more than control students who had not even seen the text. Dupuy and Krashen (1993) reported slightly better results with college-level students of French as a foreign language who, after reading 15 pages of text, following on from having watched a related video extract, achieved on average 6.6 more correct answers on a multiple-choice test of 30 items than control students who had neither read the text not seen the video extract. Both these studies used real words and did not take account of the possibility that, in any case, some of the tested vocabulary may not have been new to some of the study participants.

In none of the three L2 studies above was the long-term durability of the learned vocabulary measured. Rott (1999), investigating the incidental acquisition of 12 lexical items through reading (previously verified as unknown to the study participants) with 67 university-level learners of German as a foreign language, found that the number of newly-learned lexical items which could be actively produced deteriorated significantly over a period of four weeks. Waring and Takaki (2003), working with a group of 15 university-level Japanese EFL learners, found that on three tests measuring different degrees of receptive word knowledge of 25 incidentally-learned invented words (word-form recognition, prompted meaning recognition and unprompted meaning recognition) performance deteriorated dramatically after only one week. Thus, it seems that some of the small gains which have been reported may also be precarious.

There are some very obvious reasons why incidental word acquisition from reading in an L2 may be less successful than in an L1. One cannot learn a new word without first comprehending it. The L1 eighth-grade students in the study by Nagy *et al.* would have had, by Nation and Waring' s (1997) estimate, in the region of 13,000 to 14,000 word-families already at their command. The adult L1 participants in the original *Clockwork Orange* study (Saragi *et al.*, 1978) could most likely be expected to have an in-depth, unquestioned and firmly-established knowledge of every word in the text except the *nadsat* words. An average intermediate L2 student, on the other hand, may have a word base of some 2,000 to 3,000 word families, much of which is not yet firmly established or quickly accessed.

Nation (2001) estimates that a word-base of 2,000 words in English represents approximately 90% coverage of novels, 80% coverage of newspapers and less than 80% coverage of academic texts. This leaves rather a lot of unknown words, and the L2 reader may not have the linguistic tools to establish a definite meaning for a new word. Where a probable meaning *is* established, the L2 reader cannot have the assurance of the L1 reader that the surrounding context has been correctly interpreted, and this may have a strong psychological effect, inhibiting commitment to memory. Even in cases where the L2 reader may recognise every word in the immediate environment of an unknown word, knowledge of these context-words may not be as yet well-formed or firmly established. In addition, interpretation of the grammar of a sentence may be effortful, also causing insecurity, and the slowing down of the reading process which is caused by this will also result in the loss of inter-sentential clues, since working memory is very short-lived. There may, in fact, be a proficiency threshold, below which the kind of fluent reading which will engender incidental vocabulary acquisition cannot take place. Laufer (1992) has suggested that knowledge of 3,000 word-families (approximately 5,000 words) may be a minimum vocabulary base for comprehension of an unsimplified L2 English text. Hu and Nation (2000) have suggested that 98% coverage of text-vocabulary is needed for satisfactory text comprehension. If we assume that incidental vocabulary acquisition will not take place *before* comprehension, then these figures might also be suggested vocabulary knowledge minimums for an incidental acquisition threshold. Nor should we forget that there is more to text comprehension than just vocabulary knowledge. Knowledge of grammar may also have some impact on interpretation of unknown words.

One other strand of research has investigated the influence of number of encounters on eventual learning of a new word merely from reading. Rott (1999) found that some learning occurred after only two encounters, but that six encounters engendered more durable results. Horst, Cobb and Meara found that "sizable learning gains can be expected to occur consistently for items that are repeated eight times or more" (1998: 215). From a particularly rigorous and well-planned experiment, Waring and Takaki (2003) concluded that for a learner to have a 50% chance of recognizing a word form three months after the experiment, at least eight encounters were necessary, but that the chances of form-meaning recognition after three months were only approximately one in ten even if a word had been encountered more than 18 times, and next to zero for words encountered fewer than five times.

What the authors of the frequency studies discussed above have not explored is that figures representing estimated numbers of encounters necessary for incidental word learning may constitute partly a frequency requirement and partly, also, a comprehension requirement, such that a word needs to be met on average a certain number of times, the learner adding each time to previously gained partial knowledge, before an adequate understanding of the word is reached. Clearly, not all contexts are equal, and some contexts will provide better clues to the meaning of an unknown word than others. Not only is the reader's understanding of preceding text very important, but also, more locally, the preciseness of immediate semantic environment. Compare, for example, "She was carrying a *XXX*" with "She was wearing a *XXX*". It is likely, too, that the properties of some words lend themselves better to incidental learning than others. For instance, adjectives may be harder for the L2 reader to pin down than nouns or verbs. Consider the following three short sentences: "She was *XXX*ing a red ribbon in her hair", "She was wearing a red *XXX* in her hair" and "She was wearing a *XXX* ribbon in her hair." Assuming knowledge of all the other words, the unknown word in the first sentence can have very little range of possible meanings; the unknown word in the second sentence has a slightly bigger range of possible meanings (*e.g.* types of flower, any kind of hair ornament); the unknown word in the third sentence, however, has considerably more possibilities (*e.g.* green, pretty, large, shiny, dirty, new, torn, ridiculous, borrowed *etc*). This third sentence, in fact, may not provide a very fruitful encounter with the unknown word. Schmitt, having tracked the acquisition of 11 words over the course of a year for three adult L2 learners, found that nouns and verbs were mastered before adjectives and adverbs, tentatively concluding that "adjective and adverb forms are not so readily learned from general exposure (perhaps due to their lower frequency of occurrence) and hence might be good candidates for explicit instruction" (1998: 307). As our illustration above shows, however, the word class itself may play an equally, if not more, important role than general frequency. Previously unknown lexis which occurs in contexts which are relatively unrestrictive as to range of possible meanings has considerably less chance of being learned than lexis which occurs in more restrictive contexts. Nouns and verbs tend to appear in more meaning-restrictive contexts than adjectives and adverbs.

Although Horst *et al.* (1998) did observe that many of the words in their study which had high gain scores (*i.e.* which were more successfully learned) were concrete nouns, the experiments outlined above did not attempt to categorize words for other factors which might affect their ease of acquisition. This may help to explain any apparent disparity

amongst the findings. It could also be argued that Rott's 12 lexical items, Waring and Takaki's 25 words, and Horst *et al.*'s 23 words may have performed quite differently in other surroundings. One other point worth noting is that in *all* of the L2 incidental vocabulary acquisition studies the experimental target words may not have been the only words to have been learned incidentally during the experiments. Other words may also have been learned, varying from individual to individual, which were neither pre- nor post-tested, and a student with a recorded gain of one or two new words may, in fact, have learned 10 or 20 new words.

Where does this leave extensive reading? Whilst it is abundantly clear that we cannot expect L2 reading to lead to the same kind of incidental vocabulary gains that have been found to exist in L1, this does not necessarily mean there are no vocabulary benefits to be gained from L2 extensive reading. If it is true that incidental vocabulary learning will not occur unless there is an existing knowledge of at least 98% of the text vocabulary (since this is when satisfactory comprehension occurs), then the use of graded or simplified text makes it possible for such favourable conditions to be engineered at different stages of L2 proficiency, and this is a strong argument in favour of the use of simplified texts. Moreover, as Waring and Takaki point out, vocabulary learning need not be an all or nothing affair, and reading may help to establish already known vocabulary. The studies above focused on previously unknown words, and we are told of the effects which reading one, in some instances relatively long, text had on these. We do not, however, know the effect of reading the text on the hundreds of *other* words which also made up the text. We do not have tests which distinguish between 60% certainty and 70% certainty; we do not have tests which can distinguish between varying *speeds* of vocabulary access; vocabulary tests do not give measures of *breadth* of knowledge of words or of ability to manipulate and place a word with exactitude in varying contexts, and to recognise sub-textual connotations. Cooper (1984), for example, found that one of the features which best discriminated practised from unpractised L2 readers in a group of university undergraduates was a much improved grasp of lexical cohesion, where relationships between lexical items, often across sentences, are components of overall textual coherence. (An example of this might be "The square was filled with cars and taxis. None of the vehicles had any lights on.") The L1 reader who incidentally acquires the only four or five previously unknown words in a text may have made no greater improvement to his vocabulary than the L2 reader who has visibly acquired no new words, but who has spread his gain more thinly over a very much greater number of

words. The incidental acquiring of even one new word whilst so doing may, in fact, represent no small achievement.

2.6 Graded text as an input medium

The use of simplified text as an input medium in L2 teaching has been criticized on two fronts. The first contention is that simplified text is not *authentic*, therefore cannot engender an authentic reading experience; the second is that, in any case, simplifying a text may not necessarily render it more easily understood by the L2 reader. Whilst the first of these claims is primarily theoretical, the second has been explored by a number of researchers in tightly-controlled experiments with simplified and non-simplified texts.

Honeyfield may have been one of the first to argue that simplification processes rob a text of many of its original, some would term *authentic*, properties, maintaining that simplification produces "material which differs significantly from normal English in the areas of information distribution ... and communicative structure" (1977: 431). This may well be the case. Indeed, one of the purposes of text simplification for the L2 consumer is to produce more comprehensible text by operating checks on any kind of textual complexity, be they syntactic, lexical or rhetorical. It would be most unexpected if such controls did *not* produce different patterns of information distribution and communicative structure.

There are two answers to such a criticism. The first is that the L2 reader is not an L1 reader. The relationship between the L1 reader and his L1 text cannot, in any case, be replicated with the simple replacement of the L1 reader by the L2 reader. The text itself may retain its *authenticity*, but an L2 readership probably is not its intended, *authentic*, audience. As Claridge (2005) points out, what are (subconsciously) perceived by the L1 reader as relatively frequent and infrequent language items, with low and high information loads, giving the variety of information density which Honeyfield maintains to be a distinctive feature of the authentic reading process, may not be perceived in quite the same fashion by the L2 reader. Words and structures which are commonplace for the L1 reader and hardly need to be attended to may become a point of focus for the L2 reader. Different *kinds* of information extraction must take place. In L1 reading, assuming the mechanics of decoding and word *access* (recognition) to have been mastered, the work which is required of the

reader is the parsing of a string of accessed words into units of meaning. In L2, the reader may use accessed words to postulate meaning, but may also use postulated meaning to access words. There is a considerable difference between reading a text in one's mother tongue and attempting to understand a text through the medium of a second or foreign language in which one may or may not have a high level of proficiency.

The second answer to this kind of criticism is, accepting that there may be a difference between simplified and non-simplified texts, why should this matter? The broader question of the relationship between authentic and simplified text is one which has been commented on by a number of academics. Widdowson, for example, makes the distinction between *simplified versions* of pre-existing texts and *simple accounts*, which may be original texts or may be recasts of another, source, text, describing simple accounts as "a genuine instance of discourse, designed to meet a communicative purpose" (1978: 88). From this we may conclude that Widdowson views simple accounts as instances of *authentic* discourse, but does not view simplified versions as such. Davies makes the point that this is a difficult distinction to maintain since "most rewritings will be partly simple accounts and partly simplified versions" (1984: 183). In fact, the distinction between what is an authentic text and what is not is something of a vexed question. Is a version of a classic novel which has been rewritten for children not an authentic English text? Is an English translation of a Russian novel not an authentic English text? Many formerly opaque bureaucratic documents, such as social benefits claim forms, have now been rewritten in plain English so that the users might understand them better. Some are written differently for different target readers. Are these not authentic documents? Perhaps more politically, is English written by a non-native speaker, assuming the grammar and lexical choices to be correct, not authentic English? Does authenticity of language use arrive at the same time as full competence? How do we know — for native-speaker children as well as for L2 speakers — when full competence has been achieved? One of the best writers in the English language, Joseph Conrad, was a native speaker of Polish, not English. Are we to regard his novels as *inauthentic* whilst regarding the first attempted written birthday message of the four-year-old native speaker as *authentic* model text?

These questions are perhaps more philosophical than of any practical use. A more useful question is whether the language of a text conforms to the rules of the linguistic system. Rhetorical and discourse patterns — the "communicative structure" of Honeyfield — are not

rules, but conscious or subconscious authorial choices. Simplified texts may represent one small part of the larger linguistic system. As Tommola writes, "the code is not affected, the learners are not presented with a simpler language system but with a restricted sample of the full system" (1979, cited in Davies 1984: 183).

A number of studies have investigated the effects on L2 learners of various types of text simplification. Blau (1982) examined the effects of shortening sentences on the comprehension of ESL students in Puerto Rico. Three versions for each of 18 short reading passages were prepared, using the same vocabulary items but different syntax. The first, simplest, version was characterized as containing only short, simple sentences, the second as containing "complex sentences with clues to underlying relationships left intact" (1982: 517), and the third contained complex sentences without these clues. From the performance of groups of students at each of two proficiency levels on multiple-choice questions following each reading passage, the second version, containing longer sentences than the first version, was found to have been consistently better understood. From this Blau concluded that "the relationship between syntax and readability ... is not so strong as may have been expected" (*ibid*: 527) and that simple sentence structure was not necessarily more easily understood.

We should not conclude from this, however, that simplified text may be more difficult to understand than unsimplified text. If a group of students finds a text more difficult after it has been simplified, then, as Lynch (1996) notes, it has *not* been simplified. It has been tampered with in such a way as to make it less readable. What *was* simplified in Blau's study was syntax at the level of the individual sentence. Individual sentences were probably very *well* understood by students on a stand-alone basis. However, a text is not a series of stand-alone sentences. This is what makes it a text. What happened to the texts in Blau's experiment was that, in order to keep sentences as short as possible, nearly all connectors and cohesive markers were omitted, resulting in a lack of inter-sentence connection clues and producing a series of short, sharp, apparently unconnected sentences.

This, in fact, is exactly the kind of "simplification" which has given graded text a bad name. Shortening sentences often has the effect of separating very closely related propositions and putting these in two or more adjacent sentences. This places not just one, but two, *additional* burdens on the comprehension process. Firstly, since English grammar generally does not allow coordinators and subordinators such as *but, so* and *because* to be placed at the

beginning of a simple sentence, these are lost to the text, and the reader is left to supply the relationship between the relevant propositions for himself. (For example "He hated his grandmother even though she had bought him a car" makes little immediate sense if it is divided into "He hated his grandmother. She had bought him a car.") Secondly, in order to comprehend the full meaning of the *extended* proposition, the reader must now hold, not one, but several sentences in working memory. Dividing a single extended proposition into two or more parts simply results in the reader having to hold the same amount of text in working memory, but with the difference that reading fluency, and fluency of meaning extraction, has been disrupted by an inappropriate pause marker. Thus, chopping up a text with full-stops may not be particularly helpful.

Readability formulas, which use sentence length as one of the differentiating criteria for various difficulty levels, are often invoked as prescribing short sentences for easier, or more readable, text. This is a misrepresentation of both the intended purpose of readability formulas and the techniques which are used to derive them. These were developed as *post hoc* measuring sticks for text difficulty using correlations between surface text features, such as average word and sentence lengths, and text difficulty as measured either by native-speaker judgements or cloze performance. Short sentences were found generally to co-occur with simple text. However, short sentences also tend to co-occur with simple verb constructions, SVO word order and the absence of embedded clauses and complex noun phrases. Short sentences are also indirectly linked to the use of more commonplace vocabulary and simpler content, since longer sentences tend to appear more often in more sophisticated literary, expository or academic texts. Short sentences do not *cause* textual simplicity.

Tweissi (1998) investigated various combinations of type and amount of simplification: lexical simplification only, syntactic simplification only, these two together in *full* simplification, and these two together but with only half the number of text changes, resulting in *semi*-simplification. For a group of English learners enrolled at a tertiary college in Oman, Tweissi found that lexical simplification resulted in better text comprehension than syntactic simplification, which, in turn gave better results than semi-simplification. The difference between these was not great, however. Surprisingly, *full* simplification did not do so well as any of these, although this still produced better text comprehension than no simplification. From this Tweissi concluded that there may be such a thing as too much simplification. It should nonetheless be noted that the five groups of 40 students who

participated in the experiment were not pre-tested, but had all obtained scores of between 40% and 60% on a placement reading test administered we do not know how long before the experiment. The groups may have been of unequal reading proficiency at the outset.

Working with Japanese undergraduates, Yano, Long and Ross (1994) investigated the effects on reading comprehension of two types of text modification: simplification and facilitating elaboration of the kind observed more often in oral discourse "where redundancy and explicitness compensate for unknown linguistic items" (1994: 189). The researchers found that both types of modification facilitated text comprehension as measured by a 30-item multiple-choice test, and that there was no significant difference between the two. They further found that elaboration had a more facilitating effect than simplification on test-items requiring inference, although this may be a somewhat precarious finding, given that only two items on the test were agreed upon by judges as being *inference items*, and the example of this which is given in the appendices is highly questionable as a possible discriminator amongst the three texts. In the example we are given the correct choice of answer, out of four, is *Catfish can live on land for as long as in water* which must be extracted from: *Catfish have both gills for use under water and lungs for use on land, where they can breathe for twelve hours or more* (unmodified); *Catfish have both gills and lungs. The gills are used for breathing under water. The lungs are for use on land. The fish can breathe on land for twelve hours or more* (simplified); or *Catfish have two systems for breathing: gills, like other fish, for use under water, and lungs, like people, for use on land, where they can breathe for twelve hours or more* (elaborated). It is hard to see how any one of these texts provides more help than the others in making the inference that 12 hours is half a day.

Studies such as the above are extremely difficult to do. A text might be regarded as a kind of ecosystem within which almost any change will necessitate other changes, such that it is very hard to modify any one text feature whilst keeping others constant. This is perhaps more evident with syntax modifications than vocabulary substitutions. Blau, for example, in order to break single conditional "if-" sentences into two short sentences, fronted one of these with "suppose", as in *Suppose the manufacturer and the market are a long distance apart. This can be a big expense*. Although the intention here was text simplification, the vocabulary item "suppose" is far less likely to be known to learners than "if". Simplification can be very subjective, and what one experimenter may consider simple, another might not. Tweissi, for instance, considers *the qualified pilot faces a constant risk. The risk is that of losing his*

licence to be simpler than *the qualified pilot faces the constant risk of losing his licence.* Personally, I do not see the first text as being simpler than the second, given that it contains the structure "is that of + ...ing", which is no simpler than "(the risk) of + ...ing", the construction it replaces. The "simpler" text also requires additional referring of the substitute pro-form "that" to the noun "risk". In the examples provided by Yano *et al.*, above, the simplified text is the only one which uses the passive, and both the simplified and elaborated versions use the structure "for + ...ing", which the unmodified text does not. The elaborated version also introduces the relatively low-frequency word "system".

These experiments may be very interesting exercises, but, ultimately, they are barely generalizable beyond the actual texts and specific comprehension questions used in the experiment. It is only common sense that an L2 reader is likely to have less difficulty comprehending a text which uses grammar and vocabulary he knows than one which uses grammar and vocabulary he does not know, particularly if he is reading at some speed. If one looks closely enough, one will usually find that simplified texts which do *not* aid comprehension have had their original cohesion and coherence features disrupted in some way.

There is also the question of the reader. What we, as native speakers of English, may consider disruptive to text comprehension may not have the same effect on L2 readers. In Blau's experiment the learners were native speakers of Spanish, a language which may be considered closely analogous to English in terms of both local grammar and overall communicative structure. Other languages, however, may be quite different. For instance, some scholars maintain that Chinese does not possess the necessary grammar to form non-defining relative clauses and that these do not exist in the Chinese language (*e.g.* Del Gobbo, 2005). Thus, the flattening out of these in an English text, in the cause of simplification, may render the text structure *closer* to what a Chinese reader is used to. Similarly, Chinese does not make use of markers of cohesion in the way that English does, and such markers may be less critical to Chinese L1 speakers' text comprehension strategies, whilst reading in English, than they are to the text comprehension strategies of native English speakers.

Text features, and different types of text modification, may also have differing effects on the L2 reader's comprehension at different developmental stages. It is quite possible that what was helpful at a certain stage may be distracting and unhelpful once the reader has gone

beyond that stage. When one knows the past tense perfectly well, it may be irritating to read a text which avoids its use. If one has not yet learnt the past tense, one is not irritated. As an aid to text comprehension, even the chopped-up sentence, causing, as it does, breaks in the parsing process which will actively *hinder* the more fluent reader, may be helpful to the less fluent beginning L1 Chinese reader of English. Until a great deal more research has been done in this area we cannot categorically say in what ways the non-native English speaker's comprehension of a text will be disrupted by any particular type or instance of text modification, but must conclude, with Lynch, that "the success of simplification can only be judged by reference to a particular learner or group of learners" (1996: 29).

To date, there has been no experimental evidence to show that using simplified text in the first stages of language learning disrupts later progress on to unsimplified texts or causes damage to the learner's L2 linguistic system. On the contrary, the number of opportunities afforded by extensive reading for encountering correct, if controlled, uses of L2 morphology and syntax may help to prevent fossilization of learners' partially correct interlanguage forms. Moreover, the use in a text of low-level vocabulary and grammar structures previously studied in class clears the path for the L2 reader to engage in language skills which would otherwise be beyond his reach.

In reading, fast accessing of words is essential if higher-order reading skills are to be given a chance to develop. Parsing, or meaning formulation, at first the phrasal level, then the sentence level, cannot take place until words and local grammar have been identified. For the interacting relationships amongst words in a meaning-unit to be apprehended, and overall understanding to be achieved, all the words within that unit must be held in working memory at the same time. The L2 reader who is struggling to access individual words or make sense of an unfamiliar grammar structure will almost certainly fall victim to Perfetti's *bottleneck effect* (1985; Perfetti and Lesgold, 1977), a situation whereby too much time spent identifying words further on in a sentence results in the forgetting of the first part of the sentence. (In L1 reading this effect is caused by poor decoding skills.)

The same phenomenon applies to the larger meaning-unit of the paragraph. One of the most marked differences between the skilled and the unskilled reader is the skilled reader's ability to make use of cohesive links in text (Oakhill and Garnham, 1988). The slow reader cannot do this, as the structure of previous text is forgotten in the time it takes to comprehend the

next piece of text. We have surely all found ourselves having to re-read a paragraph because we have read it too slowly in the first instance and structural clues from the first sentences have faded from working memory by the time we reach the last sentence, leaving us unable to link these shorter text units together and construct an overall meaning for the piece of text we have just read. Thus, whilst providing the L2 reader with simplified text may reduce his chances of encountering new input, it greatly increases his chances of being able to practise his parsing skills and develop his ability to make inter-sentence connections. Whether or not the practice *text* is authentic, the reading *process* which results is far more likely to be so. In fact, the laborious reconstructing of meaning from a text containing unfamiliar words and grammar which is likely to arise when the low-intermediate L2 reader tackles unsimplified reading materials can hardly be said to be *reading*. It is merely text-based language study.

A relatively unexplored question is how ease of comprehension, whether or not brought about by text simplification, impacts on *intake*. We have already seen that a high percentage of already known words may provide a favourable environment for guessing unknown vocabulary from context. This, although not the stated aim of the research, can be deduced from the L2 incidental vocabulary acquisition studies (Pitts *et al.*, 1989; Day *et al.*, 1991; Dupuy and Krashen, 1993; Horst *et al.*, 1998), since one cannot learn a previously unknown word just from reading without first having guessed its meaning. Liu and Nation (1985) have suggested, more precisely, that successful guessing can occur at a ratio of 24 known words to one unknown, or 96% already known vocabulary, although later work by Hu and Nation (2000), suggesting that 98% of vocabulary needs to be already known in order for satisfactory text *comprehension* to occur, might imply this to be a slight underestimate. We do not, however, as yet know if helping a student to comprehend a text by simplifying it also facilitates intake of the grammar structures which are present.

In a rare experiment of its kind, Leow (1993) investigated the differential effects of reading simplified and unsimplified versions of a text on the intake of the present perfect and present subjunctive amongst undergraduate students of Spanish as a foreign language. The two versions had equal numbers of instances of each linguistic item. Students were pre-tested and post-tested using a multiple-choice recognition task. A major weakness in the study is that there appears to have been three weeks of normal classroom instruction between the pre- and post-tests, so the experimental texts may not have been the only relevant input variable, although this would have been a random effect across students assigned to both texts. Leow

found that students who had read the simplified text significantly improved their intake of the two linguistic items, but students who had read the more difficult version did not. However, the students who had been assigned to the simplified version of the text were significantly less proficient than the other group at the outset, which may have affected results in some way, and Leow chose not to see the post-test results as proof that the simpler text had facilitated intake, instead, rather bizarrely, concluding that "simplification does not appear to have a facilitating effect on learners' intake of linguistic items in the input" (1993: 342). This conclusion does not seem justified given the statistically significant gains of those students who read the simplified version, and the absence of any significant change in the performance of the students who read the more difficult text. If the two groups were too disparate at the outset to be able to prove anything, then they must also be deemed too disparate to *dis*prove anything, and *no* conclusion ought to have been made. Clearly, more research is needed in this area.

Before concluding this section, it is worth noting that none of the studies above were concerned with the effects of text simplification in connection with *extensive* reading, and in all cases a strict limit was set as to the amount of text available for input. One of the benefits of extensive reading which results from the deployment of easily understood text is that greater *quantities* of text may be processed by the reader. Thus the student who has read two pages of text may encounter twice as many words and instances of grammar structures as the student who has read one page in the same amount of time. This potential benefit of using more quickly understood text is not taken account of in these text modification experiments. For example, Leow (1993) compares the impact of eight instances of the Spanish present perfect and present subjunctive in two environments; simplified text and unmodified text. If we assume, under normal circumstances, that the students who were given the simpler text could read their text more quickly, then in real-life conditions they might have had the opportunity to encounter the target grammar structures several *more* times in the same amount of time as the readers of the more difficult text took to achieve only eight encounters. Additional exposures might then result in better learning.

In the comprehension studies, we do not know if students reading simpler texts were not disadvantaged by the experimental design, since *speed* of comprehension was not included as a variable. Perhaps these students finished the reading task more quickly; perhaps they could have answered questions on two texts, instead of one, in the time allotted, but were not

given the opportunity to do so. This is, of course, merely conjecture, but there is a difference between a student who uses all the time available and one who does not, even if scores on a test are similar. Again, in real life, the student who comprehends text more quickly will be able to comprehend more of it within a given period of time, hence attain more input, than the reader who is slowed down by more difficult text.

It is true that simplifying text can result in a restriction of input (Yano *et al.*, 1994). In exchange, however, in the context of extensive reading, it affords easy access to much greater *quantities* of input, hence provides more samples of a particular structure and more exposures, within a greater variety of contexts, to uses of particular vocabulary items (*cf.* Nagy, 1997). Increased practice in decoding and accessing words may also lead to the activation of frequency effects, and help learners towards more automatic language recognition. In addition, the use of relatively easy text permits low-level students to achieve the reading speed necessary for the deployment of higher-level reading skills, from parsing, at the level of the sentence, to integrative, whole-text comprehension. This may be of particular benefit if the L2 is later to be used as a medium for studying. Finally, there may be a number of psychological benefits to the learner arising from the provision of texts which he can read easily and independently, without having constant recourse to his dictionary and grammar books. Such an achievement may have important motivational consequences which, in turn, may impact on the amount of intake resulting from the reading of the text in question and may extend, also, beyond that text to the learning of the L2 in general.

3. THE RESEARCH CONTEXT

3.1 Background to the Hong Kong Extensive Reading Scheme in English

The Hong Kong Extensive Reading Scheme in English was launched by the Hong Kong Ministry of Education in 1991, following on from a number of government-funded reports in the eighties, which had given rise to concern amongst educationalists as to falling standards of English in primary and secondary schools. At that time (although conditions have since changed), the majority of secondary schools in Hong Kong were, nominally at least, English-medium. In practice many of the English-medium schools were really what might be termed "mixed-medium", as code-switching between English and Cantonese was common practice. The most frequent method was the use of prescribed English text-books, discussed in the classroom partly in English and partly in Cantonese; Johnson and Lee (1987) found that English talk-time in the normal subject classroom ranged from an average of 37% in Secondary 1 to 53% in Secondary 3. A glance through a text-book belonging to a student would normally reveal pages of English text with the margins full of handwritten translations into Cantonese.

The Extensive Reading Scheme in English was only one of a raft of measures recommended by the Hong Kong Education Department to boost the level of English in the secondary schools. Other projects included an expatriate teachers scheme and short, intensive bridging courses in English for primary schoolchildren about to enter English-medium secondary schools. Its implementation was the responsibility of the Hong Kong Institute of Language in Education, who issued a contract to the Edinburgh Project on Extensive Reading (EPER), headed by David Hill, to provide the reading materials and to assist in the overseeing and development of the scheme, including the design and creation of evaluation instruments.

The scheme was to take in Secondaries 1, 2 and 3 and would operate along principles already established by EPER. Since a school's participation in the scheme would necessarily entail the fairly heavy costs of purchasing the prescribed reading materials in the first place, the Ministry of Education would bear these costs and schools would be admitted gradually to the scheme — always at the beginning of the school year — over a period of six years. Participation by schools in the scheme was to be purely voluntary, and at a school's request.

Any school granted admittance to the reading scheme would undertake to guarantee that the scheme would formally become a part of the English curriculum and would be allocated a prescribed minimum number of periods per week (two to three periods in Secondary 1, two in Secondary 2, and one to two in Secondary 3). These ERS classes would replace normal English lessons, so would not increase the overall amount of time spent on English by the reading-scheme students.

3.2 Materials and classroom mechanics of the Hong Kong Extensive Reading Scheme in English

One problem facing the organizers of a large-scale, multi-level EFL reading programme, intended to cater for students throughout three years of ever-increasing language proficiency and reading ability, is how to pace the students. A related problem is how to grade the reading materials into levels of difficulty. Since many students joining the HKERS would be expected to have very low levels of English at the outset, there would be a corresponding need for very low-level reading materials. However, it would be quite likely that, even within one class, there might be students at three or four different reading levels. This could be expected to be particularly the case in Hong Kong, where the amount of exposure to English outside of school varies enormously, even within the same broad social group. In families with a Filipino, rather than a Chinese, maid — and perhaps even one who may be entrusted with the greater part of the childcare duties — English may be spoken quite routinely at home. Moreover, the recent social history of Hong Kong could sometimes result in a family member being a native English speaker, or in a Hong Kong family having relatives in an English-speaking country. Other families, however, even of the same socio-economic standing, may have little or no routine contact with English. Because of this wide range of English abilities within one peer group, a self-access system, with students reading at their own level and progressing at their own pace, seemed to be particularly appropriate.

One aspect of EPER's work is to re-group the numerous commercially available graded readers, which are written and allocated to levels according to various publishers and editors' systems, into one standardised data-base of eight levels of reading difficulty. Thus, for example, at the EPER level of "C" — an intermediate reading level — there are Heinemann New Wave level 5 readers, Heinemann Guided Readers levels 4 and 5, Longman Structural Readers level 3 and Whitman English Readers level 1. The designation of a graded reader to any given EPER level is decided holistically, upon an overall appreciation of linguistic and

also extra-textual factors, such as difficulty of plot, flow of information and length and appearance of book, and is based upon David Hill's considerable knowledge and experience of graded readers and reading schemes (*cf.* Hill and Reid-Thomas 1988; Hill 1997, 2001). With few exceptions, the majority of graded readers are narratives. Sample pages from one such reader, used by the HKERS, can be found in Appendix 2.

Having a single grading scheme for all the graded readers made it easy to allocate reading materials to the reading scheme at EPER-specified levels, and to create a "levels ladder" for students to make their way up. In order to provide for maximum flexibility of possible reading levels within a class, it was decided that each class within the ERS should have access to the full range of eight EPER reading levels. So as to keep costs down, three classes — one Secondary 1 class, one Secondary 2 class and one Secondary 3 class — would form a "cohort" and share a complete reading library of 400 graded readers (50 at each of the eight reading levels). Thus, whilst it might be expected that the majority of the Secondary 1 students would probably be reading at a low level, and the majority of the Secondary 3 students might be reading higher up the "levels ladder", with the bulk of the Secondary 2 students somewhere in between, very high level Secondary 1 students and very low level Secondary 3 students would still be accommodated. The only problem would be a timetabling one, with the three constituent classes of any one cohort unable to have their ERS lessons at the same time. However, since the ERS lessons might be fitted into the total weekly allocation of English lessons whenever was suitable, each class would be able to choose which out of eight to ten lessons should be used on the ERS.

In practice, the cohort system worked very well, and was further improved upon a few years into the reading scheme, with the provision of one hundred extra graded readers for each complete eight-level library, at the middle reading levels. This was to cater for the reading "bulge" which became apparent at those middle levels, with students progressing fairly quickly through the lower levels, but slowing down at the intermediate levels. At the higher levels, fewer books were needed, partly because — with the high level books being so much longer — students did not need to exchange their books so frequently, and also because, quite simply, not many students actually reached the highest reading levels.

Although individual silent reading was to be undertaken during the ERS lesson, it was also expected that students would read *outside* class. During the ERS lesson, students would

choose a book from the class library and start reading it silently in class, taking it home at the end of the lesson. By the next ERS lesson, the student might be expected to have finished the book, and to be ready to exchange it for another one. Although one of the principles behind the practice of extensive reading is that the reading should be a low-risk activity, with few follow-up activities and no pressures of peer or teacher judgement, EPER's previous experience of extensive reading-schemes showed it to be nonetheless still necessary to have some kind of monitoring mechanism which would allow the teacher to check whether or not the student had actually read the book and whether or not he had understood it. Additionally, in order to evaluate both the students' progress, and the effectiveness of the reading scheme, it was thought desirable to keep track of the number of books each student read and at which reading levels.

In addition to the graded readers themselves, each title was accompanied by pre- and post-reading cards. The pre-reading card was designed to orientate the student simply and succintly to the subject matter — usually a story — and was attached inside the front cover of the book. The post-reading card, obtained by the student after he had read the book, had multiple-choice, gap-fill or short-answer questions and was intended to ascertain whether or not the student had achieved global comprehension of the story. There was no intensive language work and only minimal writing required, although, at the higher levels, students were encouraged to write a few sentences summarizing the main points of the story. The students would self-check their answers from a matching answer-card, kept in a separate box and obtained by the student upon application to that week's designated responsible student. (Sample pre-reading, post-reading and answer cards may be found in Appendix 3.) The students themselves recorded which books they had read and also wrote their answers to the post-reading questions in their reading notebooks. The role of the teacher was that of general overseer and facilitator, helping to resolve practical problems and making sure all the students were actively involved in the extensive reading class, rather than that of the traditional teacher, and a student might participate in the ERS lessons for several weeks without any individual dealings with the teacher. This was a role which Hong Kong teachers initially found quite hard to adopt, although the Institute of Language in Education ran teacher-training workshops for all new ERS teachers. The teacher would conduct occasional interviews with students to talk about the books they had read, but, in a 40-minute class, or even two consecutive 40-minute classes, would certainly not be able to devote five minutes to every individual in a class of 40-plus students, whilst at the same time overseeing the

mechanics of borrowing and returning books, post-reading cards and answer-cards. In fact, the practicalities of the ERS classes actually *compelled* the teachers to relinquish the more traditional type of teacher's role they were used to, and to adopt the role of overseer and facilitator, as constraints of time and class organization did not allow for traditional teaching methods.

Students were assigned to their initial reading level via their test scores on a proficiency test provided by EPER. This test consisted of a series of graded modified cloze tests. Students could move up to a higher reading level when they themselves felt able to cope with that higher level (with the teacher's consent) or when the teacher felt that they had read enough at one particular level and they ought to move up a level. It was also possible for a student to move *down* a level, if the teacher thought it appropriate.

3.3 Implementation of the Hong Kong Extensive Reading Scheme in English

The ERS started in 1991 in Secondary 1 only of twenty secondary schools. Efforts were made, at this point, to include schools from both the top-scoring and bottom-scoring schools in the unofficial school banding system, as well as schools from the middle. (The Hong Kong school banding system is an unwritten ranking of schools, used only within the Ministry itself.) In practice, however, the lower-band schools did not, on the whole, apply to the scheme, perhaps because they did not wish to lay themselves open to possible official scrutiny.

In year two of the scheme, the original Secondary 1 students continued into Secondary 2, taking the ERS with them, and the new Secondary 1 classes below them also started the reading scheme. An ERS school would now be running the reading scheme in Secondaries 1 and 2. In year three of the scheme, Secondary 2 moved up to become Secondary 3, still continuing with the reading scheme, Secondary 1 moved up to become the new Secondary 2 and an incoming Secondary 1 would now also start the ERS. By 1994, when the large-scale evaluation was carried out — and three years into the ERS — the reading scheme was running in Secondaries 1, 2 and 3 of the original twenty schools, in Secondaries 1 and 2 of the second wave of schools which had joined the ERS in 1992, and in Secondary 1 of the third wave of schools which had joined the ERS in 1993. Schools continued to join the

scheme for a number of years after this, until, by 2000, approximately 200 of the 400 government-funded secondary schools in Hong Kong were using the scheme.

4. METHODOLOGY

4.1 Background to the data

The data which is used in the present study is part of a very much larger data set which was collected under the auspices of the Hong Kong Institute of Language in Education (now subsumed, in part, by the Hong Kong Institute of Education) as part of an initiative to evaluate the overall impact of the Hong Kong Extensive Reading Scheme in participant schools.

Although the so-called *rolling* system of schools entering the ERS was intended primarily to help spread the financial costs for the Ministry of Education, and to ease schools gradually into the reading scheme (*cf.* Chapter 3), it also had the side-effect of creating, as it went along, "natural" control and experimental peer groups, progressing through the same secondary schools with a one-year gap between them. At the end of a school's first year in the ERS, the performance (*e.g.* in English language assessment-related procedures) of students finishing Secondary 1 — the school's first Secondary 1 to have been using the ERS — might be compared to that of the equivalent students of one year earlier, who had not had the experience of the ERS — at the end of *their* year in Secondary 1. At the end of a school's second year in the ERS, the performance of Secondary 2 students, who had by this time spent *two* years with the ERS, might be compared against that of the non-ERS Secondary 2 students at the end of the previous school year. The end of a school's third year within the ERS would allow comparison between the respective performances of students who had spent three years within the ERS and equivalent students from the previous year who had not used the ERS. As both groups of students — classes using the reading scheme and their equivalent classes of the year before who did not use the scheme — had the same syllabus, the same English coursebooks, the same total number of English-lesson class hours (ERS lessons were to be fitted into the existing number of hours given to English classes), the same teachers, and indeed the same overall school environment, and came from the same socio-economic and geographic background, a potentially very rich research environment was created, affording investigation, through same-stage comparison of peer groups, of the impact of the ERS on students, and also on teachers and the school as a whole, after one year, two years and three years of participation in the Scheme.

Data, in the form of student test scores and completed student and teacher questionnaires, was collected from 19 schools using the ERS as part of the school curriculum. Eighty-four intact classes, drawn from Secondaries 1, 2 and 3, provided control (non reading-scheme) and experimental (reading-scheme) students. With an average of 40 students per class, this yielded data from approximately 3,360 students. The data was collected at the end of two consecutive school years, such that the data collected at the end of the first year came from the control groups, and the data collected at the end of the following year came from the experimental groups.

Pre-test scores for control classes were unavailable, since schools could only be asked to participate in the research project *after* they were admitted to the ERS, at which point students of the previous year became the control cohort. Schools participating in the data collection undertook to provide control and experimental data from matching classes, which is to say that, in a school where students were streamed into different ability groups, if the highest class, for example, was used to provide any particular set of control data the first year, then the equivalent highest class of the following year would then be used to provide the subsequent matching set of experimental data. Students from, for example, consecutive highest ability classes (*i.e.* the highest ability class in any particular Form in two consecutive school years) might only differ significantly in ability if the whole student intake of approximately 200 students differed significantly in ability from one year to the next. Although this possibility cannot categorically be ruled out, the rigidity of the Hong Kong schools' banding system and of individual school intake procedures made it highly unlikely. In practice, the intake of any Hong Kong secondary school is homogeneous from one year to the next.

Tests administered were the Standardized Reading Test of the Hong Kong Reading Association (HKRA, *n.d.*), which was taken by Secondary 1 students, and two tests designed by EPER specifically to accompany the HKERS: a test of extensive reading (Davies and Irvine, 1992a) and a vocabulary recognition test (Davies and Irvine, 1992b), which were taken by Secondaries 2 and 3. Secondary 3 students and one pair of experimentally matched Secondary 2 classes also wrote a timed composition. Teachers and students completed questionnaires on students' reading habits and attitudes towards reading in English and towards English in general, and on students' and teachers' opinions as to the value and effectiveness of the reading scheme as part of the school curriculum. A thesis-length

discussion of this large-scale study of the HKERS can be found in Yu (2000), of which a brief account has been given in Chapter 2 of this present thesis.

4.2 The Present Study

The data investigated in the present study consists of 392 narrative compositions written by Secondary 2 and Secondary 3 students from four schools using the ERS. These compositions were included in the large-scale study outlined above and have been previously evaluated as part of that study (Yu, 2000). In that research, two raters each provided a score ranging from 1 to 5 for the three constructs of *content*, *narrative structure* and *language and style*. This procedure did not, however, produce any clear results (*cf.* Section 2.2), and these three constructs may not have provided the best framework for evaluating these particular writing samples. The present study proposes a different approach, combining rater assessment with text-analysis techniques, with a stronger bias towards *linguistic* features, which it is hoped may provide better insight into any differences in the writing performance of the two groups.

In the present study, the issues of content and narrative structure will not be addressed. Firstly, within the context of a narrative writing task, it is not clear how *content* should be judged. Unlike expository writing, which may have a very controlled environment, unless there is a complete set of story prompts, such as a sequence of pictures, narrative content is at the discretion of the author. It would be hard to set any kind of objective benchmarks for plot. Secondly, in a genre which does not have any particular set of syntactic, lexical or rhetorical markers which might exert a certain influence on content (such as formal use of the passive in abstracts, or the conventions of writing a business letter), performance on *content* cannot be held necessarily to have a direct relationship to linguistic competence. Such constructs as are routinely used in the assessment of *creative* writing — such as *imagination* and *story development* — are not language-specific, and even a native speaker may perform poorly on these. In any case, in the present data, it is necessary to distinguish between what may be the effects of reading stories, in any language, and what may be the effects, more specifically, of reading in an L2.

The same might be said of *narrative structure*, with deviations from any kind of expected norm very difficult to connect categorically to varying levels of L2 competence. One of the most common and practical criteria used in the evaluation of *narrative structure* is whether a

story has a beginning, a middle and an end. Although this *may* be valid for assessing creative writing, it is unclear how it might be connected to language competence. Indeed, when assessing the writing of non-native English speakers with possible differing literary and rhetorical traditions, unless this has been a specified part of a course now being assessed, its use as a criterion raises many important socio-cultural and ethical questions.

In the case of the compositions used in this study, the writing task was to tell a story, in the first person, about something which happened to the student-story-teller in the past. The most probable narrative structure to obtain from this writing task was a simple first person time-sequenced past narrative, and, at the level of Secondaries 2 and 3 students, aged 12 to 13, there was not likely to be much departure from this format. (This proved, in fact, to be the case.) With such little variety of narrative approach, it is not clear how to differentiate amongst scripts. In any case, narrative time sequencing is not language-specific, and may derive from knowledge of narrative structure in an L1 quite as well as from competence in an L2. Neither is coherent (or incoherent) storytelling language-specific and this may, in fact, derive from many factors. Some structural incoherence may, of course, be the result of poor language skills (for example, the misuse of connectors or pronouns, or poor command of verb tenses or of syntax in general), and it is important to try to extract these from the equation.

The research interest of the present study is whether extensive reading in an L2 may be seen, through a close evaluation of the language used in an open production task, in which language performance is constrained by a time limit and by the genre elicited by the task, but otherwise unguided, to effect measurable improvement in *language* knowledge and use. In other words, does extensive reading contribute to increased proficiency in an L2 in morphological, syntactic and lexical domains?

Specifically, this study aims:

- to investigate, using a different assessment approach from that used in previous research on the same data, whether extensive reading may lead to rater-noticeable and subjectively assessable changes in language production as elicited through a narrative writing task

- to investigate, through an analysis of countable surface text-features of the type commonly used by researchers to attempt distinction between proficiency and performance levels, whether extensive reading may lead to measurable changes in surface text-features of language production as elicited through a narrative writing task

- to investigate the relationship between objectively-measured surface text-features and rater-assessments within this context.

4.3 Methodology

4.3.1 Schools and students

Four secondary schools were involved in the present study; three high-ability schools, and one low-ability school.[1] These rankings are those of the Hong Kong Ministry of Education's unofficial school banding system (*cf.* Section 3.3). All four schools were officially English-medium, although in practice much Chinese was spoken and written in the English-medium classroom in Hong Kong at that time, particularly in lower level schools, and, within schools, particularly in Secondaries 1 and 2 (*cf.* Section 3.1). The degree of English/Chinese mix used within the classroom was not controlled for during the initial data collection, other than by the fact that matching control and experimental classes came from within the same school and were likely to have experienced similar ratios of Chinese to English as the language of instruction and classroom interaction.

All four schools streamed their intake into Secondary 1, creating a range of higher to lower ability classes within the same year. Individual students were streamed by means which were at the discretion of each individual school, there being no standard, Ministry-endorsed, set of guidelines and criteria for dividing students into lower and higher ability groups. Classes, once created, remained constant across all subjects, with no re-setting of ability groups for specific subjects. Individual students might be promoted or demoted between the end of one school year and the beginning of the next, but, on the whole, classes remained stable throughout Secondaries 1, 2 and 3. In any case, all the students of a given year either

[1] *The results from Yu (2000), reported in Section 2.2, derived from compositions collected from five schools. Compositions from four of these schools are used in the present study. The fifth school, a medium-ability school, differs from the other four in that it was the only school which did not stream, but had mixed-ability classes.*

did or did not use the Scheme, so a student's movements between higher and lower classes whilst progressing through Secondaries 1, 2 and 3 would not affect his number of years of exposure to the reading scheme.

The low-ability school provided data from two control and two experimental classes. One high- and one very high-ability school each provided data from one control and one experimental Secondary 3 class, and a second high-ability school provided data from one control and one experimental Secondary 2 class, giving a total of five control and five experimental classes (Table 4.1).

Table 4.1 Participant schools and classes

School	Ministry Band	Year (Secondary)	Number of control (non ERS) students	Number of experimental (ERS) students	Number of years in ERS (experimental students)
1	Very High	3	33	43	3
2	High	3	41	41	3
3	High	2	42	42	2
4	Low	3	40 (higher level class)	43 (higher level class)	3
			34 (lower level class)	33 (lower level class)	3

The data was collected over two consecutive school years. Each school undertook to match control and experimental classes, as explained in Section 4.1. The total number of students was 392, of which 190 were control students and 202 were experimental students. Of the experimental students, 160 had spent three years in the extensive reading scheme and 42 had spent two years in the scheme.

4.3.2 The writing task

Students were required to write a timed narrative composition under standard examination conditions, that is in class time, under teacher supervision, with no help from dictionaries or guidance from the teacher. A narrative writing task was set since this might best capture any changes in the writing of the experimental students which could be hypothesized to be a result of their participation in the ERS. Thus the genre of the elicited (writing) output would

match the genre of the experimental (reading) input. (Over 95% of the graded readers used in the reading scheme were of the narrative genre, as described in Chapter 3.) Moreover, an open writing task would not place false constraints on students' performance, allowing the weakest and the strongest alike to produce text at their own level.

Students from the highest ability school and students from the lowest ability school were instructed to choose one of the following writing topics:

You and a friend got lost and had to spend a night at an empty house people said was full of ghosts. Describe what happened to you.

Imagine you met a magician who offered to make three wishes come true for you on one special condition. Say what the condition was and what happened when your wishes came true. Did they turn out as you planned?

Students from the remaining two high ability schools were asked to write on a third topic:

The most exciting day of my life. (It does not have to be a true story. Use your imagination to make it as exciting as possible.)

Whilst it is recognized that writing prompts may exert some influence not only on the linguistic characteristics of writing samples, but also on students' performance levels, such effects are more likely to be caused by differences between task-types, and genres, than to be caused by topic within a genre. For example, Brown, Hilgers and Marsella (1991) found a significant effect (in task difficulty) for text purpose, but no effect for topic. In the case of the three writing prompts above, it was not hypothesized that any one represented a more difficult task than the others. Although these might elicit different vocabulary, there was only one possible genre and levels of personal involvement with the task were likely to be similar across the three rubrics. In any case, matching pairs of control and experimental classes were given the same task.

Secondary 3 students from two high-ability schools were instructed to write 250-300 words. Secondary 2 students from one high-ability school and Secondary 3 students from the low-ability school were instructed to write 200-250 words. The time allocation for the writing task was 70 minutes.

4.3.3 Evaluation methods

Evaluation of the compositions falls into two distinct parts. Firstly, all 392 compositions were evaluated subjectively by three teacher-raters, for overall quality, as a primary, holistic, measure, and, since "proficiency does not necessarily develop at an equal rate in all its components" (Young, 1995: 16), for a number of language-proficiency constructs. Secondly, a subset of 150 compositions — 74 control and 76 experimental — were assessed on a range of objectively quantifiable surface text-features.

4.3.4 Evaluation by raters

4.3.4.1 Evaluation for overall quality

The first stage in the evaluation process was the assignment of a single holistic score, to be decided rapidly, and independently, by three raters. The 392 control and experimental composition scripts were each assigned a random ID number. The compositions were then typed up to eliminate the variable of handwriting, which may be distracting or difficult to read, and which was not of interest to the current research. The scripts were delivered to the raters in identical format, each script with its unique ID number, to enable initial rater discussions and permit the clerical work of recording scores, but with no information as to the school, class, gender or control/experimental status of the writer. All errors of spelling and punctuation and the students' own paragraphing were retained.

4.3.4.2 The rating instrument

There are many potential problems associated with rating scales, and a poor choice of rating scale can invalidate the most carefully planned piece of research. Rating scales are generally at their most valid when used for gate-keeping exercises, when desirable and/or necessary skills and sub-skills have been pre-identified, or for achievement testing after a specific, well-defined teaching programme, where the descriptors will correspond to what has supposedly been taught by the course. Rating scales which are claimed to differentiate between language proficiency levels in general, independently of any more specific purpose, should derive from some theoretical model of language development. In practice, many rating scales are syllabus-led, with questionable assumptions about the order, and the linearity, of second language development, and little or no account taken of the fact that the

traits within a construct may develop at different rates, or that constructs may not co-develop, but might even compete with each other, such that progress in one reduces, or is even positively detrimental to, progress in another.

Forcing a set of data to fit into a possibly inappropriate rating scale can be counter-productive, since findings are partly pre-supposed before they have been found, or, at the very least, constrained to fit into a particular framework. If, for example, a rating scale is being used to measure the effects of a treatment such as a new teaching method, sets of descriptors, theoretically representing different proficiency levels, must surely reflect the researcher's expectations of the effects of the new method. If descriptors are not appropriate to the language development which *actually* occurs, these will merely produce noise, and features of language development which *do* occur, but which were not anticipated when the rating scale was devised, may pass unrecorded.

When the effects of a particular treatment are as yet unknown, it becomes extremely difficult to establish descriptors. We do not conclusively know what prompts learning in an L2. It is probable that different features of language are more, or less, susceptible to different learning practices (and that these may also vary with the individual). Because of time constraints, and because of the traditional role of the teacher as a central learning coordinator, most classroom teaching is form-focused, with students directed into form-focused methods of processing controlled amounts of input. This is certainly the case in the Hong Kong secondary school classroom. Extensive reading, on the other hand, strictly *dis*courages focus on form, but encourages reading for meaning only, involving large amounts of quickly-processed text. Thus the ways in which text is processed by the learner-reader during extensive reading may be quite different from the ways in which text is normally processed in the classroom. Since non-form-focused learning is associated with *implicit* learning, it might be a justifiable assumption that extensive reading may favour development of constructs, or traits within constructs, which are more susceptible to implicit learning. It is even possible that some language skills may develop at the expense of others, either because of a reduced need to develop these, or because cognitive resources are being diverted into less usual tasks. Some skills might simply be less favoured by this different type of learning environment. A rating scale which has been devised to capture a more usual pattern of classroom-induced language development may not accommodate any such different learning pattern, and might even obscure it.

If there is no appropriate rating scale available to fit one's data into, it is of course possible to create a custom-made scale. This should not be done, however, if there is no sound language development theory which fits exactly the case in question. In the absence of such theory, another solution is to construct an empirically derived scale, using one part of the data to establish sets of descriptors which may then be used to evaluate the whole set of data. If this is well done, the framework constraining raters' evaluations may be a more appropriate one. This method does, however, retain the disadvantage common to all scales which use sets of descriptors, which is that descriptors — though providing a very helpful environment for maintaining inter-rater, and intra-rater, reliability — place an additional burden on the rater during the rating process. Lumley even contends that there is little evidence that raters actually *use* the descriptors to make decisions, but rather that "the task raters face is to reconcile their impression of the text, the specific features of the text, and the wordings of the rating scale" (2002: 246).

The approach adopted in the current research was to use a simple norm-referencing technique, which requires raters to evaluate a group of scripts wholly in relation to all the other scripts in the group, holistically, and with no reference to any external set of criteria. This eliminates the possible contaminating effects of poorly selected criteria and inappropriate descriptors, on raters' judgements in the first instance, and, ultimately, on the research outcome. If descriptors modify the raters' task in the way that Lumley claims, then, irrespective of any potential set of criteria, the activity of rating, in itself, might also be more valid.

One other advantage of this technique is that no empty bands are created at the extreme ends of the scale, where unrealistic criteria might exclude all but a few abnormal outliers. In other words, the data and the rating scale fit each other perfectly. With reliable raters, if a construct is valid, and if one places the average score in the middle of the score range, data should produce a normal distribution, permitting the use of a range of powerful parametric statistical tests. The scale below was used, following Hamp-Lyons (1991):

6 High/Excellent
5 Good
4 High Average
3 Low Average
2 Weak
1 Low/Very Weak

Raters were further given the instructions that:

The holistic mark is an impressionistic mark. Compositions should be assessed *rapidly*, and in relation to the others in the group.

Although Pollitt (1991) argues that a six-point scale may be overly optimistic for reliable distinction between scale points on a writing test, and advises that no more than four scale points is optimal, there was strong justification for a six-point scale (irrespective of the fact that many successful commercial tests use six or more scale points; *e.g.* TWE). Both inter-rater and intra-rater reliability are critical to the success and the validity of any rating procedure, and these are conventionally controlled for via correlational procedures. A range of three possible scores is unusable for correlational procedures. To use Pearson's product-moment correlation with a range of four possible scores is hardly better, since the direction (positive or negative) of this correlation depends on how many pairs of scores are on opposite sides of the mean. A range of four scores, awarded by two judges, could be expected to lead to many 2-3 and 3-2 pairings — since the scores 2 and 3 are adjacent and, in this situation, represent 50% of the possible scores — resulting in a disproportionate number of paired scores on opposite sides of the mean. The many negative *products* arising from this (*i.e.* the z-scores, or standardized distances from the mean, of the paired scores multiplied one by the other; a negative multiplied by a positive always gives a negative) would erode the final *moment*, or magnitude of the correlation. Moreover, a range of four scores may not provide a reliable correlational magnitude even without this problem. The fewer the scores available for selection, the greater the risk of an inflated correlation, either positive or negative. If correlations are going to be used, a minimum of five points on a scale is highly advisable. More is even better, though, as Pollitt rightly points out, it then becomes more difficult for the *rater* to behave reliably.

One other reason for preferring a six-point scale to a five-point scale in the present context was to help prevent raters tending too heavily towards the central score on a set of options, with a reluctance to classify anything other than clear-cut cases as being above or below. *Two* average scores — *high average* and *low average* — would force raters into a decision, with no choice other than to place a script either in the top half or the bottom half of the range of scores. This might spread the data more and help to reveal any differences.

4.3.4.3 The first rating process

The three raters were all current EFL teachers with a minimum of four years' teaching experience in the UK and abroad. Rating was conducted in controlled rating sessions in an unused classroom in a language teaching institute. Each rater had a complete set of typed scripts, ordered numerically using the ID numbers previously assigned randomly to each script. The re-ordering of the random ID numbers into base numerical order ensured the random distribution of control and experimental scripts, and of scripts from different schools and classes, within the set of scripts first received by the raters.

Working with the first 40 scripts from the set, each rater independently chose the ones she considered to be the strongest and the weakest. The group then discussed their choices, and agreement was reached as to which two compositions would represent the top and bottom end-points of the range of scores. Raters then individually judged the first ten compositions, in relation to each other and to the end-point compositions, deciding where to place these ten scripts within the range of six grade scores. The group then discussed these, eventual agreement resulting in sample scripts for each of the grades. This process was repeated with small numbers of scripts, until a high level of inter-rater consistency was achieved, and benchmark scripts for each grade identified.

With preliminary inter-rater consistency established, each rater then individually rated 40 more scripts. Following a break, each rater rated the same scripts again to check intra-rater consistency. This process was repeated until a high level of intra-rater consistency was achieved.

Retaining six benchmark scripts — scripts which all three raters had consistently placed at the same grade — as a guide for each of the six bands, the remaining 386 scripts were reshuffled, and each rater then graded the complete set of scripts in her own unique random order, at an approximate rate of two to three minutes per script, depending on length. Raters made no marks on the actual scripts. At the end of each session, four or five randomly selected rated scripts from each rater were re-inserted into that rater's pile of scripts remaining for the next session, to be rated again by the same rater. In this way intra-rater reliability was monitored continually as the rating process went along.

4.3.4.4 Reliability of the first rating process

All three raters produced normally distributed sets of scores, permitting the use of parametric statistical procedures to ascertain rater reliability. Inter-rater reliability for the set of 392 scripts was .853 (significant at $p < .001$), using a Pearson's correlation with a Spearman-Brown Prophecy Formula adjustment for three raters (*cf.* Henning, 1987). Comparison of the respective means and distributions of the three raters' sets of scores (ANOVA), which may be used to ascertain whether any particular rater — though still correlating with other raters — is stricter or more lenient than other raters (Alderson, Clapham and Wall, 1995), revealed no significant differences between the raters. Means and standard deviations are given below (Table 4.2).

Table 4.2 Means and standard deviations for overall quality ratings

	Mean	s.d.
rater 1	3.51	1.15
rater 2	3.42	1.12
rater 3	3.41	1.30

N = 392
Score range = 1-6

Intra-rater reliability was not susceptible to correlation, with the number of twice-marked scripts too small to provide a reliable correlation coefficient. Percentage agreements, which is another method of assessing inter-rater reliability (Cushing Weigle, 2002), were used as a measurement of *intra*-rater reliability. With "agreement" defined as same score or adjacent score, which is to say that paired scores of more than one grade point apart "disagree" (White, 1984), rater 1 had 95.4% self-agreement (N = 22), rater 2 had 91.66% self-agreement (N=24) and rater 3 had 95.8% self-agreement (N=24). According to White, 90% agreement (between two raters) is average, whilst 95% agreement is excellent.

4.3.4.5 The second rating process: within band rating

When the first stage of assessment for overall quality was completed, a second, refining, stage was undertaken. Within each grade band, raters allocated a secondary score, from 1 (lowest) to 4 (highest), to each script within that band. As in the first stage of rating, no criteria or set of descriptors were used for this, scripts being judged holistically, for overall quality, in relation to the other scripts which had also been placed (by the rater now doing

the secondary evaluation) in that band. As each of the six original bands was thus divided into four "sub-bands", this increased the spread of possible scores to 24.

Inter-rater consistency for the four bands within each original grade band was monitored at the outset in the same way as inter-rater consistency for the original six grade bands was monitored (described above), but using for initial trial and discussion only those scripts which all three raters had previously placed in the same band. Consistent inter-rater agreement, however, was found to be impossible to achieve at this fine level of distinction — and insistence on agreement where none can be reached, even after discussion, if it is not an indication of the unreliability of one or more of the raters, may in fact be a threat to the validity of rater judgements.

A new range of overall quality scores, from 1 to 24, was calculated by stretching the original range of six scores to accommodate the addition of the within-band scores. For example, a script graded at 1, and with a within-band score of 1, would remain, in the new range of scores, at the lowest score of 1. However, a script graded at 1, but with a within-band rating of 2, would have a new score of 2. A script graded at 2 with a within-band rating of 1, would have a new score of 5, and so on.

Inter-rater reliability for the set of 392 scripts with the new range of scores was found to be .90 (significant at $p < .001$), using a Pearson's correlation with a Spearman-Brown Prophecy Formula adjustment.

4.3.4.6 Evaluation of separate constructs

In her survey review of research methodology in second language writing research, Polio looks at 50 studies of EFL learners' writing, most of these published in major journals. From these, she identifies nine categories of features of L2 writing which are commonly investigated by researchers: "overall quality, linguistic accuracy, syntactic complexity, lexical features, content, mechanics, coherence and discourse features, fluency and revision" (Polio, 2001: 92).

Polio's study offers a very useful reference taxonomy. In the present research, the issue of overall quality is addressed by the procedures described above. The question of content is not considered, *a priori*, to be a valid question in an examination of linguistic features of narrative text (*cf.* Section 4.2). The study of revision processes is not feasible with the data.

This leaves the six categories of accuracy, complexity, lexical features, mechanics, coherence and discourse features, and fluency to be considered.

Whilst most researchers would agree that increased grammatical complexity co-occurs with language development, this is not necessarily the case for grammatical accuracy. Many researchers argue that levels of grammatical accuracy are not stable predictors of language development. It is true that reduction of error *may* co-occur with development, but increase in error has also been shown to co-occur with the ongoing acquisition of more advanced structures, or been seen simply as evidence of any kind of restructuring. Both accuracy and complexity are of considerable interest in the present research, as being possibly susceptible to an effect of extensive reading.

Lexical features also afforded a very interesting field of investigation. The potential of extensive reading for aiding L2 vocabulary development is where it receives most support from the literature, specifically in the notion of *incidental* vocabulary learning (*e.g.* Day and Bamford, 1998; Nation and Waring, 1997; Coady, 1997), though the experimental evidence for this is, in fact, very slight, as we saw in Chapter 2. For this research "lexical features" was re-framed, for the purposes of rater judgements, as *vocabulary range*, which is a more precise construct and therefore more likely to attract reliable ratings.

A separate category for spelling was created for two reasons — to isolate spelling as a variable in its own right, so that it would not impinge upon vocabulary range or grammatical accuracy judgements, and because spelling has been claimed to benefit from reading (Krashen, 1989), although it has been, in fact, little explored by researchers. Indeed, in rating scales, spelling is commonly grouped with punctuation, under the heading "mechanics". As Polio (2001) points out, there is no theoretical motivation for grouping spelling and punctuation together, and consequently there is no sound reason to suppose that as one improves, so does the other.

"Coherence and discourse features" gathers together under one umbrella a wide range of disparate, part-lexical, part-syntactic language devices and features. In research, and also in teaching, isolation of and focus on discourse features is highly specific to language task. It is not a common practice within research, even when evaluating academic or expository writing, for which specific types of discourse features may be highly important, to judge these holistically, as a group. Instead, performance on specified discourse features is most

commonly judged by counting instances — although counting instances does not inform as to misuse.

In narrative writing there is a less clearly tabulated set of associated discourse features than in many other genres, all the more so at lower levels where lexis and grammar have not yet been mastered, and there is less clarity of definition as to what may be considered a "discourse feature" and what is merely grammar and/or lexis. It might be argued, for instance, that connectors in general, conjunctions in particular, use of the past tenses, or use of the first person pronoun are all discourse features of narrative. However, these are also features of the basic grammar. For these reasons, "discourse features" was not considered an appropriate evaluation category for the low-level narrative writing in this study.

Coherence, on the other hand — as a single, if highly complex, construct, rather than as the sum total of the number of instances of cohesion devices — is commonly evaluated holistically, and is an appropriate category of evaluation for narrative writing. Fluency is also appropriate, although this is more commonly cited in the evaluation of oral production. In written work, fluency, which might be characterized as "ease of production", cannot be directly accessed, but must be assumed from the outcome of the production. It is often assumed, for example, that, all else being equal, longer written texts indicate a greater writing fluency than shorter written texts produced within the same time limits. Ease of reading is also used as a measure for L2 writing fluency, although, in fact, we do not know if what is easy to read was also easy to produce. It is possible that text which is easy to read comes from a greater degree of affinity with the L2, however impossible it may be to deconstruct, far less operationalize, such "affinity". Such "fluency" may be quite difficult to separate from "coherence", which also manifests itself in ease of reading. In the present research, coherence and fluency, for the purposes of holistic rater evaluation, were merged into one category of *coherence and flow*.

In total, six assessment categories were decided on and, in a third rating procedure, the same three raters who performed the *overall quality* evaluations also assessed the 392 scripts for:

grammatical complexity
grammatical accuracy
vocabulary range
spelling
punctuation and paragraphing
coherence and flow

4.3.4.7 The third rating process: rating of separate constructs

The compositions were assessed on the six constructs detailed above, using the same basic approach as for the initial overall quality ratings, with some slight modification. Scripts were rated for each construct, holistically and rapidly, on a six-point scale, from Low/Very Weak to Very High/Excellent, in relation to all the other scripts. As before, sets of level descriptors were not used.

Constructs other than spelling and punctuation — which were considered self-explanatory — were each elaborated in a short set of guidelines. As two of the raters were EFL teachers, not applied linguists, the construct of *coherence and flow* was re-written as "story reads well", with the intention of providing a less technical rubric. The rubrics and guidelines were discussed by the raters in the first closed rating session, to establish a common understanding of what was, and what was not, to be taken into consideration when making judgements. In particular, spelling was not to be taken into consideration when making judgements on *vocabulary range* or *grammatical accuracy*. Grammatical accuracy was not to be taken into account when making judgements on *coherence and flow*, unless coherence was severely disrupted by lack of accuracy. The actual events in the story were not to be taken into account when making judgements on *coherence and flow* (which had the rubric "story reads well"), but, rather, such ease of reading as came from *language* use in the story telling. (The evaluation instrument and rater instructions are to be found in Appendix 6.)

The same procedure was followed as before, with initial grade-level benchmark scripts for each construct being established through individual decision-making and inter-rater discussion, as described above. This third procedure differed from the overall quality assessments, however, in one major respect, which is that, once the rating proper started, raters made judgements on the six constructs in the course of a single reading. Separate readings for each of the six constructs would have involved a total of eight readings (including the two readings for the overall quality judgements) for each script, which might have induced a type of script-fatigue, with readers beginning to recognise scripts previously rated and no longer processing them with a fresh eye or applying the same degree of attention. Possible effects of making six judgements on the single reading, and in particular the possible effects of fixing the order in which these judgements were made, with judgements on one construct perhaps influencing judgements on other constructs, are discussed in Section 4.3.5. Within-band rating was not carried out for the six constructs. As

this was a method of refining judgements within a single band, it would have been necessary to re-group the scripts into their six bands in order for comparisons to be made against all the other scripts in that band. This would have had to be done separately for each of the six constructs, entailing a further six readings for each composition.

4.3.4.8 Reliability of the rating process for separate constructs

Of the six constructs, only five produced normally distributed sets of scores. Inter-rater reliabilities for these, in the form of a Pearson's correlation coefficient with a Spearman-Brown Prophecy Formula adjustment for three raters, are given below (Table 4.3).

Table 4.3 Inter-rater reliability for separate constructs

grammatical complexity	.772
grammatical accuracy	.818
vocabulary range	.806
punctuation and paragraphing	.585
coherence and flow	.757

N =392
All correlations are significant at p < .001

Spelling scores did not produce a normal curve, but a J-shaped distribution for all three raters. (Possible reasons for this are discussed in Chapter 5.) For this reason, a non-parametric correlation, Spearman's rho, was used, giving an inter-rater correlation, using a Spearman-Brown adjustment for three raters, of .738, significant at p < .001.

Means and standard deviations for the three raters' sets of scores are given below (Table 4.4).

Table 4.4 Means and standard deviations for separate constructs

	grammatical complexity		grammatical accuracy		vocabulary range		punctuation & paragraphing		coherence and flow		spelling*
	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Mean	sd	Median
rater 1	3.07	1.29	3.04	1.17	3.95	1.07	2.67	1.24	3.50	1.37	5
rater 2	3.35	0.95	3.07	1.19	3.55	0.96	3.35	0.96	3.54	1.09	5
rater 3	2.85	1.21	2.58	1.27	2.95	1.27	3.06	1.44	3.06	1.25	5

N = 392
Score range = 1-6
** Spelling scores did not produce a normal distribution. Medians are reported. There is no appropriate measure of dispersion.*

Comparisons of the respective means and distributions of the three raters' sets of scores (ANOVA) revealed significant differences ($p < .05$) between the raters for five of the constructs. (*Spelling* scores were not included in the ANOVA.) Possible reasons for these differences between the raters are discussed below (Section 4.3.5).

4.3.4.9 Rater debriefing sessions

italics?

After each of the three rating procedures (overall quality, within-band refinement and separate constructs) a rater debriefing session was held. Raters made observations about their own rating behaviour, discussing factors they felt might have influenced their judgements and any particular problems they encountered during the rating processes. These discussions were recorded. They have not, however, been transcribed and are not a main issue, as, apart from time and space restrictions, the focal interest for this research is not the processes of rater decision-making. Rather, these recordings provide a useful source of additional information which may be referred to in later discussion on the results of the rating procedures, if relevant.

4.3.5 Limitations of the subjective evaluation methodology

As noted above (Section 4.3.4.8), rating for separate constructs did not produce such homogeneous results from the three raters as the first two rating procedures (ratings for *overall quality*). An initial ANOVA revealed significant differences amongst the raters for all five constructs. (*Spelling* was not included, as it did not have a normal distribution.) Subsequent paired-samples t-tests revealed that rater 3 differed from raters 1 and 2 in *grammatical accuracy* and *coherence and flow*. There were no significant differences between raters 1 and 2 on these. A t-test reveals differences between distributions using the mean and the variance. From the figures in Table 4.4 it seems that the differences between rater 3 and the other two raters on *grammatical accuracy* and *coherence and flow* were caused by rater 3's being more strict.

Paired-samples t-tests for scores on *grammatical complexity*, *vocabulary range* and *punctuation and paragraphing* revealed differences on *each* pairing of raters (*i.e.* between raters 1 and 2, between raters 1 and 3 and between raters 2 and 3). Again, it seems likely that these differences may be largely accounted for by differences in strictness. Adding to

the differences in means, however, rater 2 also had noticeably lower standard deviations on *grammatical complexity*, *vocabulary range* and *punctuation and paragraphing*, indicating a narrower spread of scores on these than the other two raters. Such differences in distribution patterns may also contribute to the finding of a significant result in a paired-samples t-test.

No such differences in strictness or distribution patterns between the raters emerged during the rating for *overall quality*. That such differences emerged from the rating of different constructs may be for theoretical reasons, practical reasons, an effect of individual raters' backgrounds or an effect of the constructs evaluation instrument.

It is possible, for example, that rater 2's lower standard deviations arose from increased difficulty in differentiating levels of performance for separate constructs. This might be a reflection of the fact that forcing an evaluation on such constructs is less valid than an overall holistic evaluation. It may also be that this particular rater was unused to considering texts in terms of these constructs. Interestingly, rater 2's only standard deviation which did not diminish was that for *grammatical accuracy*. It is possible that rater 2 was more accustomed to evaluating texts on grounds of accuracy — a common focus for classroom teachers, who are expected to correct their students' mistakes. Alternatively, the reduction in rater 2's range of scores might simply have been caused by a dulling of the ability to perceive differences after having read over a thousand scripts.

It may also be, however, that rating for six constructs in one reading had an effect on rating behaviour. Rater 2 might have experienced a kind of halo effect, with a subconscious tendency towards consistency causing scores on the six constructs to be grouped more closely together than they otherwise might have been. For example, if the rater was undecided between a 3 and a 4 for *coherence and flow*, a score of 4 for *grammatical complexity*, which had already been decided upon, might have influenced the rater to also choose 4 for *coherence and flow*. The rater might have been unwilling to award scores of, for example, 2 and 5, for separate constructs, to the same composition. The construct least likely to be affected would be that of the six which the rater judged first. Any subconscious "reining in" of scores on constructs judged subsequently might ultimately reduce the overall standard deviations for these. Rater discussions disclosed that all three raters, in fact, disregarded the order of the constructs on the rating instrument. Grades were assigned to constructs either in order of personal preference or in order of which features of a particular composition were most salient, the most difficult category to judge generally being left until

last. *Grammatical accuracy* may frequently have been the most easily, and consequently the first, judged construct for rater 2.

Rater 3's increased strictness, particularly evident in *grammatical accuracy* scores, but also apparent in *grammatical complexity*, *vocabulary range* and, to some extent, *coherence and flow*, may also have been an effect of judging six constructs on one reading, but in a different way. It is possible that giving six different scores "liberated" the rater and that one high or medium score (possibly for *spelling*) permitted her to give lower scores on other constructs than she might normally have done if only a single score was to be awarded. It might also be that giving scores on six constructs per reading made it more difficult for the rater to hold in mind her overall picture of the performance of the group as a whole for each of these constructs. In this case, the rater might not have been aware of her own strictness. It is equally possible that, as a personal tendency, rater 3 simply judged specific constructs more critically than she judged general overall performance. Of the three raters, rater 1 was the only one whose rating behaviour did not vary across overall quality judgements and judgements on constructs.

Nevertheless, the inter-rater correlations for each construct were still high, indicating reliability of both constructs and raters. (Why this was less the case for *punctuation and paragraphing* than for the other constructs is discussed in Chapter 5.) Such differences between raters as are revealed by ANOVA are usually only a threat to reliability if one rater takes over and continues the task of another, or if several raters share the same rating task. (For example, in large-scale commercial testing very large numbers of scripts make it impossible for one rater to mark *all* of these. Several raters may each mark only 500 scripts out of several thousand. In such cases, in fairness to the test candidates, it is important for raters to be of a similar strictness.) In the present research, differences in rater strictness do not threaten reliability, as these differences extend across both the control and the experimental groups. If a rater judges the control students' performance harshly, she will judge the experimental students' performance equally harshly. It is important, however, to be aware of any effects which the rating instrument might have had on rater judgements.

4.3.6 Objective analyses

After the evaluation by raters of the 392 scripts, 150 scripts were subjected to a number of objective text-analysis and error-analysis techniques. These 150 scripts were from two

control and two matching experimental classes from the same school. The reasons why these scripts, and not others, were selected for further evaluation are discussed in detail in Chapter 5. It was hypothesized that — if differences were found, through the *subjective* analysis, to exist between the control and experimental scripts — *objective* analyses would help ascertain more precisely what these differences were. If no differences were found, then these analyses might help to explain *why* no differences had been found, or might reveal differences which were not captured by the raters' evaluations.

To optimize the possibilities of such triangulation, measures which previous research has associated with four of the six constructs were selected. (Punctuation, paragraphing and spelling are not complex enough to have any such associated "synthetic" measurement techniques, and, in fact, these are written-language presentation skills rather than linguistic skills.) Some of the selected measures have been associated with several of the constructs. In the present research, use of certain of the measures was adjusted in some way to better reflect the specific properties of the data, as described below.

4.3.6.1 Quantity of production

Quantity of production, within a particular period of time, has often been associated in the research literature with *fluency* (Wolfe-Quintero, Inagaki and Kim, 1998). Number of words (*e.g.* Hirano, 1991; Ishikawa, 1995) is the most straightforward measure. Number of T-units (*e.g.* Tedick, 1990; Ishikawa, 1995) and number of clauses (*e.g.* Ishikawa, 1995) have also been used as measures of fluency. All three of these measures are, however, subject to the possibly competing effects of *complexity*. Is it a justifiable proposition, for example, that a 200-word composition containing only short, one-clause sentences and very low level vocabulary should be judged as exhibiting greater fluency than a 180-word composition full of complex, much longer sentences and higher level vocabulary (if written in the same amount of time)? Are 100 short T-units equal to 100 longer T-units? Is the production of 100 independent main clauses within a specified time an indicator of the same level of fluency as the production of 50 main clauses and 50 associated dependent subordinate clauses? As with accuracy, it is possible that fluency does not develop in linear fashion, but that levels of fluency may fluctuate with ongoing language development, decreasing as new structures are engaged with and increasing as these structures are mastered, only to decrease again as more new structures enter the learner's repertoire. Quantity of production, and apparent levels of fluency, must be weighed carefully, in context, with due regard to other relevant factors.

The number of words in each composition was retrieved using a computerized word count facility. T-units were coded using a slightly modified version of Lennon's (1990) definition. Hunt's original definition of a T-unit as "one main clause, with or without subordinate clauses [which may be] punctuated with terminal marks at both ends" (1970: 199) does not sufficiently clarify the position of coordinate clauses. This may be clear when coordinate clauses are two full main clauses joined by a coordinator. (*e.g.* "He lives next door and his mother is a musician" might be punctuated as two sentences, if one either removes the "and" or accepts it as the first word of a sentence), but Hunt's definition does not take into account the case of coordinated clauses with ellipted subjects. Lennon, on the other hand, states that "a coordinate clause is a T-unit itself unless conjunction reduction occurs" (1990: 406). Lennon further provides the example of "He goes to the bookmaker and gets some money" (*i.e.* a main clause with an ellipted coordinate clause) as being *one* T-unit. In fact, the effect of this common ellipsis of the subject in the second of two coordinated main clauses is little discussed by researchers, although — if the genre of learner text being studied is likely to include much post-coordination ellipsis — it may make a substantial difference to the counts and mean length of T-units. In affected sentences, whilst the number of T-units is halved by such ellipsis, the mean length of T-unit is doubled.

Sentences which do not contain verbs are not well integrated into descriptive grammars of English. These may range from one-word units such as "Thanks" to longer units such as "What an ugly old man", and are referred to variously in grammar books as *interjections*, *formulae*, *phrasal exclamations* and *minor sentences*. They are not clauses since they do not have a verb. Although standard definitions of a T-unit, including those of Hunt and Lennon, specify a main clause as the first necessary condition, the definition was extended in this study to include the verbless sentence as also representing a production unit. Whilst such units are not strictly T-units, they are treated as such for the purposes of measuring quantity of production.

The complete coding guidelines for T-units may be found in Appendix 19, as may a sample coded composition. The word "so" is deserving of special mention since it is classed differently by different grammarians. Quirk and Greenbaum (1979) class "so" as a conjunct, whereas Crystal (1988) classes "so" as a subordinator. If "so" is classed as a subordinator, then a sentence such as "He was tired, so he went to bed" represents one T-unit, with a main clause and a dependent adverbial clause of result. If "so" is classed as a conjunct, then the same sentence would represent *two* T-units. The solution adopted was to respect the

ideational units, which is to say the sentence boundaries, of the learner. When "so" was used in the middle of a sentence it was classed as a subordinator, and the sentence above would have been coded as one T-unit. When "so" was used at the beginning of a sentence, it was classed as a conjunct. Hence "He was tired. So he went to bed." would have been classed as two T-units.

4.3.6.2 Syntactic complexity

Clause-type

Clauses were also counted, and, in addition, coded into four principal categories. Whereas clause-type may not be of immediate relevance if number of clauses produced is merely being used as a fluency measure, it is of interest when investigating text sophistication. As a measure of complexity in spoken language, Skehan and Foster (2001) found that, compared to measures of range of structures used, subordination correlated with, but was more robust than these. In written texts, Homburg (1984) found that number of dependent clauses (subordinate and relative) significantly discriminated between two ability groups. Hirano (1991) found that the ratio of dependent to independent clauses significantly differentiated the writing of students at three programme levels.

The four principal categories of clause were:

main clause
full coordinate clause
"reduced" coordinate clause
subordinate clause

A main clause was an *independent* clause which "could in principle stand as a sentence on its own" (Crystal, 1988: 176). Coordinate clauses were split into two types as it was felt that there was a qualitative difference between a full coordinate clause, which is simply a main clause prefaced by one of the three coordinators, *and*, *but* or *or*, (*e.g.* "He got up and he made breakfast"), and a coordinate clause (termed here "reduced") which exhibits closer syntactic and ideational links to the main clause through use of ellipsis, usually of a subject pronoun (*e.g.* "He got up and made breakfast"), but sometimes also of a verb or even of an adverbial phrase. (*e.g.* "He went to a restaurant with his friend, and then to the cinema"

might be recovered as "He went to a restaurant with his friend and then he went to the cinema with his friend". "And then to the cinema" is thus a reduced coordinate clause.)

Subordinate clauses were further coded into three categories of *complement* clauses, *adverbial* clauses and *relative* clauses (Hurford, 1994: 29). These were also divided by type, with four types of complement clause, seven types of adverbial clause and two types of relative clause (which is to say defining and non-defining). Since individual subordinate clause types were in many cases relatively too infrequent or erratically distributed to allow reliable between-groups comparisons, these were grouped together for the main analysis. However, a full taxonomy of identified clause types, with examples, is given in Appendix 22. Ratios of subordinate and coordinate clauses to total number of clauses (as measures of syntactic complexity) and of "reduced" coordinate clauses to full coordinate clauses (as a measure of the use of ellipsis) were calculated. Relative clauses were included in the counts for subordinate clauses, but were also investigated separately.

Length of syntactic unit

Although mean length of syntactic unit has been associated with writing fluency (*e.g.* Yau, 1991), it is more commonly used as a measure of syntactic complexity. Mean length of sentence, mean length of T-unit and mean length of clause have all been used by researchers seeking to establish differences amongst writing samples (Ortega, 2003). Mean length of T-unit is often preferred to mean length of sentence, as this eliminates the contaminating effects upon data of unstable or non-standard use of punctuation. A common problem, for example, is the so-called *run-on* sentence, where a learner produces a series of main clauses separated by commas (*e.g.* "He got up, he had breakfast, his friend arrived."). Such misuse of punctuation can drastically affect the validity of sentence length as a measure of syntactic complexity. T-units overcome this problem — and, indeed, were originally designed specifically to do so — by disregarding punctuation.

The reverse side of this coin is that, although T-units may have the advantage of being under the immediate control of the researcher, they do not, however, respect author intention at either the syntactic level *or* the meaning level. Although coordinators *may* operate as unmarked connectors of main clauses, making a division of the joined clauses into two T-units more theoretically viable, we should not lose sight of the fact that coordinators are used *intentionally*, and there is very often a relationship, for example of cause and effect, between

the coordinated clauses. For instance, the use of "and" in the sentence "The baby cried and she picked it up" implies cause and effect, and it might be argued that "and she picked it up" does not, psychologically, represent an independent second main clause. There may be other relationships, such as of time, or of purpose, between two coordinated clauses. In any case, the T-unit discounts the learner's knowledge of coordination, and may not reflect accurately the learner's grammar. It is, in fact, the sentence which is the *intentional* unit of the learner.

Both measures were calculated, as was mean length of clause. In addition to the measures themselves, this would allow for computation of complexity ratios such as clauses per T-unit and clauses per sentence. Since the clauseless sentences described above (Section 4.3.6.1), included in calculations for *quantity* of production, occurred almost exclusively in direct speech, and were typically much shorter than standard T-units, these were excluded from the calculation of mean *length* of T-unit, in order not to disadvantage those compositions which had made more use of direct speech. Whilst some compositions contained relatively extended tracts of directly reported dialogue, others made no use of direct speech. Including these short, verbless expressions in the calculation may lower the mean length of T-unit and lead to a possible false conclusion as to the ability, or not, of the learners to produce longer T-units. Such text units are not symptomatic of language proficiency levels, as they do not become any longer with increased proficiency, but are simply markers of the spoken genre. The amount of direct speech within a narrative composition cannot be held to have any relationship with language proficiency. Two-word units which contained a verb (for example, "He left", "Come here!") were not excluded from the calculation.

Mean length of T-unit was calculated by removing the number of words contained in the excluded units from the total number of words in a composition and dividing this by the number of standard T-units. A full list of excluded units, and the numbers of each, for control and experimental compositions separately, may be found in Appendix 20. These were excluded only from calculations of mean T-unit length. Other calculations using T-units, for example numbers of words contained in error-free T-units, used integral texts.

Punctuation in the compositions was very unstable, and in some cases distinction did not appear to be made at all between commas and full-stops. (This is a common feature in the writing of Hong Kong secondary schoolchildren.) For this reason punctuation was corrected before calculating the number of sentences and mean length of sentence. The rules for

correcting punctuation, and a sample uncorrected and corrected composition are set out in Appendix 18.

A similar problem with the mean length of sentence as that concerning the mean length of T-unit was envisaged. Narrative text and informal spoken dialogue of the type reproduced in the compositions are two quite different genres. The second, with its high proportion of characteristically short sentences (examples from the data include Wh-questions such as "Who are you?", requests such as "Please help me" and commands such as "Give it to me") might obscure the mean sentence length of the first, if included in the same calculation. For this reason, two calculations were made: one for overall mean sentence length, obtained by dividing the total number of words in a composition by the total number of sentences, and a second for narrative text only, obtained by dividing the total number of words contained in the narrative text by the number of sentences in the narrative text. Sentences which were part narrative text and part directly reported speech (*e.g.* I took the book and said him "I want the book!") were also excluded from the second calculation.

Clauses were the least likely units to be affected by genre, since they are the shortest of the three units and, unlike T-units and sentences, are not composed of varying numbers of smaller sub-units. All clauses were counted. Length of clause was calculated by subtracting the number of words in a composition which were not contained in clauses (for example, "No", "But why?") from the total number of words and dividing by the number of clauses.

4.3.6.3 Coding reliability

All coding was blind, which is to say that scripts were not identifiable as to their origin, and was done in random order. Clauses were coded twice, using unmarked scripts, with an interval of 12 months, giving a Time 1/Time 2 correlation (Polio, 1997) of .989 (N = 150). This was judged as being sufficiently high to use the first set of data without adjustment. The reason for such a high correlation was most likely that clauses benefit from a very strict definition, and may be confirmed by the real or inferred presence of no more than one acting verb. A small number of clauses had no verb due to ellipsis (*e.g.* I was very frightened, *and Mary too.*) or to grammatical inaccuracy (*e.g. The second wish a lot of money.*).

Setting sentence boundaries which were not marked by the learners (*e.g.* dividing run-on sentences; removing full-stops after fragments) was more susceptible to subjective variation. These were also coded twice, with an interval of 12 months, giving a Time 1/Time 2 correlation of .967 (N = 150). The number of sentences calculated at Time 1 and Time 2

were then added together and halved. For example, if a composition produced 32 sentences at Time 1, and 33 sentences at Time 2, the figure used for calculations involving number of sentences (*e.g.* sentence length; number of clauses per sentence) would be 32.5.

T-units were not double-coded. Number of main clauses, full coordinating clauses and verbless sentences, coded on a separate occasion, were added together and the resulting total for each composition was correlated with counts of T-units — the inclusion of one, and only one, of any of these three syntactic units being the necessary condition for a T-unit. The correlation thus obtained was .986 (N = 150), and the coded data was used in preference to the calculated data.

Other coding reliability checks were also possible by triangulating the data sets in this way. Number of clauses and number of verbs (*cf.* Section 6.13) showed a correlation of .985 (N = 150). Number of main clauses + verbless text units showed a correlation of .98 (N = 150) with number of sentences.

4.3.6.4 Measures of accuracy

Apart from the counting and classifying of individual errors, the most common measure of linguistic accuracy used in text-analysis is the *error-free T-unit*. Accuracy may be evaluated in terms of total number of error-free T-units or as the ratio of error-free T-units to total number of T-units (*e.g.* Robb, Ross and Shortreed, 1986; Casanave, 1994). Total number of words contained in error-free T-units and proportion of words in a text which are contained in error-free T-units have also been used (*e.g.* Polio, Fleck and Leder, 1998; Robb *et al.*, 1986).

T-units were coded as either "error-free" or "with-error". Distinctions were not made as to type or degree of error, or number of errors within the same T-unit. "Error" was defined as grammatical or lexical error, and did not take account of spelling mistakes, minor errors of usage or incorrect punctuation. The reason for this relative laxity of definition of "error" was that, with such low-level learner text, a stricter definition would have meant that many compositions contained no error-free T-units, resulting in a possible "floor" effect and a less reliable variable. (Coding guidelines and examples of *error-free* and *with-error* T-units may be found in Appendix 21.) Error-free T-units and number of words contained in error-free T-units were counted. The percentage of whole text included in error-free T-units, written as number of words contained in error-free T-units per 100 words, was then calculated.

Spelling mistakes were recorded separately. Two counts were made for each composition. The first was the total number of instances of misspelt words, irrespectively of whether these included the same words repeatedly misspelt. In the second count, words which were repeatedly misspelt were only counted once. Thus each composition had a *token* and a *type* count of misspelt words.

4.3.6.5 Past tense verb forms

Patterns of error in past tense verb forms seemed a potentially particularly fertile ground for investigation in this study. In psycholinguistics, many practitioners "take it for granted that linguistic expressions are either computed or stored, and that, if a form can be computed, it is not stored" (Dabrowska, 2004: 20). In English, almost all the work which has been done in this area has been on regular and irregular inflections, of plural forms and simple past forms (*e.g.* Pinker and Ullman, 2002). According to the declarative-procedural model proposed by Ullman (2001a), regular simple past forms (*e.g.* walk*ed*, look*ed*), may be computed, whilst irregular forms (*e.g. saw*, *took*) cannot be computed, but must be memorized and stored.

Dabrowska suggests that one of the "tell-tale signs of memory storage is sensitivity to frequency" (2004: 20). It was hypothesized that extensive reading, which may greatly increase students' exposure to both regular and irregular simple past forms (as opposed to normal classroom exposure), would have a greater effect on the acquisition of irregular past forms than regular past forms. As these are otherwise well-matched in terms of complexity and usage there may be few other intervening factors. The data sample (compositions) was especially suited to such an investigation because of the narrative genre of both the extensive reading materials and the writing task, which might be expected to contain high levels of simple past. In addition, because participation in the reading scheme had been long-term, memory effects might be expected to be more apparent.

A data-driven approach was adopted, following the first three stages of Corder's (1974) procedure for error analysis — selection of corpus, identification of errors and classification of errors. (The last two of Corder's stages — explanation of the causes of the errors and ranking the errors for gravity — are outwith the scope of this study.) The simple past verb forms from an initial sample of ten compositions were analysed. Only verbs contained within the narrative text were used for the study. Verbs contained, or which should have been contained, within reported speech quotation marks were not included. (These were

rarely instances of the simple past, but, in general, a mixture of present tenses, futures, conditionals and imperatives.)

A categorization system was established, with simple past verbs classed in the first instance as regular simple past, irregular simple past, simple past BE, or "past" modal (*e.g. could, had to*). A secondary level of classification established these as declaratives, negatives or interrogatives and a third level of classification as correct or incorrect in form. A taxonomy of recurring types of incorrect forms for each category was created and a hierarchical coding system developed. For example the code "IDC" indicated an irregular, declarative, correct form. The code "RNE3" indicated a regular, negative form with error "3" (an example of this error-type being "didn't believed"). Use of the simple present form when simple past was required was included as one of the error types. Other error-types included forms such as "bringed", modal errors such as "can went" and negatives such as "don't brought".

The simple past verb forms in all the compositions were then tagged using the coding system. If an error was identified which did not fit into the existing categories, a new error code was created. In instances where it was unclear what type of error was displayed, or if a combination of errors was apparent, the verb was coded as "uncoded error". (The complete coding system and examples of coded compositions are available in Appendix 24.) Verbs which were contained in the narrative text but which were not simple past were recorded only by tense and whether or not they were correct in use and in form. In practice, there were very few instances of these.

Counts were made, using WordSmith Tools (Scott, 1998), of each error-type and correct to incorrect ratios were calculated for verb forms within each of the different categories (*e.g.* regular declarative, regular negative, etc) for each of the four groups as a whole. Group performance would afford a more coherent picture of any overall emerging patterns of correct and incorrect simple past forms than examining the performance of individuals.

4.3.6.6 Vocabulary measures

Lexical variety

It is now largely accepted by the research community that the type-token ratio — the ratio of unique words (*types*) to total words (*tokens*) in a text, which has been mooted as a measure

of *lexical variety* — is neither a reliable nor a necessarily valid indicator of proficiency level in the area of vocabulary. The report from the BAAL/CUP Seminar on Vocabulary Knowledge and Use: Measurement and Applications (2004) stated that "there was unanimous agreement that especially one measure that has been used widely in research (the type-token ratio) is extremely unreliable" (Treffers-Daller, 2004: 24). Technically, it is an inherent feature of the type-token ratio that it is sensitive to text-length. The longer a text is, the more likely lexical items are to be repeated, leading to a lower type-token ratio, and it is not justifiable to compare the type-token ratios of two texts of different lengths. Some researchers have attempted to bypass this problem by applying the measure to segments of text of a standard length, although the question is then raised of comparability of segments. In comparing, for example, the first hundred words of a short narrative composition, which might have only 200 words in total, against the first hundred words of a longer narrative composition, which might have 400 words, we are comparing one half of the content of the first composition against one quarter of the second. So we might be comparing the beginning and middle of one composition with only the beginning of the second. Moreover, type-token ratios may vary across different segments of the same text. Malvern and Richards further argue that using short segments is "likely to distort results because they are not sensitive to repetition of words beyond the boundary of their own segment" (2002: 88). In any case, measuring lexical variety in this way is simply monitoring levels of repetition. There are no theoretical grounds for contending that higher type-token ratios, indicating lower levels of word repetition, necessarily indicate greater quantities of known vocabulary.

Lexical density

Lexical density is concerned with the ratio of lexical words to non-lexical, or grammar words, sometimes referred to as "functors", and is calculated as the number of lexical tokens divided by the total number of tokens. Since this is a ratio measure, as the number of lexical items proportionately *in*creases, so must there be a corresponding *de*crease in the proportion of non-lexical words. This, in fact, is just as likely to be an artefact of the syntactic structures used, and of writing style, than of the fact of knowing more lexical items. Whilst lexical density may be of interest in comparing genres, there is no reason to suppose that, within a given genre, increased lexical density means increased vocabulary knowledge. As Engber remarks of the results of her study comparing lexical density indices to quality scores of L2 student writers, "simply piling up lexical words did not affect the quality scores" (1995: 148).

The most appropriate measures for the present study, which might produce more theoretically valid and also technically reliable results, were hypothesized to be *lexical sophistication* and *lexical originality*.

Lexical sophistication

Lexical sophistication measures distinguish between knowledge of common or "easy" words (defined as such by their general high frequency within the English language) and knowledge of less common, or "more difficult" words. Thus, using words from lower frequency categories may be seen as indicative of qualitative vocabulary development, and knowledge of ten "easy" words will not be considered comparable to knowledge of ten "difficult" words. Rather than simply being seen as acquiring (or using) more and more words, learners may be seen as progressing through the first stages of common vocabulary knowledge to more advanced stages.

Two lexical sophistication evaluation tools were applied to the data. The first was Web VocabProfile (Cobb, *n.d.*), a publicly available web version of Laufer and Nation's Lexical Frequency Profile (LFP). As its name suggests, the LFP provides information on numbers of high and low frequency words in a given text, using frequency ratings based on West's General Service List of English Words (West, 1953) to discriminate between "the first 1,000 most frequent words, the second 1,000, and any other vocabulary" (Laufer and Nation, 1995: 311).

As Bauman points out, although the General Service List (GSL) may be old enough to differ significantly from the English of today, "the core vocabulary of English changes more slowly, so at the frequency level of the first 2,000 words this may be less of a problem" (Bauman, 1996: *unpaginated*). Indeed, a programme based on the GSL may still be one of the best available computerized options, since "other corpus-based lists need substantial adjustment to make them appropriate as vocabulary standards. These adjustments have already been made to the GSL" (*ibid*). Additionally, most of the graded readers used by the ERS were connected to the GSL, being based on vocabulary grading schemes derived from it, and any increased sophistication of vocabulary awareness and use on the part of the ERS students may have evolved within, and been to some extent influenced by, this framework.

Web VocabProfile was thus used to obtain type and token counts at each frequency level for each composition, with the adjustments described below.

Words which may be low frequency in general may be very high frequency in certain environments. Web VocabProfile recognises this fact and provides a "recategorization" facility, allowing the user to customise the programme for any particular data set by moving words initially identified by the programme as low frequency into a higher frequency category. Both *classmate* and *exam*, for example, appeared in the "off-list" category of Web VocabProfile, which is to say words which are less frequent than the first 2,000. Hong Kong students, however, in English-medium schools, are likely to be frequently exposed to these words, and they were moved to the highest frequency category.

Words which were moved from lower frequency categories into the highest frequency category may be characterized as falling into three main types. These were classroom and school vocabulary, some transport vocabulary (all notices concerning public transport in Hong Kong were, at the time the compositions were written, in both Chinese and English and so exposure to words such as *airport* and *ferry* was liable to be high) and "global" English words, often related to popular culture and recreation, such as *cinema* and *television*, which may be used even by people speaking in Chinese. In addition, proper names — of Hong Kong locations and of the student writers' fictional companions — and words contained in the writing task rubric which might otherwise be classed as low-frequency words, such as *ghost* and *magician*, were also moved to the highest frequency category. (A complete list of recategorized words may be found in Appendix 25.)

The LFP has been designed for a broad use, across the widest range of proficiency levels. In practice this may render it less likely to make finer distinctions of vocabulary progress, particularly within the lower proficiency levels. In the case of lower intermediate learners, for example, Morris and Cobb (2004) have suggested that approximately 90% of lexis may be expected to derive from the 1-1000 band. Although Web VocabProfile further divides the 1-1000 band into the first 500 and second 500 words, it was felt that a more finely-tuned evaluation instrument was needed to capture any more subtle shifts in vocabulary use. To complement the use of Web VocabProfile as a measure of lexical sophistication in absolute terms, providing an anchor to external norms of proficiency and also a potentially useful link to other research, an *internal* frequency list was devised, allowing patterns of vocabulary use across the four groups within the data set to be studied more closely.

The compositions from all four classes were combined to create one master corpus of learner text (N compositions = 150; N words = 41,992). The computer programme WordSmith Tools (Scott, 1998) was then used to generate a list of all the words in the corpus, in order of frequency of use. From this master list, ten frequency levels were set, deriving from what appeared to be natural breaks in frequency patterns and considerations of what might be workably-sized vocabulary groups for optimal discrimination between students' performance levels (Table 4.5). For example, there seemed little point in working at a level below the first 25% of text coverage, since, surprisingly, only seven words accounted for this. At the other end of the frequency scale, however, it seemed important to discriminate as finely as was practical, since this appeared likely to be where differences might emerge, and boundaries consistent with only approximately 5% text coverage were set. Word frequencies were also considered in relation to the group size. With data from only 150 students, words at level 2 (161-283 uses) were quite likely to have been used by the majority. At level 3, however, approximately half the students are likely to have used a word, at level 5, approximately a quarter, and at level 9, only one or two.

Table 4.5 Initial internal word frequency level settings

Level	0	1	2	3	4	5	6	7	8	9
N unique words	7	14	19	32	52	97	104	149	293	1,203
cumulative % of text coverage	25.53%	40.57%	50.28%	60.38%	70.28%	80.19%	85.76%	90.07%	94.47%	100%
frequency range in corpus *(number of times used)*	2,237-1,052	614-314	283-161	155-109	106-61	60-31	30-17	16-10	9 -5	4 -1

Word counts made by WordSmith, upon which the decisions described above were based, are mechanical, and require exact matches. Associations between, for example, singular and plural forms, or first and third person forms, are not made. *Friend*, *friends* and *friend's*, for instance, are taken as different words, as are *eat* and *eats*, or *like* and *liked*. These words may appear at different frequency levels in WordSmith's output. To regard them as different

types may not produce an accurate picture of vocabulary knowledge. Instead, they may inflate the number of types merely by a student's application of basic grammar rules.

The master frequency list was examined, and, following a set of criteria detailed in Appendix 26, pairs or groups of associated word forms were re-counted as a single type. Only words from the same word class were grouped together. For example, *old* and *older* were counted together; *quick* and *quickly* were not. The rationale behind this was that no additional *vocabulary* knowledge is required in order to apply morphological changes to a root word within the same word class. In the above example, for instance, assuming *old* is known to be an adjective, which is an integral part of knowing its meaning, even if such knowledge is not declarative, no further knowledge of the word is needed in order to compute the comparative form *older* (although the *grammar* knowledge that short adjectives form the comparative in this way is necessary). On the other hand, not all adjectives may form an adverb by the addition of *-ly* (for example *big*, *old* and *fast* do not), and the use of an adverb implies additional, or different, vocabulary knowledge from that implied by use of the associated adjective. Similarly, the infinitive *rain* and participle form *raining* were counted together, but the noun *rain* retained its identity as a separate type.

Necessary adjustments to the level assignments of affected words were then made. An initially low-frequency word may thus have been re-assigned to the higher-frequency level of an associated word form, or two (or three) associated word forms counted together may have resulted in such a combined number of uses that both (or all three) words were re-assigned to a higher-frequency level. In this way numbers of types and percentages of total text coverage at each level shifted slightly from the initial blueprint. Names of people (the writer's fictional friends) and the Chinese names of Hong Kong locations (written in Roman alphabet; for example *Lantau*) were removed from the word list. Names of countries were not. Spelling mistakes were disregarded, unless the word was unidentifiable, in which case it was removed from the count.

Verbs raised a particular set of questions, firstly because many verbs have irregular past forms which cannot be computed via taught grammar rules, and secondly because of the constraining nature of the task rubric on verb forms. In *regular* verbs, all present and past tense forms display knowledge of the common lemma, and it might be justifiable to count all forms together as exhibiting knowledge of a single lexical *type*. However, use of *eat*, or *eating*, does not necessarily imply knowledge of *ate* or *eaten*. The root form, the *-s* form and

the *-ing* form were therefore counted as one type, and the simple past was considered as a different type, for both regular and irregular verbs, since its use displayed additional knowledge of the lexical item. With irregular verbs, it showed knowledge of the different form, and with regular verbs, it showed knowledge of the fact that the verb used the regular form in the past. Past participles of irregular verbs were also treated as new types, displaying different knowledge from that required to produce the irregular simple past form, but past participles of regular verbs were not, since it is a basic grammar rule that the regular past participle is the same as the simple past form. Knowledge of either the regular past participle or the regular simple past implies knowledge of, or the means of computing, the other, and the two forms were grouped together.

Where *lexical* items used in the writing task operated under conditions of free variation, choice of verb *tense* did not, since students were required to write a past narrative. At an intermediate level, a very high frequency of simple past forms might be expected, with such forms exhibiting higher frequency of use than present tense forms. These relative frequencies, rather than representing students' free preferences, may be merely an artefact of the genre.

Since the Hong Kong English syllabus, like most language syllabuses, did not teach past forms before teaching present forms, it must be regarded as counter-intuitive for present tense forms to represent higher level vocabulary knowledge than past tense forms of the same verb. However, simply to count all forms as exhibiting equal knowledge of a verb would be to discount additional knowledge shown by the production of a past form. The following rules were therefore adopted:

if the simple past form was more frequent in the data, this form subsumed all present and root forms, and all instances of production except irregular past participles (which were extremely rare in the data) were counted together;

if the simple present was more frequent than the simple past — in a genre which *invited* the use of the simple past and gave relatively little opportunity for present tenses — this was taken to indicate that the past form may not be widely known by the group as a whole. In this case, past forms were counted separately, and may have achieved a higher (lower frequency) band in the internal frequency list.

Using the revised word frequency levels, the data set was tagged and token counts at each frequency level were extracted from each composition using the editing facilities of the computer programme *Word*. The relative performances of control and experimental groups were then compared, and information thus obtained was also compared with the information obtained by the application of the GSL-based programme Web VocabProfile. (Appendix 27 contains the complete internal frequency list.)

In addition to the vocabulary use displayed by individual compositions, vocabulary *pools*, for each of the four classes, were created. In an open writing task, assessing vocabulary proficiency is possibly more problematic than assessing proficiency in the other language constructs of complexity, accuracy and coherence, in the sense that the task may not allow students to demonstrate their full competence. It is possible to demonstrate high levels of complexity, accuracy and coherence within a very limited space, even within one paragraph. However, whilst a short narrative composition may allow students to display their full command of syntax, such a task — indeed *any* short writing task — might not allow an accurate representation of the breadth of their vocabulary knowledge. Vocabulary which is irrelevant to the task simply will not be used. In order to give a broader picture, the vocabulary from all the compositions of each class was pooled at each Web VocabProfile frequency level and at each internal frequency list level, allowing comparison between the pools of vocabulary used by control students and experimental students at each of these levels. This may offer insights into any differences between the vocabulary of the control and experimental students brought about by participation in the ERS which may be obscured by the practical limitations on vocabulary use imposed by the writing task.

Lexical originality

Lexical originality is usually taken as a measure of the number of words which are unique to one writer in relation to a particular group. Whilst it may be possible for each student in a group to produce a similar number of types — thus seemingly demonstrating like performances — many of these students may, in fact, be producing the same types as each other. Lexical originality gives credit for producing types which others in the group do not.

In the present research, lexical originality was re-defined as being that of the group rather than of the individual. Lists of types produced by each class were compared manually and types which had been used by only one class were extracted. Since the four classes

investigated were not of equal size, and it might be expected that 40 students may produce more *unique-to-group* types than 30 students, *lexical originality* was calculated as the number of words unique to a group divided by the number of students in that group.

4.3.7 Significance testing

Although some researchers have criticised the application of multiple significance tests to the same data (*e.g.* Brown, 1990), maintaining that, as the number of significance tests increases, so should the p-value be correspondingly set lower, this kind of adjustment is neither theoretically valid nor mathematically possible with correlated variables (Tabachnick and Fidell, 1996). The significance level was therefore set at $p < .05$, and actual p-values are reported where these may provide useful information to the reader.

5. PRELIMINARY RESULTS

5.1 First range of overall quality ratings

The *overall quality* ratings (on a scale of 1 to 6) of the three raters were combined to give one *overall quality* score for each composition, ranging from a lowest possible score of 3 to a highest possible score of 18. This and all subsequent data was entered into an SPSS database. As had been expected from the choice of rating procedure described in Section 4.3.4.2, the scores produced a normal curve. (A bar chart for the data distribution is reproduced in Appendix 7.) Descriptives obtained are given below (Table 5.1).

Table 5.1 Overall quality scores: descriptives for whole data set (control and experimental)

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
overall quality score *N = 392*	3	18	10.34	3.14	10	11

Range of possible scores = 3 to 18

Splitting the data into control and experimental groups yielded the information in Table 5.2.

Table 5.2 Overall quality scores: descriptives for control and experimental groups

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
control group *N = 190*	3	18	10.36	3.51	10	10
experimental group *N = 202*	3	18	10.32	2.76	11	11

Range of possible scores = 3 to 18

These results suggested that there was little difference between the *overall quality* scores of the control and experimental groups. Although the experimental group had a higher median and mode than the control group — by one point only in a range of 16 possible scores — the means of the two groups differed by only 0.04. The lower standard deviation of the experimental scores did, however, indicate that there was more centring around the mean for

that group. Examination of individual score frequencies revealed that whereas 18.4% of the control students had scores within the ranges of either the top three scores (8.4%) or bottom three scores (10%), only 8% of the experimental students had scores within these ranges (4.5% in the top three scores and 3.5% in the bottom three). An initial interpretation might be that the reading scheme had the effect of homogenizing students somewhat, by favouring weaker students whilst having no effect, or even a detrimental effect, on the writing performance of stronger students. An independent-samples t-test found, however, no significant difference between the two groups ($t = .129$; $p = .897$).

5.2 The second rating process

A second scoring procedure, described in Section 4.3.4.5, required raters to make a further judgement on each composition within its originally assigned grade band. Each grade band was divided into four sub-bands, and the range of possible scores given by one rater to a composition was thus expanded from 1 to 24. Scores from the three raters were combined to give a second *overall quality* score, ranging from a lowest possible score of 3 to a highest possible score of 72 (Tables 5.3 and 5.4).

Table 5.3 ***Second range of overall quality scores: descriptives for whole data set (control and experimental)***

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
overall quality score *N = 392*	3	72	36.90	13.34	37	37

Range of possible scores = 3 to 72

Table 5.4 ***Second range of overall quality scores: descriptives for control and experimental groups***

	Minimum	Maximum	Mean	Std. Deviation	Median	Modes*
control group *N = 190*	3	69	36.97	14.73	37	36 & 37
experimental group *N = 202*	3	72	36.84	11.93	37	31 & 39

Range of possible scores = 3 to 72
** Both data sets were bimodal*

The rationale behind extending the range of possible scores in this way was to increase the validity of the rating process by allowing raters to make distinctions between, for example, a low level 5 composition, which may have only just achieved that level, and a high level 5 composition which the rater may even have considered placing in level 6. This secondary grading procedure would allow for the possibility that the reading scheme *may* have caused differences between the writing of the control and experimental students, but which were very subtle, either not of sufficient magnitude to justify a composition's being placed at a higher level — which might result in most of the lower compositions at a given level being those of control students and most of the higher compositions at that same level being those of experimental students — or in ways which might not be of primary influence in the raters' initial perceptions of *overall quality*, but which might subsequently come into play in the secondary appraisal. Such differences might entail features not normally considered by classroom teachers to be of primary importance — although the raters may not deliberately, or even consciously, have weighted their decisions more on certain features than on others — or simply features which might be less salient.

All three raters reported during debriefing that they found the second rating procedure less stressful than the first. With the first, and what they felt to be the main, judgement already made, the raters felt a certain liberation from the responsibility of possibly making a wrong judgement. Later standard multiple regression procedures, using as independent variables subsequent scores in the five categories of *grammatical complexity*, *grammatical accuracy*, *vocabulary range*, *punctuation and paragraphing*, and *coherence and flow*,[1] and the first and second *overall quality* scores, separately, as dependent variables, revealed that the second rating procedure represented the five variables slightly better, particularly *vocabulary range* and *punctuation and paragraphing*. This suggested that the raters may have weighted their immediate primary judgements on accuracy, complexity and coherence, leaving vocabulary range and punctuation to be subsequently factored in as a refinement of the initial judgement. The differences between the two sets of scores revealed by the standard multiple regression were, however, subtle. Whereas 75.3% of the initial *overall quality* ratings could be explained by the five variables ($R^2 = .753$), 79.2% of the second *overall quality* ratings could be explained ($R^2 = .792$). In other words, the second rating procedure allowed the raters to factor in another 3.9% of the five variables.

---

[1] *Spelling* was not included, as the distribution did not meet requirements for the statistical procedure.

The premise that the second rating procedure would either enhance any differences between the control and experimental groups suggested by the first rating procedure, or would reveal differences which the first rating procedure did not, did not prove to be the case. The second procedure may have allowed finer distinctions to be made, but this was distributed equally across both control and experimental compositions. Of the control compositions, an average of 49.6% (across the three raters; *i.e.* 54.7%, 43.7% and 50.5%) were assigned to the lower half of their original grade level. Of the experimental compositions, an average of 51.2% (*i.e.* 53.5%, 49.5% and 50.5%) were assigned to the lower half of their grade level.

Additionally, when the same standard multiple regression procedure as described above was performed on the control and experimental compositions separately, results were similar for the groups. For the control group, R^2 was increased from .786 to .822, indicating that 78.6% of the first *overall quality* ratings and 82.2% of the second *overall quality* ratings could be explained by the five constructs, and from .723 to .763 for the experimental group, indicating that 72.3% of the first *overall quality* ratings and 76.3% of the second *overall quality* ratings could be explained by these constructs. In the case of the control compositions, the second rating procedure had allowed the raters to factor in another 3.6% of these five variables, and in the case of the experimental compositions, it had allowed for another 4%.

The second rating procedure produced virtually the same picture as the initial rating procedure. The mean for the group as a whole remained fractionally below the mid-point in the range of possible scores (0.16 below the mid-point of 10.5 for the first rating procedure and 0.6 below the mid-point of 37.5 for the second rating procedure — which magnified any differences by four because of the way it was constructed). When the data set was split, the experimental group had in both cases a negligibly lower mean than the control group (0.04 lower for the first rating procedure and 0.13 for the second), and a considerably lower standard deviation.

Although bimodal distributions may sometimes be an indication of two distinct populations within a group, in this case they were most likely an artefact of the way in which the range of scores had been created. Whilst the overall range of scores from 1 to 6 created a normal distribution, the sub-range of four scores within each band also tended towards a normal distribution, with raters more commonly awarding the middle scores of 2 and 3 (32.9% and 29.8% respectively), and awarding fewer of the top and bottom end scores of 1 and 4 (17.5% and 19.8%). This had the effect of creating a series of mini-peaks on the line of distribution,

where scores arising from combinations involving the sub-ratings of 2 or 3 were more common than scores which arose from combinations involving the sub-ratings of 1 and 4.

This may be illustrated from the distribution of a single rater. Within the range of four new scores derived from band 1 in the original scoring system — *i.e.* the scores of 1, 2, 3 and 4 in the new scoring system — scores 2 and 3 were more often awarded, creating a within-band mode of one or other of these two scores. From the *next* band — band 2 in the original scoring system — arose the new scores of 5, 6, 7 and 8. However, within that range of four scores, the two central scores of 6 and 7 were more frequent, since raters tended towards the within-band scores (out of 4) of 2 and 3. So a within-band mode of either 6 or 7 would result. The next within-band mode, arising from the original band 3, would be the score of either 10 or 11, and so on. The within-band modes would have a minimum of two points and a maximum of four points separating them. However, since the final range of scores was derived by summing the scores of *three* raters, not only was each composition score amplified, but the distance between any two adjacent within-band modes was similarly amplified. Thus the observed distance of 8 points between the two modes of the experimental distribution was caused quite simply by each of three raters contributing 2 or 3 points to the distance between the two modes of adjacent original bands (*i.e.* bands 3 and 4), and did not necessarily point to the presence of two distinct populations.

Whilst the second rating procedure did enhance inter-rater reliability (.90, compared to .853 in the first rating procedure; see Sections 4.3.4.4 and 4.3.4.5), it did not reveal any additional information about differences between the two groups. The results of an independent-samples t-test ($t = .101$; $p = .919$) were very similar to the results obtained from the t-test performed on the first set of results and showed that the initial rating procedure had not favoured either of the two groups. Additionally, the second rating procedure showed that the original six-point scale was both valid and robust, since the distributions it produced resisted change even when a favourable condition for change was subsequently applied. For the data set, a six-point scale proved appropriate.

5.3 Inter-rater reliability for control and experimental groups

Apart from the diminished standard deviation of the experimental group, one other interesting difference between the two groups which emerged from the first stages of data

analysis was that inter-rater correlations were lower for the experimental group than for the control group.

Inter-rater reliability was calculated using the Pearson correlations for each pairing of raters over the complete data set of 392 compositions. These three correlations were averaged (converting them first to an interval scale using a Fisher Z transformation) and a Spearman-Brown Prophecy Formula adjustment was then made (Henning, 1987). This adjustment is used when there are more than two raters, and reflects the fact that any agreement between two judgements is given further credibility if a third judgement corroborates it. Thus the combined agreement of three raters is more reliable than the agreement between any two of those raters, and — assuming there is some overall agreement amongst all three raters — the adjusted reliability correlation will be higher than any of the correlations obtained by any single pairing of raters.

Inter-rater reliability for the whole data set was .853 for the first rating procedure and .90 for the second. Tables 5.5 and 5.6 show the correlations between raters when the data set was split into control and experimental compositions.

Table 5.5 Inter-rater correlations for control and experimental compositions: first overall quality rating

pair of raters	1 and 2	1 and 3	2 and 3	overall reliability correlation
control compositions *N = 190*	.683	.733	.748	.885
experimental compositions *N = 202*	.508	.619	.582	.799

All correlations are significant at p < .001

Table 5.6 Inter-rater correlations for control and experimental compositions: second range of overall quality scores

pair of raters	1 and 2	1 and 3	2 and 3	overall reliability correlation
control compositions *N = 190*	.777	.806	.815	.923
experimental compositions *N = 202*	.641	.723	.704	.870

All correlations are significant at p < .001

The fact that correlations for the experimental data were consistently lower than equivalent correlations for the control data suggested that this may not have been a random effect. Nor were these differences negligible. In the first set of ratings for *overall quality*, for example, raters 1 and 2 obtained a correlation of .683 for the control compositions, but of only .508 for the experimental compositions. Raters 2 and 3 obtained a correlation of .748 for control compositions, but of .582 for experimental compositions. Since correlations may only range in magnitude from 0 to 1, such differences may be seen as relatively large.

A procedure for testing the statistical significance of the difference between correlation coefficients obtained by two independent groups for the same variables (Pallant, 2001) confirmed that, in each case, the difference between the inter-rater correlations for control and experimental compositions was statistically significant. This showed that the level of inter-rater agreement for the control data was *significantly* higher than for the experimental data. Z_{obs} values obtained by the procedure are given in Table 5.7.

Table 5.7 ***Z_{obs} values for differences between control and experimental data inter-rater correlations for overall quality scores***

	raters 1 & 2	raters 1 & 3	raters 2 & 3
first rating procedure	2.611	2.003	3.056
second rating procedure	2.817	2.150	2.552

N control = 190
N experimental = 202
Z_{obs} is significant when greater than or equal to 1.96

It is the case that low standard deviations depress correlations, as the magnitude of a positive correlation is more heavily influenced by items (or scores) which are further from the mean than by items which are close to the mean. The lower standard deviations observed in the distributions of the experimental data may have — technically, as an artefact of the correlation equation — *caused* the inter-rater correlations to be lower for the experimental data than for the control data. There is, however, no standard procedure to control for varying standard deviations when correlating two sets of scores, and so it is not possible to investigate *mathematically* whether these differences between the inter-rater correlations reflect *true* differences, or simply mathematically-created differences, concerning the level and nature of agreement between the raters.

A possible procedure would be to reduce the difference between the standard deviations of the control and experimental data sets by cutting off the tails of the distributions and using a more central set of scores for the correlations. This might allow for comparability of control and experimental inter-rater correlations within more similar distributions — *i.e.* for the more central scores. This would, however, at the same time depress both control *and* experimental correlations, by restricting the ranges of scores, and would discount information provided by the excluded scores.

Another approach was to investigate whether similar differences in size of correlation coefficients for control and experimental would occur in other correlations. Distributions for the six constructs revealed considerably bigger control than experimental standard deviations for three of these — *grammatical complexity*, *vocabulary range* and *coherence and flow* (Table 5.9). These differences were comparable to the difference between control and experimental *overall quality* standard deviations for the same range of possible scores (Table 5.2). As with *overall quality*, inter-rater correlations for *grammatical complexity*, *vocabulary range* and *coherence and flow* were significantly higher for control than experimental data (Tables 5.12 and 5.13). Again, this might simply be a mathematical effect of the higher standard deviations. However, correlating any of the variables *grammatical complexity*, *vocabulary range* or *coherence and flow* with *overall quality* ratings (Table 5.15) did *not* produce a significantly higher correlation coefficient for the control data set than the experimental, despite the larger standard deviations for all four of these control distributions. Thus it appears to be very unlikely that the differences between control and experimental data inter-rater reliabilities were simply mathematical artefacts of higher and lower standard deviations.

Nor could this effect have been an artefact of raters' attitudes towards either of the two groups, with possibly one rater favouring the experimental group, consequently behaving differently towards that group and lowering overall group agreement, since raters did not know if they were rating a control or an experimental composition. Furthermore, the lower correlations were evident for each of the three rater pairs, indicating that at least two of the raters must have behaved less reliably whilst judging experimental compositions, and this in different ways from each other.

It seemed that either it was harder for the raters to behave reliably when rating experimental compositions, or that, whilst each rater might still be behaving consistently as an individual rater, the experimental compositions invoked more disagreement amongst the raters as a

group. Either of these two situations might arise where a rating instrument does not suit a particular set of data. However, since the rating instrument did not indicate expected performance outcomes for different grade bands, and had no sets of descriptors — specifically to prevent any such mismatch between the data set and the rating instrument — but required only a single impressionistic judgement, in relation to the other compositions in the data set, the increased variability of scores within the experimental data must have arisen from the interaction between the raters and the compositions. There may have been features in the experimental compositions, but which were not present in the control compositions, to any one of which not all three of the raters were sensitive. Alternatively, the experimental compositions may have displayed the same features, possibly within the same ranges of proficiency, but with a different internal balance amongst these, and the raters, all experienced classroom teachers, may have been more used to the kind of text produced by the non reading-scheme students. In other words, the reading scheme may have affected the writing output of the experimental students but the raters either were not equipped to perceive any such changes, or reacted to them erratically and differently from each other.

5.4 Scores for the six constructs

The three scores for each construct (on a scale of 1 to 6) were summed to give each composition one score for each construct (Table 5.8).

Table 5.8 ***Scores on individual constructs: descriptives for whole data set (control and experimental)***

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
grammatical complexity	3	18	9.27	2.86	9	9
grammatical accuracy	3	18	8.69	3.09	9	9
vocabulary range	3	18	10.45	2.80	11	11
spelling*	3	18			15	17
punctuation & paragraphing	3	17	9.08	2.69	9	9
coherence and flow	3	18	9.97	3.05	10	11

Range of possible scores = 3 to 18
N = 392
**Mean and standard deviation are not reported for spelling, as the distribution was not normal. The median is a more appropriate measure of central tendency.*

Grammatical complexity, *grammatical accuracy*, *vocabulary range*, *punctuation and paragraphing*, and *coherence and flow* produced normal distribution patterns. The distribution of *grammatical accuracy* scores was very slightly positively skewed (which is to say that there was a slight tendency on the part of the raters to award lower scores), and the distribution of *coherence and flow* scores was very slightly negatively skewed. Both of these distributions, however, could be classed as normal. Bar charts for all six distributions may be found in Appendix 8.

Spelling scores produced a J-shaped distribution, with only 11.5% of the scores in the lower half of the range of possible scores, and 57.1% of the scores being amongst the top four possible scores (in a range of 16 possible scores). Raters reported after the rating process that they had, against their initial expectations, found it difficult to compare the spelling of the compositions. Whilst a few compositions at the lower end of the scale contained a high proportion of spelling mistakes and required the use of the lowest possible score of 1, the majority of compositions had few spelling mistakes, and a considerable number had none at all. Moreover, what might on first consideration seem to have been the simplest, and easiest to operationalize, of the six constructs, and consequently the easiest to judge, aroused the most rater quandary. Raters could not decide if it was worse to misspell an easy word than a more difficult word, or a technically easy to spell word, but which was rare in the corpus (for example "tank"); if it was worse to misspell two words in a short composition than in a longer composition, or than to misspell *three* words in a longer composition; if a student should be penalised for misspelling the same word multiple times, or if that should be regarded as one mistake; what to do if a student had misspelt a word on one occasion, but written it correctly on another occasion.

Although the inter-rater correlations for *spelling* scores were all significant at $p < .001$ (.440, .606 and .392, using Spearman's rho, which is a rank order correlation, as the distribution was not normal), spelling was not a very suitable construct for rater judgements with this particular data set. Spelling mistakes were so relatively scarce that raters, in an attempt to differentiate amongst the compositions, began to assess mistakes individually, rather than giving a rapid impressionistic judgement of the body of spelling mistakes as a whole. Global, impressionistic judgement of spelling may be more appropriate in data sets where spelling mistakes are more common.

Possible reasons for the scarcity of spelling mistakes in this data are, firstly, the influence of the formal L1 writing system, Chinese Putonghua characters, and secondly, the effects of an

L2 language teaching methodology which greatly discourages inaccuracy of word form. These two influences may even be related. Accuracy of form is very important in written Chinese, and this may have a psychological influence on subsequent L2 learning, particularly writing. The common pedagogical practice in Hong Kong EFL classrooms of a very strict insistence on accuracy of word form may, in fact, be a legacy of Putonghua learning methods.

The problems caused by the low numbers of spelling mistakes (as regards the distribution of scores) had not been anticipated in the first stages of monitoring inter-rater and intra-rater reliability (*cf.* Section 4.3.4.7), since the issue at that time had been rater agreement, and not the relative scarcity of low-scoring and middle-scoring compositions. At that point it had been possible to identify both a lowest-scoring and a highest-scoring composition from the initial sample of 40 scripts. The relatively higher frequency of higher-scoring scripts had been remarked upon by raters whilst identifying benchmark scripts, but was not seen, at that stage, to be an indication of any inappropriateness of the rating instrument for spelling.

Table 5.9, below, shows descriptives obtained for the six constructs when the data set was divided into control and experimental groups.

Table 5.9 ***Scores on individual constructs: descriptives for control and experimental groups***

	Group status	Minimum	Maximum	Mean	s.d.	Median	Mode
grammatical complexity	control	3	17	9.19	3.15	9	11
	experimental	4	18	9.34	2.56	9	9
grammatical accuracy	control	3	16	8.66	3.14	9	9
	experimental	3	18	8.72	3.06	9	7
vocabulary range	control	3	18	10.61	3.07	11	12
	experimental	4	18	10.30	2.52	10	11
spelling	control	5	18			15	18
	experimental	3	18			15	15
punctuation & paragraphing	control	3	17	9.09	2.67	9	9
	experimental	3	16	9.07	2.71	9	7
coherence and flow	control	3	18	9.93	3.29	10	11
	experimental	3	18	10.01	2.82	10	12

Range of possible scores = 3 to 18
N control = 190
N experimental = 202

The same pattern as was observed for *overall quality* judgements was observed for *vocabulary range*, with a slightly lower mean for the experimental group and a markedly lower standard deviation, indicating a more closely grouped distribution. The pattern differed slightly for *grammatical complexity*, *grammatical accuracy* and *coherence and flow*, in that for these it was the control group which had the lower mean, though the difference in all three cases was of less than 0.2 points. The pattern for the standard deviations of these three constructs, however, remained similar to that for *overall quality* and *vocabulary range*, which is to say it was smaller for the experimental group than for the control group, although the difference was not large for *grammatical accuracy*.

Punctuation and paragraphing was the only variable for which the control and experimental distributions seemed to be a close match, although the mode was lower for the experimental group. The mode, however, was not in this case a good indicator of central tendency. Examination of the distribution chart (Appendix 8) showed that this was a chance artefact of the data, and that the distributions differed only in that the experimental curve was slightly flatter, with kurtosis of -.649 compared to the kurtosis of -.222 observed in the control data curve.

For *spelling*, which did not have a normal distribution, the most appropriate indicator of central tendency was the median, which was 15 in both cases, indicating that approximately half of each group had scores of 15 or below. When the median itself is a common score, it is inevitable that the split into two halves will fall somewhere *amongst* the group of cases which have that score. Exact figures showed that 55.8% of control students scored 15 or below, and 57.9% of experimental students.

Independent-samples t-tests for each of the variables showed no significant differences between the two groups (Table 5.10).

Table 5.10 Comparison of control and experimental groups for the six constructs

	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow	spelling*
t-value	.525	.190	1.087	.093	.269	Z =.156*
significance level	.600	.849	.277	.926	.788	.876

N control = 190
N experimental = 202
** Because the spelling data curve was not normally distributed, a Mann-Whitney U test, comparing ranks and medians, was used.*

5.5 Inter-rater reliability for separate constructs

Overall inter-rater reliability correlations for the six constructs were reported in Section 4.3.4.8, but are included in the table below for ease of reference.

Table 5.11 Inter-rater correlations for separate constructs: whole data set

	raters 1 and 2	raters 1 and 3	raters 2 and 3	overall reliability correlation
grammatical complexity	.457	.611	.509	.772
grammatical accuracy	.530	.681	.584	.818
vocabulary range	.527	.667	.519	.806
spelling*	.440	.606	.392	.738
punctuation & paragraphing	.427	.303	.223	.585
coherence & flow	.526	.550	.452	.757

N = 392
**A Spearman rank-order correlation was used for spelling, which did not have a normal distribution*
All correlations are significant at p < .001

Grammatical accuracy showed the highest level of rater agreement, followed by *vocabulary range*. This may have been because these two constructs are in some ways more definite, hence easier to evaluate, than *complexity* and *coherence and flow*. Vocabulary and grammar are also a more common classroom focus, possibly partly *because* they are more readily definable and lend themselves more easily to instruction. This widespread teaching focus on grammar and vocabulary may have contributed to the raters' apparent lesser ability to differentiate amongst, and agree as to, varying levels of complexity and coherence. *Punctuation and paragraphing* had the lowest agreement. Over the set of six constructs, the best level of agreement was between raters 1 and 3.

The most likely reason for the comparatively low level of inter-rater agreement on *punctuation and paragraphing* was that it did not prove to be a homogeneous construct. The correct use of punctuation and an appropriate use of paragraphing did not always co-exist in

the same composition. Some compositions demonstrated correct sentence-level punctuation, but consisted of only one, or two, very long paragraphs. Other compositions demonstrated competent use of paragraphing, but the sentences within them were not correctly punctuated.

Particularly prevalent in the data was the use of run-on sentences, in which multi-clause "sentences" consisted of a series of main clauses, with or without a change of grammatical subject, separated by commas and with no capitalization to provide any evidence that the comma may have been intended as a full-stop. (One such example was "I heard a noise, it was a man, he opened the door.")

It is difficult to interpret such run-on sentences. They may derive from the influence of an L1 — in this case Chinese — which has different conventions of phrasing and punctuation from European scripts. They may also be evidence of a kind of punctuation-related interlanguage where the student wishes to connect the clauses, aiming to present them as one extended ideational unit and to tie them more closely together than would be the case if they were three separate sentences. The student may not have mastered the use of connectors, or his grammar might be as yet too simple for the task. (To take the previous example, the student may be attempting to represent "I heard the sound of a man opening the door" or, more basically, "I heard a noise, which was a man, who opened the door".) On the other hand, it may be, quite simply, that the student *has* attempted to make three correctly punctuated sentences, in a written code which he knows to have different punctuation and phrasing conventions from his L1, but failed, through inattention or misconception. Since it is impossible for raters to know what a student was attempting, they judge what they see. With raters focusing more on actually-present grammar than on possibly-intended sentence units, run-on sentences are most likely to be regarded as text where the grammar may have succeeded but the punctuation has failed, rather than as a single sentence where the grammar has failed.

Inter-rater correlation coefficients for the six constructs, for control and experimental compositions separately, are given below.

Table 5.12 ***Inter-rater correlations for separate constructs: control and experimental compositions***

	Group status	raters 1 and 2	raters 1 and 3	raters 2 and 3	overall reliability correlation
grammatical complexity	control	.569	.679	.641	.836
	experimental	.315	.541	.357	.676
grammatical accuracy	control	.547	.678	625	.830
	experimental	.513	.684	.546	.809
vocabulary range	control	.596	.709	.629	.848
	experimental	.449	.628	.373	.742
spelling*	control	.474	.698	.494	.792
	experimental	.405	.516	.284	.667
punctuation & paragraphing	control	.521	.209**	.215**	.585
	experimental	.325	.395	.235**	.585
coherence and flow	control	.583	.579	.585	.806
	experimental	.458	.524	.311	.694

N control = 190
N experimental = 202
**A Spearman rank-order correlation was used for spelling, which did not have a normal distribution*
*All correlations are significant at $p < .001$ except those marked ** which are significant at $p < .01$*

Apart from *punctuation and paragraphing*, which was the least reliable of the constructs, inter-rater reliability was uniformly lower for the experimental compositions than for the control compositions. This was a repeat of the finding for the two *overall quality* rating processes. Z_{obs} values for statistical significance of differences between control and experimental averaged inter-rater correlations are given in Table 5.13.

Table 5.13 ***Z_{obs} values for differences between control and experimental averaged inter-rater correlations for separate constructs***

	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow
Z_{obs}	3.053	0.530	2.317	0.0	1.963

N control = 190
N experimental = 202
Z_{obs} is significant if it is greater than 1.96
Z_{obs} was not calculated for spelling

The differences between the averaged inter-rater correlations for *grammatical complexity*, *vocabulary range*, and *coherence and flow* were statistically significant, confirming that these were not merely chance differences. This suggested that the underlying turbulence found between the experimental texts and the raters' evaluations — but which did not exist in the case of the control texts — may be local to these constructs. *Grammatical accuracy* appeared to be a more constant variable, whilst *punctuation and paragraphing* showed little variation across the two groups. Z_{obs} was not calculated for *spelling* — although the Spearman correlations also exhibited lower values for the experimental data set than for the control data set.

Punctuation, paragraphing and spelling are relatively trivial constructs when compared to the grammar and vocabulary of a language and the complexity and coherence of language production. These three features are confined to written representations of the language, and are not part of the underlying linguistic structure. Co-development of punctuation and spelling skills with language competence is not likely to arise from ongoing interaction with other language components, but is more likely to be the result of the incidental training and practice in these skills which co-occur with the learning of the language system. Such written-text presentation skills might be considered superficial to underlying language knowledge.

Punctuation, paragraphing and spelling may not be subject to the kind of interactive restructuring affecting the more dynamic components of the language system, but may simply increase in scope as knowledge of the language develops. This may be one reason why overall inter-rater correlation for *punctuation and paragraphing* was not affected by the control or experimental status of the compositions in the way that inter-rater correlations for *grammatical complexity*, *vocabulary range* and *coherence and flow* were. In other words, the type of input offered by the reading scheme (non form-focused) may have led to a restructuring of the language system in ways that provoked rater disagreement but did not affect punctuation skills. This is not to say that the reading scheme had no effect on such skills, nor, indeed, that it *did* have any effect, but that these skills, if they *were* affected, were not affected in such a way as to cause problems for rater agreement. Such skills may simply have developed more linearly.

Although spelling may also evolve incidentally to language development, rather than integrally, it is nonetheless linked to vocabulary growth. For example, more recently

acquired words may have less stable spelling, or new words may destabilize the spelling of other, similarly spelt or similar sounding words. The lower inter-rater correlations for *spelling* for the experimental compositions may have been a by-product of vocabulary adjustments engendered by restructuring and which may have caused more instances of those judgement contexts found problematic by the raters (as discussed in Section 5.4.)

Of the four more strictly linguistic constructs, *grammatical accuracy* alone did not exhibit any significant difference between the levels of rater agreement for control and experimental compositions. One possible reason is that grammatical accuracy is less susceptible to any effects of extensive reading than complexity, vocabulary or coherence, but is more susceptible to the effects of standard classroom instruction. (Both groups had lessons using a coursebook, although the experimental group had fewer.) That *grammatical accuracy* offers an exception to the pattern, may, in fact, strengthen the case for the real significance of the differences between the levels of rater agreement for the other three constructs.

5.6 Relationships between constructs

Table 5.14 Correlations between scores on five constructs*and overall quality scores for whole data set

grammatical accuracy	.767				
vocabulary range	.822	.767			
punctuation & paragraphing	.669	.670	.655		
coherence and flow	.845	.787	.812	.773	
overall quality	.783	.828	.771	.647	.780
	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow

N = 392
All correlations are significant at p < .001
**Spelling is not included*

The (Pearson) correlation matrix for the five constructs (*spelling* is not included since it did not have a normal distribution) shows a high level of inter-construct correlation. The lowest

correlations were obtained for *punctuation and paragraphing*, which was the only construct to obtain correlation coefficients of less than .7. This was not surprising, as it was the least reliable of the constructs, and its lower correlations with the others may have derived partly from the observed lower inter-rater reliability.

Correlations may be *disattenuated* , or corrected, for such observed unreliability of the initial measurement procedure. Indeed, *all* the correlations in the above matrix might be disattenuated in this way, to allow for the differing inter-rater reliabilities for the different constructs. However, when inter-rater reliability is lower than the observed correlations (*i.e.* between variables) which are to be corrected, the disattenuation equation tends to over-correct and to produce inflated correlations. Furthermore, in the case of *punctuation and paragraphing* it was the construct itself which was unreliable, as it did not represent a single dimension, but two features which might operate independently of each other. The lower reliability of the measurement (inter-rater correlations) was as likely to be a consequence of the inherent instability of the construct as a flaw in the measurement technique. When unreliability is inherent to the construct, disattenuation would not be an appropriate procedure.

Table 5.15 ***Correlations between scores on five constructs* and overall quality scores for split data set***

grammatical accuracy	.803 .731				
vocabulary range	.838 .805	.797 .741			
punctuation & paragraphing	.675 .671	.657 .684	.662 .657		
coherence and flow	.863 .822	.789 .787	.844 .775	.754 .800	
overall quality	.800 .759	.862 .796	.789 .748	.630 .678	.784 .778
	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow

Non-shaded areas contain control coefficients: shaded areas contain experimental coefficients
N control = 190
N experimental = 202
All correlations are significant at p < .001
** Spelling is not included*

Splitting the data into control and experimental groups revealed little difference between the two sets of correlations. Except for *punctuation and paragraphing*, all the experimental correlations were lower than the equivalent control correlations, but these differences were slight and non-significant. In *this* instance, unlike the case for the inter-rater correlations, where the differences were very much bigger, this may have been a mathematical artefact of the lower standard deviations of the experimental group and not indicative of any true differences between the strength and patterns of relationships amongst the different constructs for the two groups.

To investigate further, a standard multiple regression procedure was performed using *overall quality* ratings as the dependent variable and the four linguistic constructs of *grammatical complexity*, *grammatical accuracy*, *vocabulary range* and *coherence and flow* as independent variables. *Spelling* could not be used, as the distribution did not meet the requirements for the statistical procedure. Moreover, it is a feature of multiple regression that reliability and generalizability of results are very dependent upon the ratio of independent variables to sample size. (The fewer the independent variables and the larger the sample, the better.) For this reason, *punctuation and paragraphing* was also excluded from the model, as it was the least interesting of the five normally distributed variables in that it was the least theoretically liable to interact with other variables and not, properly, a linguistic construct.

Table 5.16 ***Beta coefficients for standard multiple regression: control and experimental compositions: dependent variable = overall quality***

	Control compositions		Experimental compositions	
	beta coefficient	significance of contribution	beta coefficient	significance of contribution
grammatical complexity	.167	.030	.186	.014
grammatical accuracy	.546	.000	.389	.000
vocabulary range	.137	.056	.157	.025
coherence and flow	.095	.214	.197	.011
R^2	.782*	.000	.721*	.000

N control = 190
N experimental = 202
** R^2 values are not beta coefficients*

The beta coefficients provided by standard multiple regression indicate the size of the *unique* contribution of each of the independent variables. (Unlike R^2, these cannot be rewritten as percentages, and can best be judged by comparing against each other.) The unique contribution of any independent variable is that part of the dependent variable — in this case, *overall quality* judgements — which may be attributed solely to that particular independent variable, independently of all other variables. Any contribution made to the dependent variable by the *combined* effect, or overlap, of more than one variable is not taken into account. For example, the beta coefficients for *grammatical complexity* were .167 for the control compositions and .186 for the experimental compositions. These values did not include the potential contribution to the dependent variable of any combination effects caused by *grammatical complexity* and *grammatical accuracy*, *grammatical complexity* and *vocabulary range*, or *grammatical complexity* and *coherence and flow*. R^2, conversely, represents the combined contribution of all the variables, including all overlap and all unique contributions.

The four linguistic constructs together explained 78.2% and 72.1% (R^2 converted) of the variance in *overall quality* for the control and experimental compositions, respectively. That the model accounted for a lower percentage of the total variance in the experimental compositions was consistent with the view that there was at least one dimension in the experimental compositions which acted outside the other variables and which the raters may have perceived — since it was present in the *overall quality* ratings — but could not categorize and had more difficulty in perceiving consistently.

Grammatical accuracy had the largest beta value for both control and experimental data. This does not necessarily mean that *grammatical accuracy* made a greater overall contribution to *overall quality* than any of the other variables, but that, acting by itself, and not in combination with any other variable, it made a greater *unique* contribution. This may have been because accuracy did not act so well in tandem with other variables, but, conversely, operated better as a predictor on its own account than did *grammatical complexity*, *vocabulary range* or *coherence and flow*.

All four variables made a significant unique contribution ($p < .05$) to the *overall quality* judgements for the experimental compositions. Although *grammatical accuracy* made the strongest unique contribution, the contributions of the other three variables were evenly matched. This was not the case for the control compositions, where neither *coherence and*

flow nor *vocabulary range* made a significant unique contribution. *Coherence and flow* in particular made a very low unique contribution. *Grammatical accuracy*, on the other hand, made a considerably greater contribution than was the case for the experimental compositions.

These differences in the relative unique contributions of accuracy and coherence cannot simply be explained by assuming that the control group may have exhibited higher levels of accuracy whilst the experimental group may have exhibited greater levels of coherence, causing the raters to rely on these two constructs to a greater or lesser degree, for the two groups respectively, when giving *overall quality* ratings. Referring to Table 5.9, we may see that the mean ratings for *both* of these constructs were marginally higher for the experimental group than for the control group (with a difference between the means of 0.06 points for *grammatical accuracy* and 0.08 points for *coherence and flow*).

The difference between the two standard regression models could not, then, be accounted for by any better performance by one of the groups on any one of the independent variables. Nor could the two models be seen as representing two different proficiency levels, since the difference between the means for *overall quality* for the two groups was 0.04 points (in favour of the control group), suggesting a very similar overall proficiency. These differences between the models might thus be taken as further evidence of a difference in internal linguistic texture between the control and experimental compositions, brought about by the reading scheme, which, whilst causing more rater disagreement amongst the experimental compositions, did not cause one group to be rated more highly than the other.

5.7 Comparisons between schools

Since language learning is a multi-dimensional process, and does not simply progress in linear fashion, but is developmental and interactive, any language teaching method may have different effects at different stages and upon different proficiency levels.

The data in the present study originated from four schools previously designated (by the Hong Kong Ministry of Education) as being at different levels. Moreover, one school provided data from Secondary 2 students whereas the other schools provided data from Secondary 3 students (*cf.* Table 4.1). Whilst the inclusion of a range of levels in a study may

increase the generalizability of any results obtained, it may also have a weakening effect on differences between groups arising from a treatment.

If an effect is local to one proficiency level, the strength of the effect may be diluted by taking the sample as a whole. The effect may even become invisible, particularly if there are competing effects at different levels, which may cancel each other out. For example, at one language learning stage a particular treatment might cause an increase in grammatical accuracy accompanied by an increase in sentence complexity. At a different level, the same treatment might cause an increase in grammatical accuracy accompanied by a *de*crease in sentence complexity. Looking only at the mean scores for grammatical accuracy and sentence complexity for the group as a whole (*i.e.* the two proficiency levels combined) could lead to the conclusion that the treatment resulted in an increase in accuracy but had no effect on sentence complexity. Thus the differing effects of the treatment on sentence complexity at different proficiency stages would have been cancelled out by each other.

Splitting the data into groups, by school, and ranking the four schools according to their mean scores for *overall quality* (Table 5.17) resulted in the same rank ordering as the Ministry of Education's band system, with the school designated as "very high" level achieving the highest mean score, and the school designated as "low" level achieving the lowest mean score. Of the two schools in the middle, both designated "high" level, one school had provided students from Secondary 3 (as had the highest and lowest scoring schools) and one school had provided students from Secondary 2. The school providing Secondary 2 students had a lower mean score than the school providing Secondary 3 students. However the "high" Secondary 2 students still achieved a mean score which was higher than that of the Secondary 3 students from the "low" school.

Table 5.17 Overall quality means for four schools

School	Ministry Band	Mean	Standard deviation	Year (Secondary)	N students
1	Very High	12.53	2.25	3	76
2	High	12.05	2.40	3	82
3	High	11.00	2.51	2	84
4	Low	7.92	2.54	3	150

Range of possible scores = 3 to 18

A two-way between-groups analysis of variance was conducted to investigate differences between schools and whether the reading scheme impacted differently on these. *Overall quality* scores constituted the dependent variable with *school* and *status* (*i.e.* control or experimental) as two independent variables. Whilst there was no significant effect for *status* (p = .152), demonstrating that, as has already been seen, participation in the reading scheme did not, by itself, produce an effect, there were statistically significant effects for both *school* (p = .000) and the interaction between *school* and *status* (p = .001).

These results showed that there were real differences between the schools — supporting the Ministry of Education's band system and at the same time confirming the ability of the rating procedure in this study to discriminate between levels — and that the reading scheme did, in fact, impact differently on different schools, which may have meant, in practice, at different levels. Eta squared for *school* was .404, showing a very large effect, indicating that there were big differences between the schools, regardless of the reading scheme, and .043 for interaction between *school* and *status*, showing a modest effect.

Tukey's Honestly Significant Difference test provided the further information given in Table 5.18.

Table 5.18 Mean differences between schools for overall quality ratings

	school 2 *N = 82*		school 3 *N = 84*		school 4 *N = 150*	
	mean difference	sig	mean difference	sig	mean difference	sig
school 1 *N = 76*	.48	.601	1.53	.000	4.61	.000
school 2 *N = 82*			1.05	.028	4.13	.000
school 3 *N = 84*					3.08	.000

Score range = 3 to 18

There was no significant difference between schools 1 (very high level) and 2 (high level), although the higher level school did present a slightly higher mean. This may have been because the difference between "high" and "very high" ability was not very great. The significant differences between schools 1 and 3 (also high level) and between schools 2 and

3 could be accounted for by the fact that school 3 had used its Secondary 2 students to provide the data, whilst the other schools had used Secondary 3 students. There were significant differences between school 4, which was markedly the weakest, and each of the other schools.

In order to locate the interaction effect between reading scheme and school revealed by the two-way between-groups analysis of variance, an analysis of simple effects was conducted for each of the schools. The results of independent-samples t-tests are given in Tables 5.19 to 5.22.

Table 5.19 ***Comparison between control and experimental students for overall quality ratings: school 1 (very high ability; Secondary 3)***

	N students	mean	standard deviation	t-value	significance
control	33	12.58	2.35	.166	.868
experimental	43	12.49	2.19		

Range of possible scores = 3 to 18

Table 5.20 ***Comparison between control and experimental students for overall quality ratings: school 2 (high ability; Secondary 3)***

	N students	mean	standard deviation	t-value	significance
control	41	12.73	2.56	2.663	.009
experimental	41	11.73	2.04		

Range of possible scores = 3 to 18

Table 5.21 ***Comparison between control and experimental students for overall quality ratings: school 3 (high ability; Secondary 2)***

	N students	mean	standard deviation	t-value	significance
control	42	11.52	2.61	1.942	.056
experimental	42	10.48	2.31		

Range of possible scores = 3 to 18

Table 5.22 ***Comparison between control and experimental students for overall quality ratings: school 4 (low ability; Secondary 3)***

	N students	mean	standard deviation	t-value	significance
control	74	7.39	2.63	2.551	.012
experimental	76	8.43	2.36		

Range of possible scores = 3 to 18

Standard deviations for the experimental classes were uniformly lower than those for the equivalent control classes, showing that this effect was common to all levels, and suggesting that, at whatever proficiency level it was operating, the reading scheme had a slight homogenizing effect on the students.

Schools 2 and 4 showed significant differences between the control and experimental groups ($p < .05$) whilst schools 1 and 3 did not. In the case of school 4, the lowest level school, this difference was in favour of the experimental group. In the case of (high level) school 2, however, the difference was in favour of the control group. An initial interpretation might be that the reading scheme had had a positive effect on language gain at the lower level, but a negative, or detrimental, effect at a higher level, possibly because at the higher level normal classroom instruction was more effective than extensive reading in promoting language improvement. Use of the ERS as part of the English curriculum meant that a class had two, or three, fewer lessons (out of eight to ten, depending on the school) of coursebook-based instruction per week, as this lesson time was used for the reading scheme instead.

Why the better performance of control students was not equally the case for all three high level schools was not clear. School 3 was different from the other two high level schools in that, as noted earlier, it had used Secondary 2 students to provide data. This may have meant that coursebook-based instruction was more effective for the higher Secondary 3 class, but at Secondary 2 the two instruction methods — coursebook-based instruction and extensive reading practice — did not differ significantly in their effects. This would not, however, explain why the highest level school (school 1), which did not differ significantly from school 2, did not produce similar results. It was possible that the evaluation procedure had not been reliable for one or the other of the two schools, or that some unknown variable had entered the equation at the level of either one of these.

All four schools streamed their students into a highest ability class and progressively lower ability classes. Subsequent investigation into the provenance of the data revealed that school 2 had used its highest level class to provide the control compositions, but had not used the equivalent class of the following year to provide the experimental compositions. Instead, it had used a class which was not even adjacent in level, but was separated from the control class by an intervening class. In other words, control data had been supplied from the highest class and experimental data had been supplied from the *third* highest class. Thus the two groups had not been matched from the outset, and, not only did the control group have a previously proven higher ability, but they were less likely to have been subjected to the same treatment, both in their English lessons and in other subject lessons, than two classes of the same status within the school. They could, for instance, have had the best teachers, or used different coursebooks, or have been more strictly encouraged to use English in their other subject classes. They could also have had a more elite family background, with increased exposure to English outside school.

It also emerged that school 1 had collected the data wrongly. Control data had been supplied from the highest class — an elite class of 33 students — and experimental data had come from the second highest class — a class of 43 students. Similar hypotheses to those made concerning school 2, about differing treatments for classes of different status within the school, are given weight by such an obvious difference in class size.

The reasons for this mistake in the collection of data were not discovered. The schools may not have been aware of the gravity of such an error. It is possible that, in the first year of data collection (*i.e.* the control data) by a government body, the schools had used the opportunity to put their best classes forward. (Schools in Hong Kong at that time feared being judged and were anxious to maintain their position in the bands system.) When the second batch of data (*i.e.* experimental data) was requested one year later, it may have been brought home better to the schools that *they* were not under scrutiny, but that the data collection was part of an overall evaluation of the reading scheme and that individual schools would not be named. They may then have been more prepared simply to use the class which was most convenient, regardless of instructions. Alternatively, the mistake may have arisen from some miscommunication, or even a clerical error.

For whatever reason, the data from schools 1 and 2 was not just *possibly* contaminated by differing circumstantial treatments for control and experimental groups, it actually flouted the experimental design of the study by the use of higher ability students as control groups and lower ability students as experimental groups. With no other set of data available as to students' ability in English immediately *prior* to entry into Secondary 1 — and into, or not into, the reading scheme — it was impossible to control for this difference in abilities at the start of the project.

It might be argued that for the very high level school, with very little difference between the highest level (control) class and the lower level (experimental) class ($t = .166$; $p = .868$), the reading scheme had "raised" the lower class to the same standard as the highest class. Since it cannot be known, however, what original difference, if any, might have existed between these two groups at the outset, this can only be conjecture. For school 2, the difference between the highest class and the third class remained significant, despite the reading scheme. All that can be said in this instance is that any effects of the reading scheme were not powerful enough to raise the third class to the standard of the highest class. It might also be said, however, that the difference in status between the top class and the third class was liable to be even greater than the difference between the first and second classes of the very high level school, and that it was very likely that the two classes had received quite different treatments within the school, which may have compounded the original differences in ability.

The data from schools 1 and 2 was shown to be invalid for use in comparing control and experimental performance levels, and therefore invalidated the conclusion (drawn from the results of the significance tests reported above) that the reading scheme had had no effect on the experimental students' compositions, as regards either *overall quality* or performance on individual constructs. The investigations into inter-rater agreement, however, may stand. Unlike the case for significance testing, the inter-rater correlations would not be affected by whether high and low scores in the data set derived from control or experimental compositions. In any case, extrapolating from the t-test results, the mis-collection of data appears to have resulted in approximately equal numbers of high and low scores for control and experimental, discrediting any contention that raters may have found it easier to agree at higher levels than at lower levels, or vice versa, and that the differences in levels of inter-rater agreement for the control and experimental compositions might be an artefact of this.

Thus the differences observed between levels of inter-rater agreement for control and experimental compositions remained valid.

The exploration of the strength and variability of relationships between constructs was also unaffected. The standard multiple regression would, likewise, not have been affected, and the conclusions drawn from that procedure also remained valid. The inclusion of the improperly collected data in these calculations could only have increased the reliability of results by increasing the size of the data set, which would have been substantially reduced had data from the two schools been excluded at that stage.

The observation regarding the smaller standard deviations for the experimental distributions might also still stand, as this effect did not seem to be influenced by the magnitude of the means, nor whether the control group had a lower or a higher mean than the experimental group, but was consistently a property of the experimental distributions. This would suggest it was properly an effect of participation in the reading scheme, and not of level.

It was useful to have examined the larger data set, including the data from schools 1 and 2, since this data set — almost certainly because of the use of the highest level classes for control groups and lower level classes for experimental groups by these two schools — ultimately provided two groups of the same measured ability. Observed differences between the variabilities of the two groups (*i.e.* distribution variability and inter-rater variability for *overall quality* and separates constructs) were thus shown to be properly effects of the difference in treatment between the two groups (*i.e.* reading-scheme or non reading-scheme), and not merely level effects, linked to different stages of proficiency, and which might be expected to occur at given levels regardless of the language learning methods and environment. Additionally, that the results from the rating procedure had indicated a clear difference in level between the control and experimental groups of school 2, in a direction not expected, but subsequently confirmed by the finding that these classes *did* differ in the direction indicated by the results, was further evidence that the rating procedure was a valid and reliable discriminator.

Of the two schools which *had* provided equivalent control and experimental classes, only one showed a significant difference between control and experimental groups. This was the lowest level school, school 4, which had provided the most students. The reading scheme had been in operation in this school for nearly three years, as opposed to nearly two years in

school 3. The larger number of students, and the greater length of participation in the reading scheme, made this school technically more reliable than the other school.

Additionally, I had visited both these schools during the first years of their participation in the ERS. School 4 had shown a more positive attitude towards the reading scheme, and an acceptance of it as part of the curriculum, which school 3 had not, some English teachers viewing it more as an inconvenience imposed upon them. (Although participation by schools in the reading scheme was voluntary, this decision may have been taken without consulting the English teachers.)

Although school 4 was the lowest of the four schools which provided the data for this study, it was not one of the lowest in Hong Kong. Most of the schools which had joined the reading scheme in the first wave had been the elite, band one, schools. With limited places available, only 20 schools at a time were admitted into the scheme, because of the cost of the books and the necessary training of the teachers, and, with more schools applying for the scheme than there were places, the Ministry was able to be selective. Thus, in the first few years of the reading scheme, participation was seen as an achievement in itself on the part of a school. Moreover, only schools confident enough to present themselves for official scrutiny applied to join the scheme. The lower band schools did not wish to provoke any such official scrutiny. School 4 was a weak school within a select group. It was, in fact, a more typical school than the others, particularly schools 1 and 2, which were top Hong Kong schools.

School 4 therefore can be said to be the best school to use for the more detailed part of this study. Not only had the data been properly collected, but the school represented a more typical school than the others, where there might be fewer influences on students of the kind engendered by the elite socio-economic family background prevalent in the top Hong Kong schools. For example, many of the better-off families had English-speaking Filipino maids or nannies, or the children might have better access to English language books, videos and computer games. Additionally, school 4 had provided four classes: a control and experimental class at each of two levels. This allowed comparison of effects across levels within the same school, minimizing possible effects of uncontrolled contaminating variables between schools. Class sizes were also matched across control and experimental. Lastly, I had personally visited the school on a number of occasions and found the staff to be generally positive about the reading scheme.

5.8 Summary of Chapter Five

Data from four schools (N students = 392) was investigated for differences between control (N = 190) and experimental (N = 202) students on *overall quality*, *grammatical complexity*, *grammatical accuracy*, *vocabulary range*, *spelling*, *punctuation and paragraphing*, and *coherence and flow*. No significant differences in performance level between the two groups were found.

Investigations into the variability of the data using correlational procedures, however, revealed differences in distribution patterns and rater assessment patterns for the two groups. Standard deviations for the experimental group were consistently lower than for the control group, indicating a more homogeneous range of ability within the experimental classes. Inter-rater agreement was significantly lower for the experimental group than for the control group for *overall quality*, *grammatical complexity*, *vocabulary range*, and *coherence and flow*, indicating that raters found it harder to judge the experimental compositions consistently. Results from a standard multiple regression procedure suggested that raters perceived the experimental texts differently from the control texts, and formed their judgements on different principles.

Two of the schools were found to have provided improper data, invalidating the previous finding that the reading scheme had caused no significant differences between control and experimental students on performance level. The lowest level school was selected as the most appropriate for further investigation, because of its technical reliability. Coincidentally, this school was also a better representative of a normal Hong Kong school. The results of this further investigation are presented in Chapter 6.

6. INVESTIGATION OF SCHOOL FOUR

Score distribution patterns and rater assessment patterns for school 4 were examined, to ascertain whether these were similar to those found in the larger group of all four schools, (as discussed in Chapter 5). Similar patterns emerged, with few, and only slight, differences.

The major differences in distribution arose from the fact that this was the lowest level school of the original group of four schools. Means for *overall quality* and for separate constructs were uniformly lower, and the ranges of actual scores achieved (out of possible scores) were slightly restricted, in that the maximum scores of the ranges of possible scores were not achieved. Standard deviations were correspondingly smaller.

6.1 First range of overall quality scores

The distribution of *overall quality* scores for school 4 was flatter than that for the complete data set (4 schools), with kurtosis of -.859 (compared to .023) and a slight negative skewness of -.248 (compared to -.055). This matches what would be expected from a lower scoring group extrapolated from a larger group with a broader range of abilities.

Negative skewness indicates a tendency of scores towards the higher end of a scale. In this case, the tendency was not towards the top scores of the original range of scores, since these scores had effectively been removed, but towards the central scores of that range (which constituted the bulge of the normal curve of the original distribution). This tendency, caused by the rating procedure itself, towards the original central scores — the values of these scores deriving from the broader range of abilities used in the establishment of the original six bands — now manifested itself in a tendency towards the higher scores in the new truncated range of scores. In other words, the middle scores of the original distribution became the higher scores of school 4's distribution, carrying with them a higher frequency of occurrence.

The distribution for school 4 did not, however, violate the assumptions of normality necessary for the use of parametric procedures, and the use of these to investigate distribution and variability remained appropriate. (The distribution chart may be found in Appendix 9.)

Table 6.1 Overall quality scores: descriptives for school 4

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
overall quality score *N = 150*	3	13	7.92	2.55	8	11

Range of possible scores = 3 to 18
Range of actual scores = 3 to 13

Whilst the original distribution for all four schools presented a mean of 10.34 and a standard deviation of 3.14 (*cf.* Table 5.1), school 4 presented a mean of 7.92 and a standard deviation of 2.55. Splitting the group into control and experimental gave the descriptives in Table 6.2.

Table 6.2 Overall quality scores: descriptives for school 4; control and experimental groups

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
control group *N = 74*	3	12	7.39	2.63	7.5	9
experimental group *N = 76*	3	13	8.43	2.36	9	11

Range of possible scores = 3 to 18
Range of actual scores = 3 to 13

All three measures of central tendency were higher for the experimental group. The higher median indicated that, whereas approximately half the control students had scores higher than (a theoretical) 7.5, approximately half the experimental students had scores higher than 9. Actual figures showed that 37 control students (out of 74) had scores of 8 or above, whereas 49 (out of 76) experimental students had scores of 8 or above. As we saw in Chapter 5 (Table 5.22), the t-value from an independent-samples t-test for the difference between the two means was 2.551 ($p = .012$), showing a statistically significant difference between the groups.

The standard deviation for the experimental group was slightly lower than that for the control group, indicating a tighter grouping of scores, and consequently a more homogeneous group. This was consistent with the pattern already noted for the larger group

of all four schools, although the effect was smaller, possibly partly because of the diminished standard deviations (consistent with the reduced range of scores) for both groups.

6.2 Second range of overall quality scores

Table 6.3 Second range of overall quality scores: descriptives for school 4

	Minimum	Maximum	Mean	Std. Deviation	Median
overall quality score *N = 150*	3	47	26.30	10.45	28

Range of possible scores = 3 to 72
Range of actual scores = 3 to 47

When the range of scores was stretched to 45 (3 to 47), following the application of the second rating procedure, the distribution became slightly more skewed (-.280, compared to -.248). This was because the raters awarded a greater number of the lower within-band scores of 1 and 2 (63.1% across all three raters) than of the higher within-band scores of 3 and 4. As this was common to all the original bands, the effect was to extend the tail of the distribution at the low end, by stretching out the lowest scores, whilst compacting the tail at the high end, by not stretching the highest scores in the upward direction to the same extent. This inequality of the effect of the second rating procedure on the two tails of the distribution, stretching one tail but not the other, resulted in the slightly enhanced skew.

As had happened with the larger data set, and as an artefact of the way in which new scores were achieved, (discussed in Section 5.2), the distribution for the new range of scores became bimodal. However, the flatter curve of this smaller number of cases (150 compared to 392) permitted within-band modes to become whole-distribution modes more easily. This effect was exacerbated when the data set was split into control and experimental groups, with group sizes of only 74 and 76 respectively, for a range of 45 scores. Both distributions, although still within the range of acceptable normality, became flatter and multimodal. In the case of the control data, a group of four identical scores sufficed to create a mode. These modes were not reliable indicators of the real distribution patterns, and are therefore not reported here.

Table 6.4 ***Second range of overall quality scores: school 4; descriptives for control and experimental groups***

	Minimum	Maximum	Mean	Std. Deviation	Median
control group *N = 74*	3	44	24.16	10.97	25.5
experimental group *N = 76*	3	47	28.38	9.53	29.5

Range of possible scores = 3 to 72
Range of actual scores = 3 to 47

As with the larger data set, the second range of *overall quality* scores, originally intended to enhance any differences between the control and experimental groups, offered no information not already obtained from the first range of *overall quality* scores. The standard deviation for the experimental group was again slightly lower, and the mean and median higher.

As might be inferred from the lower mean and larger standard deviation, the control curve was flatter than the experimental curve, with kurtosis of -.964 (*i.e.* .964 less than the value of 3 found in the perfectly normal curve), compared to kurtosis of -.428 for the experimental curve. The t-value from an independent-samples t-test for the difference between means was $t = 2.515$ ($p = .013$), which was very similar to that obtained by the first range of *overall quality* scores ($t = 2.551$; $p = .012$). The first rating procedure, as noted for the larger group of four schools, had favoured neither the control nor experimental students. Thus the second rating procedure did not impact differently at lower levels of performance than it did overall.

Again, however, the second rating procedure, whilst not benefiting one group above the other, did allow the raters to factor into the *overall quality* judgements a slightly higher percentage of the five independent constructs of *grammatical complexity*, *grammatical accuracy*, *vocabulary range*, *punctuation and paragraphing*, and *coherence and flow*, in exactly the same way as with the larger group (*cf.* Section 5.2). Standard multiple regression procedures revealed that these five variables accounted for 64.2% ($R^2 = .642$) of the first range of *overall quality* scores and 69.4% ($R^2 = .694$) of the second range of *overall quality* scores. This represented a gain of 5.2%, compared to the 3.9% gain achieved by the larger data set.

The variance in any of these sets of *overall quality* scores *not* accounted for by the five constructs (*e.g.* 35.8% and 30.6% of school 4's two ranges of *overall quality* scores) must be accounted for by a combination of unknown contributors and by error of measurement variance. Since the larger data set obtained stronger inter-rater correlations (*cf.* Section 6.3), there may have been less variance due to error in its first range of *overall quality* scores than in that of school 4. This hypothesis is borne out by the lower R^2 value of the standard multiple regression for the first rating procedure for school 4. For the larger data set this was .753, accounting for 11.1% more of the variance than did the R^2 (.642) for school 4. This, in turn, might be why the second rating procedure had a slightly greater effect on R^2 (in relation to R^2 of the first rating procedure) for school 4 than for the larger data set. Put another way, there was more error to redress in the former case, and so the second rating procedure had more scope to redress it. The difference, however, between the effects of the second rating procedure on each of the two data sets was relatively subtle, amounting only to a 1.3% difference in reduction of unaccounted variance.

6.3 Inter-rater reliability for overall quality scores

Inter-rater reliability was slightly lower for the smaller data set than for the four schools together. The first rating procedure obtained an inter-rater reliability coefficient of .805, using Pearson product-moment correlation coefficients with a Spearman-Brown Prophecy Formula adjustment, as compared to .853 for the larger group. It is probable that the lower level of inter-rater reliability derived from the combination of a technical effect of the reduced range of scores and the effect of the smaller sample, which would tend to magnify error. Again, the second rating procedure enhanced inter-rater agreement, raising the inter-rater reliability coefficient to .864 (compared to .9 for the larger group).

6.3.1 Inter-rater reliability for control and experimental groups

When inter-rater reliability was investigated for control and experimental data separately (Tables 6.5 and 6.6), the same pattern was observed for school 4 as for the larger data set. Inter-rater correlation coefficients were uniformly lower for the experimental correlations than for equivalent control correlations, indicating that raters, as they had also done for the larger data set, found it harder to rate the experimental compositions consistently with each other. (Suggested reasons for this were proposed in Section 5.3.)

Table 6.5 ***Inter-rater correlations for control and experimental compositions: first overall quality rating: school 4***

pair of raters	1 and 2	1 and 3	2 and 3	overall reliability coefficient
control compositions *N = 74*	.648	.583	.705	.848
experimental compositions *N = 76*	.435	.516	.508	.742

All correlations are significant at p < .001

Table 6.6 ***Inter-rater correlations for control and experimental compositions: second range of overall quality scores: school 4***

pair of raters	1 and 2	1 and 3	2 and 3	overall reliability coefficient
control compositions *N =74*	.736	.725	.774	.897
experimental compositions *N = 76*	.536	.621	.594	.809

All correlations are significant at p < .001

Z_{obs} values were calculated for the differences between equivalent correlation coefficients (Table 6.7). In only two cases did these reach significance. The observed differences between equivalent control and experimental inter-rater correlation coefficients were, however, very similar to the differences observed for the larger data set (*cf.* Tables 5.5, 5.6 and 5.7). In the case of school 4, these were prevented from reaching significance by the smaller sample size. (The equation for Z_{obs} is extremely sensitive to sample size.) Although the differences between control and experimental averaged inter-rater correlations for school 4 failed to reach significance for either of the two rating procedures, the phenomenon of consistently lower inter-rater correlations for the experimental data was nonetheless clear.

Table 6.7 ***Z_{obs} values for differences between control and experimental inter-rater correlations for overall quality: school 4***

	raters 1 & 2	raters 1 & 3	raters 2 & 3	averaged correlations
first rating procedure	1.853	0.599	1.883	1.446
second rating procedure	2.057	1.157	2.087	1.769

N control = 74
N experimental = 76
Z_{obs} is significant when greater than or equal to 1.96

6.4 Scores for the six constructs

All constructs displayed a normal distribution except *spelling* which, as in the larger data set, was severely negatively skewed. (Distributions are to be found in Appendix 10.) *Spelling* was also the only construct where the highest possible score of 18 was achieved. Although *spelling* scores were distributed throughout the range of 16 possible scores, 51.3% of the scores achieved were 14 or over. Medians and modes were identical for all six constructs (Table 6.8).

Table 6.8 ***Scores on individual constructs: descriptives for school 4***

	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
grammatical complexity	3	13	7.18	2.32	7	7
grammatical accuracy	3	12	6.09	2.14	6	6
vocabulary range	3	15	8.17	2.20	8	8
spelling*	3	18			14	14
punctuation & paragraphing	3	13	7.27	2.06	7	7
coherence and flow	3	14	7.65	2.60	8	8

N = 150
Range of possible scores = 3 to 18
**Mean and standard deviation are not reported for spelling, as the distribution was not normal. The median is a more appropriate measure of central tendency.*

Splitting the data set revealed that means for the four language-related constructs of *grammatical complexity*, *grammatical accuracy*, *vocabulary range* and *coherence and flow* were all higher for the experimental group (Table 6.9). Standard deviations were slightly lower. *Punctuation and paragraphing* — not inherent to the whole language system, but related only to conventions of written presentation — showed a negligibly higher mean for the experimental group, and a larger standard deviation. *Spelling* did not present a normal distribution, and so the median was the best measure of central tendency. With reference to the lower of the two medians, whilst 53.4% of the control *spelling* scores were 13 or above, 63.8% of the experimental scores were 13 or above.

Table 6.9 ***Scores on individual constructs: descriptives for control and experimental groups: school 4***

	Group status	Minimum	Maximum	Mean	Std. Deviation	Median	Mode
grammatical complexity	control	3	12	6.70	2.41	6.50	5
	experimental	4	13	7.64	2.10	7.50	9
grammatical accuracy	control	3	12	5.97	2.31	6	3
	experimental	3	11	6.20	1.98	6	6
vocabulary range	control	3	12	7.89	2.20	8	8
	experimental	4	15	8.43	2.17	8	8
spelling	control	5	18			13	14
	experimental	3	18			14	15
punctuation & paragraphing	control	3	11	7.26	1.98	7	7
	experimental	3	13	7.29	2.14	7	7
coherence and flow	control	3	14	7.20	2.63	7	8
	experimental	3	14	8.09	2.51	8	7&9

N control = 74
N experimental = 76
Range of possible scores = 3 to 18

Independent-samples t-tests were performed for the five normally distributed variables (Table 6.10). A Mann-Whitney U test, which compares ranks and medians, was used to compare performance on *spelling*.

Table 6.10 Comparison of control and experimental groups for the six constructs: school 4

	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow	spelling*
t-value	2.531	.639	1.516	.097	2.116	Z = 1.049*
significance level	.011	.524	.132	.923	.036	.294

N control = 74
N experimental = 76
**A Mann-Whitney U test was used for spelling*

A significant difference was observed between the two groups for *grammatical complexity* and *coherence and flow*. The least effect was that observed for *punctuation and paragraphing*. Participation in the reading scheme seemed not to have influenced performance in this latter category. This construct was, in fact, the least stable of the five normally distributed variables, (as noted in Section 5.5). The lack of any clear results may have been, in part at least, a function of the lesser reliability of the construct itself. It may also, however, have been that extensive reading had no effect on punctuation or paragraphing skills.

Reading appeared to have had extremely little effect on *grammatical accuracy*, and only a non-significant effect on *vocabulary range*. This might point to the initial conclusion that the reading scheme did not benefit these aspects of language learning to any greater extent than normal classroom instruction. There was also, however, the possibility that extensive reading did have some effect on grammatical accuracy, but this was masked by the observed gain in grammatical complexity. New and more complex structures may be subject to error in the first stages of their acquisition. *Grammatical accuracy* scores were based on a rapid, impressionistic assessment of the compositions, rather than on more detailed analysis of individual error types. Raters may not have distinguished between grammatical error at a basic level and a higher level, but may have simply judged each composition for accuracy at its own level. However, students using more sophisticated structures may have achieved

improved grammatical accuracy in lower level structures which was not taken into account by raters focusing on errors arising in attempts at higher level structures.

6.5 Comparison at two levels within school 4

School 4, out of all four schools, was alone in providing data from four classes. (Other schools had provided data from two classes.) Whilst all four classes were at the same educational stage, nearing the end of Secondary 3, one pair of matched control and experimental classes comprised the top classes of two consecutive years, and the other pair comprised the third highest classes of those same years.

To investigate whether the observed effects were stable across the two different levels within the school, separate significance tests, for each construct, were performed for each pair of experimentally matched classes (Tables 6.11 and 6.12).

Table 6.11 ***Results of independent-samples t-tests for control and experimental groups for overall quality and individual constructs: school 4; lower level***

Classes 1 & 2 (lower level)	Group status	mean	standard deviation	t-value	significance level
overall quality	control	5.94	2.43	2.403	.019
	experimental	7.63	2.40		
grammatical complexity	control	5.32	1.60	3.291	.002
	experimental	6.85	2.15		
grammatical accuracy	control	4.59	1.70	1.916	.060
	experimental	5.48	2.10		
vocabulary range	control	6.91	2.09	1.428	.158
	experimental	7.67	2.23		
spelling	control	12*		1.513**	.130
	experimental	13*			
punctuation & paragraphing	control	6.35	1.84	.214	.831
	experimental	6.45	2.04		
coherence and flow	control	5.71	1.69	3.035	.003
	experimental	7.63	2.69		

N control = 34
N experimental = 33
Range of possible scores = 3 to 18
** The median is reported for spelling*
*** This is the Z value from a non-parametric Mann-Whitney U test*

Table 6.12 ***Results of independent-samples t-tests for control and experimental groups for overall quality and individual constructs: school 4; higher level***

Classes 3 & 4 (higher level)	Group status	mean	standard deviation	t-value	significance level
overall quality	control	8.63	2.13	1.390	.168
	experimental	9.26	2.00		
grammatical complexity	control	7.88	2.37	.802	.425
	experimental	8.26	1.94		
grammatical accuracy	control	7.15	2.10	.968	.336
	experimental	6.74	1.70		
vocabulary range	control	8.73	1.96	.693	.490
	experimental	9.02	1.95		
spelling	control	14*		.106**	.906
	experimental	14*			
punctuation & paragraphing	control	8.03	1.79	.226	.822
	experimental	7.93	2.01		
coherence and flow	control	8.48	2.63	.328	.744
	experimental	8.65	2.25		

N control = 40
N experimental = 43
Range of possible scores = 3 to 18
** The median is reported for spelling*
*** This is the Z value from a non-parametric Mann-Whitney U test*

Division into the higher and lower level groups revealed that it was at the latter level that the statistical significance of all three significant effects — for *overall quality*, *grammatical complexity* and *coherence and flow* — was located. In addition, the t-value for *grammatical accuracy* for the lower pair of classes was relatively high, indicating that reading may have had more effect on grammatical accuracy at this level than for the group as a whole.

Whilst the higher-level experimental group had achieved slightly higher mean scores than its peer control group for *overall quality*, *grammatical complexity*, *vocabulary range* and *coherence and flow*, it had achieved a *lower* mean score for *grammatical accuracy*.

Participation in the reading scheme seemed to have had a *detrimental* effect on the grammar of the higher-level experimental students.

A possible reason for this might be that gains in grammatical complexity had led to restructuring, with the intervention of new structures into the grammar destabilizing previously mastered grammar, causing error. However, the raters had perceived no such major gain in the grammatical complexity of the experimental compositions. Nor had they perceived any major gain in vocabulary range, which may have led to the possible interpretation of the students focusing on vocabulary, to the detriment of their command of grammar.

Punctuation and paragraphing was unaffected by the split into lower and higher level, with negligible differences between control and experimental at both levels. *Spelling*, however, was very slightly affected, showing that the non-significant difference between control and experimental groups was perceptible only at the lower level, whilst no difference had been perceived by the raters at the higher level.

6.6 Inter-rater reliability for separate constructs

Table 6.13 ***Inter-rater correlations for separate constructs: school 4***

	raters 1 and 2	raters 1 and 3	raters 2 and 3	overall reliability correlation
grammatical complexity	.436	.535	.438	.727
grammatical accuracy	.440	.540	.459	.737
vocabulary range	.537	.496	.464	.750
spelling**	.423	.644	.309	.718
punctuation & paragraphing	.436	.028*	-.026*	.355
coherence & flow	.540	.423	.400	.715

N = 150
*All correlations except those marked * are significant at p < .001*
***A Spearman rank-order correlation was used for spelling*

Inter-rater agreement for the separate constructs for the smaller data set followed a very similar pattern to that for the larger data set (*cf.* Table 5.11). Correlations were slightly reduced. For example, the overall rater agreement coefficient for *grammatical complexity* was .772 for the larger data set and .727 for the smaller data set. Other constructs, apart from *punctuation and paragraphing*, showed similar differences. These slight reductions in coefficient values were probably, like those for *overall quality*, due to the smaller sample size and the reduced ranges of scores (since school 4 did not achieve the full range of possible scores in any construct other than spelling). The similarity of rater agreement correlations for the two data sets showed that, when making judgements on the separate constructs, raters did not operate less or more reliably for the lower level data than for the complete range of levels.

Punctuation and paragraphing was an exception to this. Whilst the inter-rater overall reliability correlation was .585 for the complete range of levels, it was only .355 for the lower level data. *Punctuation and paragraphing* may have presented more of a problem for the raters to judge consistently at the lower level than at higher levels. As may be seen in Tables 6.13 and 6.14, this construct not only produced extremely low coefficients between raters, but also negative correlations — although two out of three of the inter-rater correlation coefficients for *punctuation and paragraphing* in Table 6.13 and four out of six in Table 6.14, including all negative coefficients, were so low that they could not be said to indicate any relationship at all. The strength of the overall reliability correlations came solely from the agreement between raters 1 and 2.

This suggests that raters experienced even more problems with this construct at the lower level than in general (*cf.* Section 5.5), and that there was greater disparity between students' correct use of punctuation and their ability to apply rules of paragraphing at this level. Rater debriefing discussions revealed that raters 1 and 2 weighted heavily on paragraphing, deciding, *prima facie*, on a low score if a composition could immediately be seen to consist of only one paragraph. Rater 3, on the other hand, considered insufficient paragraphing as a once-off misjudgement, and a fault of no greater weighting than a single wrongly punctuated sentence. Hence rater 3's decisions were weighted on punctuation. This may explain why there is some level of agreement between raters 1 and 2, but no apparent agreement between rater 3 and either of the other two raters.

Table 6.14 ***Inter-rater correlations for separate constructs: control and experimental: school 4***

	Group status	raters 1 and 2	raters 1 and 3	raters 2 and 3	overall reliability correlation
grammatical complexity	control	.573	.694	.489	.815
	experimental	.282**	.380	.382	.617
grammatical accuracy	control	.499	.626	.632	.812
	experimental	.377	.473	.269**	.642
vocabulary range	control	.536	.607	.556	.796
	experimental	.526	.407	.385	.706
spelling***	control	.577	.722	.510	.824
	experimental	.226**	.550	.042**	.550
punctuation & paragraphing	control	.499	-.080**	-.029**	.337*
	experimental	.367	.158**	.068**	.428
coherence and flow	control	.598	.425	.483	.754
	experimental	.463	.410	.302*	.662

N control = 74
N experimental = 76
*All correlations except those marked * or ** are significant at p < . 001*
**Significant at p < .01*
***Not significant at p < .01*
****A Spearman rank-order correlation was used for spelling, which did not have a normal distribution*

The same phenomenon of markedly higher inter-rater agreement for the control compositions than for the experimental compositions as observed in the larger data set was observed for the lower-level school. This showed that level was not a factor for this effect. These differences achieved significance (Table 6.15) for *grammatical complexity* and *grammatical accuracy*, but not for *vocabulary range* and *coherence and flow*. (Possible

reasons for this effect were discussed in Chapter 5.) *Punctuation and paragraphing* showed no such effect.

Table 6.15 ***Z_{obs} values for differences between control and experimental averaged inter-rater correlations for separate constructs: school 4***

	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow
Z_{obs}	2.549	2.213	1.247	.6719	1.146

N control = 74
N experimental = 76
Z_{obs} is significant if it is greater than 1.96
Z_{obs} was not calculated for spelling

6.7 Relationships between constructs

The inter-construct correlation matrix for school 4 (Table 6.16), shows a pattern of relationships between constructs virtually identical to that produced by the complete range of levels (*cf.* Table 5.14).

Table 6.16 ***Correlations between scores on five constructs* and overall quality scores: school 4***

grammatical accuracy	.698				
vocabulary range	.748	.701			
punctuation & paragraphing	.548	.595	.469		
coherence and flow	.788	.711	.739	.648	
overall quality	.710	.727	.693	.513	.721
	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow

N = 150
All correlations are significant at $p < .001$
** Spelling is not included*

All inter-construct correlation coefficients were slightly lower for school 4 than for the whole data set, but by a uniform ratio, which is to say that no pair of constructs was affected more, or less, than any other pair of constructs by the general decrease in size of correlation coefficient. As was the case for inter-rater correlation coefficients, this general slight decrease in magnitude may have been caused by the effects of the smaller sample size and the technical effects of the reduced ranges of scores.

The strongest correlation was obtained between *grammatical complexity* and *coherence and flow*, whilst *grammatical accuracy* had the strongest relationship with *overall quality* judgements. *Punctuation and paragraphing* showed the weakest relationship with *overall quality* judgements and also with other constructs. These findings replicated exactly those of the matrix for the larger data set, indicating a very high level of stability in the interrelationships of the constructs across levels, since these were not affected by level. The constancy of the observed relationships (as exemplified by the correlation coefficients) further illustrated both the validity of the constructs and the inherent reliability of the rating procedure.

Splitting the matrix into two — a control data matrix and an experimental data matrix — did not provide any further insights (Table 6.17). A random pattern emerged whereby, for four out of the five variables, differences between control and experimental correlation coefficients were small and might operate in favour of either the control data or the experimental data, without any clear pattern. *Punctuation and paragraphing* was the exception, and the coefficients obtained by this construct in the control data were uniformly lower than those obtained in the experimental data. Since these coefficients could be affected by inter-rater reliability, this was most likely the effect of the lower inter-rater reliability coefficient for *punctuation and paragraphing* in the control data (Table 6.14). This, in turn, may have meant that the control compositions presented more of the single-paragraph compositions which provoked inter-rater disagreement.

Table 6.17 ***Correlations between scores on five constructs* and overall quality scores for split data set: school 4***

grammatical accuracy	.691 .719				
vocabulary range	.737 .753	.679 .729			
punctuation & paragraphing	.516 .606	.530 .673	.382 .555		
coherence and flow	.805 .754	.697 .735	.761 .707	.563 .764	
overall quality	.702 .691	.740 .724	.699 .677	.457 .591	.710 .714
	grammatical complexity	grammatical accuracy	vocabulary range	punctuation & paragraphing	coherence and flow

Non-shaded areas contain control coefficients: shaded areas contain experimental coefficients
N control = 74
N experimental = 76
All correlations are significant at p < .001
** Spelling is not included*

6.8 Quantity of production

Quantity of production was measured in three ways: number of words, number of T-units and number of clauses. The distribution charts for number of T-units and number of clauses showed normal distributions (Appendices 12 and 13). The greater spread for numbers of words, which ranged from 132 to 568, resulted in a very wide distribution, and the majority of actual scores observed (*i.e.* word counts) were achieved by only a single composition. (A total of 150 compositions shared a range of 437 possible scores.) Only six scores were each attained by three or more compositions, the maximum number of compositions achieving the same score being five.

A simple transformation was applied, dividing the number of words by ten and discarding any digits after the decimal point. For example, 365 words transformed to 36; 422 words transformed to 42. In effect, this simply grouped the scores into bands, each band representing ten scores. For example, the score of 18 included all original word counts between 180 and 189 inclusive; 19 included all word counts between 190 and 199, and so on. The transformed scores, or bands, produced a normal distribution (Appendix 11). Means and standard deviations for the four classes are given in Table 6.18.

Table 6.18 ***Means and standard deviations for number of words, number of T-units and number of clauses: school 4; four classes***

	lower level control class N = 34		lower level experimental class N = 33		higher level control class N = 40		higher level experimental class N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
N words*	23.55	5.66	27.21	7.11	26.70	5.83	33.31	8.80
N T-units	32.27	9.13	35.91	10.30	35.23	8.45	42.18	13.15
N clauses	45.97	12.93	55.94	15.71	52.67	11.79	65.44	19.67

** For a true picture, figures given for words should be multiplied by 10*

Participation in the reading scheme had a strong impact on quantity of production. At the lower level, the mean number of words was 235 for the control compositions and 272 for the experimental compositions. This represented a 15.7% greater output for the reading-scheme students. At the higher level the difference was even greater, with a mean of 267 words for the control compositions and 333 words for the experimental compositions, representing a 24.7% greater output. Additionally, the quantitative output of the lower-level experimental class exceeded that of the higher-level control class, suggesting that, strikingly, the reading scheme affected the production fluency of the lower experimental group enough to raise it to a slightly higher level than that of the non reading-scheme higher class.

Results of significance tests for the three measures of quantity, at lower and higher levels, are given in Table 6.19

Table 6.19 ***Results of independent-samples t-tests for number of words, number of T-units and number of clauses: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
N words	2.323	.023	3.959	.000
N T-units	1.285	.203	2.829	.006
N clauses	2.840	.006	3.615	.001

At the higher level, number of words was the best discriminator between control and experimental, whereas at the lower level number of clauses was the best discriminator. Number of T-units failed to discriminate between control and experimental groups at the lower level, and did not discriminate quite so well, though only relatively so, as number of words and number of clauses at the higher level.

The intervening factor was complexity. Since increasing complexity manifests itself, at the level of the T-unit, in longer units, their use as a simple measure of quantity of output may disfavour more complex texts. Thus the lower-level experimental group, judged by raters as exhibiting significantly higher levels of complexity, was disadvantaged by the use of number of T-units as a measure of quantity of output. Number of clauses, on the other hand, gave an advantage to the lower-level experimental group, increasing the t-statistic from 2.323 (when number of words was used) to 2.840.

Quantity of output is often seen as a measure of fluency of production. In more cognitive terms, it can be seen as a measure of speed of retrieval or of levels of *automaticity*. (These are not necessarily the same thing: *cf.* Chapter 7.) Neither fluency of production nor automaticity featured in the range of constructs which the raters were asked to, or indeed, *could*, evaluate. Whilst speed of production may be a good indicator of these, this is not *directly* observable from the finished product, but can only be inferred from quantity of output relative to a specified period of time.

The issue of composition length was one which raters raised repeatedly, particularly with regard to accuracy, as they found it hard to make comparisons of accuracy across compositions of very different lengths. For example, should a long composition with twenty or so grammatical errors be given the same rating as a composition half that length but with the same number of errors? Should it receive the same rating as a composition half that length but with *half* that number of errors? (Numbers of errors were not actually counted by the raters; the concept is used here only for illustration.) The answer given to the raters was that they should trust their first, impressionistic, judgement of overall accuracy. In practice, this came down to overall "texture" of errors, or how soon one error followed another and how long were the intervening stretches of error-free text, with nonetheless some slight adjustment for text length. Effectively — to give a very approximate idea — a 200-word composition with an error every 10 words might be judged similarly to a 300- to 400-word composition with an error every 12 or 13 words.

Of the four linguistic constructs of *grammatical complexity*, *grammatical accuracy*, *vocabulary range* and *coherence and flow*, and taking *overall quality* as a fifth, *grammatical accuracy* was the one that correlated least well with composition length (Table 6.20).

Table 6.20 ***Correlations between composition length and scores on linguistic constructs: school 4***

	overall quality	grammatical complexity	grammatical accuracy	vocabulary range	coherence and flow
composition length	.661	.556	.403	.599	.512

N = 150
All correlations are significant at p < .001

Two points may be made about the correlation coefficients above. Firstly, composition length correlated well with *overall quality* judgements — in fact, almost as well as the linguistic constructs themselves did (*cf.* Table 6.16.). This suggests that raters may have taken length into consideration, consciously or unconsciously, when making judgements on *overall quality*. (Raters reported that whilst they tried not to be improperly influenced by length when making judgements on *overall quality*, they were aware of consistently awarding more of the higher grades to longer compositions.) This could, however, also have been because greater length coincided with better text quality.

Secondly, whilst *grammatical accuracy* did, in absolute terms, increase with length (otherwise there would be no significant correlation, or possibly a negative one), it clearly did not keep constant pace with length, and there may not have been an absolute linear relationship between the two. The lower correlation between *grammatical accuracy* and length could not simply be accounted for by any greater degree of unreliability in either of these two variables than in the other variables of *grammatical complexity*, *vocabulary range* and *coherence and flow*, since *grammatical accuracy* was the second most reliable of the six constructs (*cf.* Table 6.13), and composition length was a direct measurement. A strictly linear relationship between what were two of the most reliable of the variables in Table 6.20, should have shown one of the *highest* correlation coefficients.

To address the first of these points, which is to say the high correlation between composition length and *overall quality* scores, a standard multiple regression was performed, using *overall quality* as the dependent variable and the five normally distributed constructs and length as independent variables. Whilst the five constructs together accounted for 64.2%

(R^2 = .642) of the first range of *overall quality* scores, the addition of length as an independent variable increased this to 71.9% (R^2 = .719).

Composition length was thus seen to be one of the contributing factors to the unexplained variance in *overall quality* scores observed previously (Section 6.2). For the second range of *overall quality* scores, R^2 was increased from .694 to .767 by the inclusion of length as an independent variable. Since the increase in total variance which could be accounted for by the combined independent variables was very similar for the first and second ranges of *overall quality* scores (7.7% and 7.3% respectively), it could be concluded that, whilst length was an influential variable, it was so equally for both ranges of scores. It was not a variable which was factored in at some greater, or enhanced, level to the second range of *overall quality* scores.

These increases of approximately 7% should not be seen as representing the total contribution of length to *overall quality* scores. Rather, they represent the *unique* contribution of length. Any contribution of length in combination with any other variable would already have been included in the equation, as part of the contribution of the other variable, attributed, in the absence of length as a discrete variable in its own right, simply to that variable on its own.

Addressing the second point, that grammatical accuracy may not have had a constant relationship with length, offered some clue to the *de*crease in grammatical accuracy observed in the higher-level experimental group (*cf.* Table 6.12). This group showed a greater increase in composition length, in relation to its experimentally matched control group, than did the lower-level experimental group. The effect size (eta squared) was .162, representing a large effect. (According to Cohen, 1988, an eta squared value of .06 indicates a moderate effect and .14 indicates a large effect.)

If the decrease in grammatical accuracy of the higher-level experimental group could not be explained by an increase in complexity, which might have a damaging effect (either perceived, because of the less accurate use of more complex structures, or actual) on accuracy, it may have been a result of increased length. Greater length may have led to greater scope for quantitative display of error. Raters may have been influenced by sheer numbers of errors, made possible by much longer compositions. Alternatively, this may have been a *real*, proportional, increase in error.

To investigate more closely any relationship between length and accuracy, a cluster analysis was performed on the highest scoring compositions. Cluster analysis is not a mathematical statistic, but makes use of the computational possibilities afforded by a computer programme to order data into "best fit" categories.

In the present data, a composition may have achieved a high score via a variety of routes. For example, it may have been short, but very coherent and very accurate. Conversely, it may have been much longer, and not very accurate, but have demonstrated a wide range of vocabulary. Either of these cases — and other permutations of length, complexity, accuracy, vocabulary use and coherence — may have led to a rater giving a high score. Only the most exceptional compositions might show high performance levels in *every* aspect.

The 30 top-scoring compositions, for *overall quality*, out of the whole group, irrespective of control or experimental status, were classified using SPSS's "quick cluster" option. Only the high scoring compositions were of interest for this particular analysis, as these might show patterns which low scoring compositions could not. (Low scoring compositions may have relatively low performance levels in nearly all aspects, and are thus more difficult to discriminate amongst for performance on individual constructs.) The 30 compositions had all been awarded at least 4, out of a possible 1 to 6, by more than one rater. The range of *overall quality* scores was 11 to 13, with a mean score of 11.27. (The range for the whole group of school 4 compositions was 3 to 13, with a mean of 7.92.) Only the linguistic constructs and composition length were used in the classification. Classification into two clusters was selected as optimal for enhancing any differences.

Table 6.21 Cluster centres for 30 top-scoring compositions: school 4

	Cluster	
	1	2
number of words	309	457
grammatical complexity	9	10
grammatical accuracy	9	7
vocabulary range	10	10
coherence and flow	10	10

Range of possible scores for linguistic constructs = 3-18

Table 6.21 shows the two patterns for the top-scoring compositions which were extracted from the data by the clustering procedure. Under the headings *cluster 1* and *cluster 2* are

presented a set of cluster *centres*. These *centres* represent the nearest approximate value of each construct. They are not strictly means, but may be regarded in the same light for the purposes of cluster interpretation. Thus each set of cluster centres represents the "typical" or "average" profile for a composition in that group, or cluster.

The two typical profiles which emerged were that of a composition with an average of 309 words, and that of a composition with an average of 457 words. These two "prototype" compositions did not differ in terms of *vocabulary range* or *coherence and flow*, both presenting a cluster centre (or typical score) of 10 for these. However, the typically longer composition had a higher *grammatical complexity* score, indicating that longer compositions, in the set of 30, tended to score better on this. On the other hand, these tended to have lower *grammatical accuracy* scores (a centre of 7, compared to 9 for the typically shorter composition).

The associated ANOVA produced by the clustering procedure (which is descriptive only, and may not be used as a significance test *per se*) found the differences in length and *grammatical accuracy* to be significant ($p = .000$ for length; $p = .044$ for *grammatical accuracy*). In other words, when asked to arrange the high-scoring compositions into two groups with the best possible sets of differentiating characteristics, the computer programme found the best separating features to be length and *grammatical accuracy*, and that compositions which tended to be long tended to have lower scores for *grammatical accuracy*, whilst compositions which tended to be short tended to have *higher* scores for *grammatical accuracy*. Although longer compositions were also judged, on average, as being slightly more complex, *grammatical complexity* was not a significant differentiating factor. These findings strongly supported the hypothesis that grammatical accuracy and composition length did not act in tandem, and that, whilst increased complexity may have had a slight detrimental effect on *grammatical accuracy* scores, increased length was likely to have had a much greater effect.

As the higher level experimental class had produced compositions which were, on average, over 24% longer than those of the experimentally matched control class, this may explain why that group also had a lower mean score for *grammatical accuracy*. Of the actual compositions assigned by the clustering procedure to cluster 2 — the longer, but less grammatically accurate group — only one was a control composition. Thus, length appeared to have had a detrimental effect on accuracy scores, and the higher-level experimental group

was the group most affected by this. This may have been a real effect on accuracy itself, somehow related to the fact of producing more output, or of producing output more quickly, or the compositions may have merely attracted lower ratings because of their length and consequently greater number of errors.

6.9 Length of syntactic unit

Two values for mean sentence length are shown in Table 6.22. These are mean sentence length for whole composition, and mean sentence length for narrative text only. (The rationale behind using narrative text only was discussed in Section 4.3.6.2.) It should be noted that figures given are not actual means (*i.e.* they do not derive from the total number of words produced by a class divided by the total number of sentences), but are the means of means (*i.e.* the mean number of words per sentence for each composition added together and divided by the number of compositions). Thus, the mean sentence length of a short composition has the same value as that of a longer composition. Although this does not represent a true mean, it gives a clearer picture of the abilities of individuals. Distributions are included in the Appendices.

Table 6.22 ***Means and standard deviations for mean length of sentence*: whole composition and narrative text only: school 4; four classes***

	lower level control class N = 34		lower level experimental class N = 33		higher level control class N = 40		higher level experimental class N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
whole composition	8.34	1.42	9.18	1.84	9.16	1.93	9.61	2.43
narrative text only	8.62	1.62	9.88	2.42	9.22	1.97	10.15	2.62

**Length of sentence was defined as number of words per sentence*

When whole compositions were used, the difference between mean sentence length for control and experimental groups was of less than one word for the lower pair of classes and less than *half* a word for the higher pair of classes. The scale of these differences is partly a reflection of the shortness of the mean sentence lengths overall, 87% of which were within the range of 7-12 words. However, the fact that observed differences between control and experimental mean sentence lengths for narrative text only (*i.e.* excluding directly reported

dialogue) were half as much again for the lower level and over twice as big for the higher level than for integral text, illustrates the susceptibility of this measure to genre.

Mean length of sentence, or of any other syntactic production unit, may not in fact be well-suited to revealing differences in *ability to produce* longer units, and may not do so very sensitively. In a research synthesis article, Ortega (2003) found that of 14 between-proficiency comparisons only six revealed a significant difference in mean sentence length. Not only were these comparisons between groups of previously recognised different proficiency levels, but they were not necessarily comparisons between adjacent levels, and included comparisons between maximally different proficiency levels. Unfortunately, Ortega did not indicate the difference in proficiency levels between the two groups in each comparison. However, since the 14 comparisons derived from only four multi-level studies, it can be assumed that the six significant differences found were most likely not between adjacent proficiency levels.

Means are themselves the products of a range of scores, and, unless the complete range of scores is uniformly moved in the same direction, more effort is needed to raise a mean than to raise an individual score. This uniform movement may not be the case for range of sentence (or T-unit) lengths produced by a learner, as the ability to produce longer and longer sentences does not necessarily preclude the continued use of shorter sentences.

If there is no need to produce a long sentence, even a very proficient student might not do so. He *may* do so, but this is not a very reliable condition, and whether or not he does so may depend as much on genre, amongst many other things, as on the student's ability. What may happen is that, as the top score (or sentence length) increases, the bottom score remains relatively stable, since the student does not, with increased proficiency, simply stop using shorter sentences. The mean cannot depend on movement at both ends of the distribution, but must depend for its increase on expansion at the top end of the distribution. Thus the mean of the complete range of produced sentence lengths does not increase directly in keeping with increased ability to produce longer sentences. In approximate terms, scores at the top end of the scale must work twice as hard to raise the mean as would be the case if *both* ends of the distribution were working to raise the mean. The range of sentence-length means obtained by a whole group of students may be more compact than the range of actual abilities, and this might be one reason why researchers have found it difficult to obtain significant results when comparing sentence lengths of differing proficiency levels.

Independent-samples t-tests comparing the mean sentence lengths of control and experimental groups gave the t-values in Table 6.23.

Table 6.23 ***Results of independent-samples t-tests for mean length of sentence: whole composition and narrative text only: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
whole composition	2.101	.040	.912	.364
narrative text only	2.517	.014	1.808	.074

Despite the difficulty of obtaining significant results for differences between mean sentence lengths (Ortega, 2003), means for both integral text and narrative text only showed a significant difference between control and experimental for the lower-level pair of classes. Calculating mean sentence length for narrative text only, compared to using complete compositions, had a greater effect on the higher pair of classes than the lower, to the extent that it doubled the t-value, although this remained non-significant. This was because the higher-level experimental group tended to reproduce direct speech more frequently than the other groups. Thus the exclusion of text enclosed within direct speech quotation marks affected that group more.

Table 6.24 ***Means and standard deviations for mean length of T-unit and mean length of clause: school 4; four classes***

	lower level control class N = 34		lower level experimental class: N = 33		higher level control class N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
mean length of T-unit*	7.72	1.03	8.08	1.02	7.93	1.14	8.37	1.75
mean length of clause	5.16	.57	4.84	.43	5.03	.41	5.09	.51

Length of T-unit and of clause was defined as number of words per T-unit or clause
**Clauseless expressions were not included in calculating mean length of T-unit*

No significant differences were found between control and experimental groups for mean length of T-unit at either the higher or the lower level, although in both cases the experimental group produced slightly longer T-units. When a similar weak effect is observed in comparable groups, one is recommended to increase the sample size by combining the groups. Any effect which is not a true effect will be weakened by increased sample size, whilst any true effect will gain in stability. Increasing the power of the t-statistic by increasing the sample size (*i.e.* by combining the higher and lower groups) resulted in a t-value of 1.928 ($p = .056$), only tending towards significance. Thus mean length of sentence, the ideational unit of the writer, appeared to differentiate better between control and experimental groups than did mean length of the more strictly technical T-unit.

Mean length of clause did not discriminate between levels nor between control and experimental groups. In fact, *length* of clause may have little to do with clausal sophistication. Clauses are very strictly defined, and may include only one action. Any new action must result in a new clause. Sentences and T-units may carry as many actions as the writer can, and wishes to, incorporate into a grammatically correct unit. They are subject to ever-increasing possibilities of expansion, which clauses are not. Moreover, syntactic techniques which result in longer sentences and T-units may actually have the effect of *shortening* the clauses which are combined to produce these longer sentences and T-units, since main clauses often carry a base of information which may then be ellipted in subsequent subordinate clauses. Thus more sophisticated clauses may be *shorter* than less sophisticated clauses.

Additionally, clauses carry an inherent third dimension of sophistication which sentences and T-units do not. Sentences and T-units may increase their complexity by adding on and embedding, re-structuring grammar and syntax as they do so. Clauses, however, have an interior, more psychological, sophistication, reflected in the variety and classification of clause types (*cf.* Section 6.10).

Correlation coefficients calculated between mean lengths of syntactic units and rater judgements on *overall quality*, *grammatical complexity*, *grammatical accuracy* and *coherence and flow* showed that the strongest relationship was between *grammatical complexity* and mean sentence length (Table 6.25). Mean sentence length and *overall quality*

showed the second highest correlation. Mean sentence length also appeared to have a stronger association with *coherence and flow* judgements than did mean length of T-unit.

Table 6.25 ***Correlations between mean length of syntactic units and rater evaluations: school 4***

	overall quality	grammatical complexity	grammatical accuracy	coherence and flow	length of composition
mean length of sentence***	.243**	.279**	.074	.175*	.064
mean length of T-unit****	.184*	.235**	.038	.123	-.005
mean length of clause	-.053	-.030	-.077	-.131	-.152

N = 150
**Significant at p < .05*
***Significant at p < .01*
****Sentence length for narrative text only*
***** Excluding clauseless expressions*

Mean sentence length may thus be seen as having had a closer association with rater judgements of *grammatical complexity*, *overall quality* and *coherence and flow* than did mean length of T-unit. Whilst this may be because mean length of sentence is, in fact, more reliably connected to overall text quality than is mean length of T-unit, it may also be that raters did not evaluate texts against a conceptual background of T-units, but against a background of sentences, and were not sensitive to the operations of T-units.

Mean length of clause produced spuriously low correlations, all of which were negative, supporting the contention — since, although none of these correlation coefficients was significant, all five of them were negative — that increasing *length* of clause is not a good indicator of sophistication, but that clauses *may* (for narrative writing at this level) be just as likely to become shorter as to become longer as text increases in sophistication. Length of syntactic production units did not appear to have any clear relationship with *grammatical accuracy* judgements.

Wolfe-Quintero *et al.*'s (1998) contention that length of syntactic unit is in fact more closely linked to, and may be used as an objective measure of, fluency, and not complexity, was not borne out by the correlational patterns in Table 6.25. Observation of text fluency was part of

the rating rubric for *coherence and flow* judgements. However, the relationships between mean lengths of sentence and T-unit and *grammatical complexity* judgements were stronger than those between these mean lengths and *coherence and flow* judgements. Increased length of syntactic unit may be seen to have had a closer association with text complexity judgements than with text fluency judgements. In the case of *production* fluency, the best available measure was composition length. There did not, however, seem to be any relationship between this and length of syntactic unit, as there was no correlation between them.

A clustering procedure using composition length and mean lengths of the three types of syntactic units as four variables discriminated largely on the basis of composition length (Table 6.26), producing two profile compositions of different lengths but of quite similar mean lengths of syntactic unit. The associated ANOVA produced by the clustering procedure found the only significant factor to be composition length. Cluster 2, representing a typically longer composition, did, however, also have a slightly higher mean sentence length and T-unit length, and a very slightly shorter mean clause length. Although these differences were not great, they may represent tendencies, and a difference in sentence length of nearly half a word — given the difficulty of discriminating between performance levels in terms of mean sentence length — may not be entirely negligible. This pattern was also symptomatic of increasing text complexity. Thus increasing composition length did not appear to be in strong competition with increasing, objectively measured, syntactic complexity. This was also the case for increasing composition length and rater-judged complexity (*cf.* Table 6.21).

Table 6.26 ***Cluster centres for 150 compositions for length of composition and length of syntactic unit: school 4***

	Cluster	
	1	2
number of words	240	376
mean length of sentence*	9.4	9.8
mean length of T-unit**	8.2	8.3
mean length of clause	5.0	4.9

**Narrative text only*
*** Not including clauseless expressions*

6.9.1 Clauses per sentence and clauses per T-unit

Ratios of clause to sentence and clause to T-unit were also calculated. Integral texts were used for these calculations, since ratios were not likely to be so affected by the inclusion of text representing spoken dialogue as mean length of syntactic production unit might be. Possible effects of shorter sentences (which may be characteristic of spoken dialogue) on any overall picture may be constrained by the fact that ratios are produced by dividing one syntactic unit by its own component syntactic units. (For example, a two-word sentence consisting of a two-word T-unit will give the same T-unit per sentence ratio as a much longer sentence consisting of a much longer T-unit.) Table 6.27 gives mean numbers of clauses per sentence and per T-unit. As with mean numbers of words per sentence, per T-unit and per clause (Tables 6.22 and 6.24) these are not true means, but the means of values which are themselves means. (Distribution charts are reproduced in Appendix 15.)

Table 6.27 ***Means and standard deviations for mean numbers of clauses per sentence and per T-unit: school 4; four classes***

	lower level control class N = 34		lower level experimental class: N = 33		higher level control class N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
clauses per sentence	1.60	.24	1.87	.34	1.81	.39	1.86	.39
clauses per T-unit	1.47	.19	1.62	.18	1.55	.24	1.60	.26

Both mean ratios of clause to sentence and clause to T-unit discriminated between control and experimental groups. The differences were small and not significant at the higher level, although the experimental group may have produced slightly more complex sentences than the control group. At the lower level, however, relatively strong effects were apparent for both sentences and T-units (Table 6.28). Eta squared was .17 for the difference between mean number of clauses per sentence for control and experimental groups and .14 for the difference in mean number of clauses per T-unit, showing a large effect in both cases (Cohen, 1988).

Table 6.28 ***Results of independent-samples t-tests for mean numbers of clauses per sentence and per T-unit: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
clauses per sentence	3.709	.000	.898	.559
clauses per T-unit	3.314	.002	.894	.372

Effects at the lower level may have been particularly pronounced because not only was there an increase in both mean sentence length and mean T-unit length between control and experimental groups, but there was also a *de*crease in mean length of clause (*cf.* Tables 6.22 and 6.24). Ratios were thus achieved by dividing longer sentences and T-units by shorter clauses, enhancing any effect. There is also a possibility that using integral texts could have inhibited any effect at the higher level. If characteristic ratios of clause to sentence and T-unit for the spoken genre *were*, in fact, lower than for narrative text, the higher-level experimental group may have been disfavoured, since this group used more direct speech than the other three groups.

6.9.2 Relationships between clause per sentence and clause per T-unit ratios and rater evaluations

The strongest correlations for clause per sentence and clause per T-unit ratios were obtained with *grammatical complexity*, reaffirming the association of such objective complexity measures with the raters' complexity judgements (Table 6.29). However, whereas mean sentence length and mean T-unit length both had a slightly stronger association with *overall quality* judgements than with *coherence and flow* judgements (*cf.* Table 6.25), clause per sentence and clause per T-unit *ratios* seemed to each have an equal relationship with *overall quality* and *coherence and flow*. As with mean sentence length and mean T-unit length, neither of the two ratio measures could be associated — either positively or negatively — with grammatical accuracy or production fluency (as represented by composition length).

Table 6.29 ***Correlations between mean clause per sentence and mean clause per T-unit ratios and rater evaluations: school 4***

	overall quality	grammatical complexity	grammatical accuracy	coherence and flow	composition length
clauses per sentence	.188*	.297**	.074	.184*	.042
clauses per T-unit	.174*	.252**	.050	.174*	.047

N = 150
**Significant at p < .05*
***Significant at p < .01*

6.10 Clause type

Whereas mean *length* of clause did not prove to be a reliable discriminator between differing levels of text complexity, clause *type* — an indicator of text sophistication — revealed some interesting differences between control and experimental compositions. All clauses, including those within direct speech text, were classified into three principal types of main clause, coordinate clause and subordinate clause. Within these types, subordinate clauses were further classified by sub-type and purpose. Coordinate clauses were divided into *full* coordinating clauses and *reduced* coordinating clauses (*cf.* Section 4.3.6.2).

Numbers of main clauses, full coordinating clauses and reduced coordinating clauses produced normal distributions (Appendix 13). The distributions for both full coordinating and reduced coordinating clauses were positively skewed, with skewness of 1.09 and 1.47 respectively, indicating a bunching of scores (*i.e.* mean numbers of clauses) towards the lower end of these two distributions. Thus more compositions achieved a more restricted range of low scores at the bottom end of these distributions with fewer compositions achieving a wider range of scores at the higher end. In other words, most students did not use many coordinate clauses, although a small number *did.*

The division of subordinate clauses into smaller groups of sub-types resulted in very erratic distributions, with single compositions exhibiting instances of only some of the sub-types. Indeed, many of these were evident in only a few compositions and there were consequently many cases of zero instance. Subordinate clauses of all types were therefore grouped together into one generic category of subordinate clause in order to create a more workable and useful variable. Although this meant some loss of information concerning *type* of

subordinate clause, the property of being subordinate is in itself an important defining feature and a useful measure of text sophistication at a different dimension than that captured by simple length of syntactic unit. (A full table of frequencies for individual sub-types of subordinate clause for each of the four classes may be found in Appendix 23.) Subordinate clauses, when grouped together, produced a normal distribution (Appendix 13).

Table 6.30 Mean numbers of clause types per composition: school 4; four classes

	lower level control class N = 34		lower level experimental class: N = 33		higher level control class N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
main clauses	29.09	8.59	31.09	10.59	30.12	8.26	36.44	12.18
full coordinating clauses	2.59	2.10	4.21	2.78	4.55	2.58	5.53	3.67
reduced coordinating clauses	1.65	1.53	3.09	1.77	2.95	1.82	4.05	3.23
subordinate clauses	12.65	5.46	17.55	7.36	15.05	6.11	19.42	7.89
total number of clauses	45.97	12.93	55.94	15.71	52.67	11.79	65.44	19.67

As might be expected, given the observed differences in length between control and experimental compositions at both the lower and higher levels, with no accompanying proportional increase in clause length (but even a slight *de*crease in clause length at the lower level), there was a steady increase in number of clauses of all four types, from control to experimental, at both levels. In addition, with the exception of full coordinating clauses, the output of the lower-level experimental class outstripped not only that of its peer control class, but also that of the higher-level control class. (A similar observation was made for composition length; *cf.* Table 6.18.)

Independent-samples t-tests showed that in the case of the lower-level group only mean number of main clauses did not produce a significant difference, and in the case of the higher group there were no significant differences between mean numbers of full coordinating and of reduced coordinating clauses (Table 6.31). However, when full coordinating and reduced coordinating clauses were added together to form one superordinate group of coordinating clauses, a t-value of 2.100 ($p = .040$) was obtained.

Table 6.31 ***Results of independent-samples t-tests for numbers of clause types: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
main clauses	.851	.398	2.744	.007
full coordinating clauses	2.700	.009	1.421	.159
reduced coordinating clauses	3.566	.001	1.921	.059
all coordinating clauses	3.659	.001	2.100	.040
subordinate clauses	3.099	.003	2.804	.006

The internal distribution patterns of clause types — the clausal "texture" of the text — was investigated in the form of ratios of each clause type to total number of clauses. Table 6.32 provides a general overview, giving percentages of total number of clauses for each clause type within the overall corpus of text for each class. (Thus, these percentages are true raw percentages rather than mean percentages derived from calculations involving the percentage figures from individual compositions.) Table 6.33 gives similar information, using as variables the mean number of each type of clause per 100 words of running text per composition. (Group mean values in this latter case are the means of means, as each individual composition itself produced a mean.) Reduced and full coordinating clauses were combined to give one value. Distributions were normal, enabling the use of independent-samples t-tests, and t-values for these are also reported in Table 6.33.

Table 6.32 ***Clause types as percentages of total number of clauses: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
main clauses	63.3%	55.6%	57.2%	55.7%
full coordinating clauses	5.6%	7.5%	8.6%	8.4%
reduced coordinating clauses	3.6%	5.5%	5.6%	6.2%
subordinate clauses	27.5%	31.4%	28.6%	29.7%

From the above table, it is clear that for both levels there was a decrease, from control to experimental, in proportional use of main clauses to other types of clause. This was more marked in the case of the lower-level pair of classes, where the experimental group showed a 3.8% higher proportional use of coordinate clauses (full and reduced) than its peer control group, and a 3.9% higher proportional use of subordinate clauses. When counted as mean number of each clause type per individual composition per 100 running words (Table 6.33), the differences between relative frequency of coordinate clause and subordinate clause use by control and experimental groups at the lower level were both statistically significant (p = .001 and .018 respectively).

The differences in proportion of use of each clause type were more subtle for the higher-level pair of classes, where the experimental group showed only a 0.4% higher proportional use of coordinate clauses, and a 1.1% higher proportional use of subordinate clauses. This accompanied a 1.5% lower proportional use of main clauses by the experimental class, as opposed to 7.7% lower proportional use of main clauses by the lower-level experimental class (compared to its matched control class). None of these changes in proportional use of clause types at the higher level reached statistical significance.

Table 6.33 ***Means and t-values for mean numbers of clause type per 100 words: school 4; two levels***

	Lower level group				Higher level group			
	control *N = 34*	experi-mental *N = 33*	t-value	sig	control *N = 40*	experi-mental *N = 43*	t-value	sig
main clauses	12.46	11.47	1.922	.059	11.44	11.09	.714	.477
coordinating clauses	1.74	2.87	3.343	.001	2.86	2.84	.066	.947
subordinate clauses	5.41	6.48	2.431	.018	5.69	5.91	.532	.596

What is apparent from the results presented in the above four tables is that differences between control and experimental compositions were not the same at the two levels. At the lower level, the clausal texture in the experimental compositions was different from that of the control compositions, with more coordination and subordination. This was brought about by a significant increase in the numbers produced of coordinate and subordinate

clauses, which outstripped the increase in numbers produced of main clauses. In other words, as students produced longer compositions, they also produced relatively more frequent coordinate and subordinate clauses. Thus text improvement evolved on two fronts at the same time.

The higher-level experimental students, on the other hand, produced longer compositions than their control counterparts, but production of the three clause types was affected more equally. Numbers of main clauses, coordinate clauses and subordinate clauses were higher in these longer compositions in very similar proportion, with only a slight reduction in percentage of main clauses balanced against a similar slight increase in percentage of subordinate clauses, whilst the proportional use of coordinate clauses remained stable.

Two explanations might be suggested for this. It is possible that, for the higher-level experimental group, the deployment of cognitive resources to produce greater quantity of output inhibited any accompanying increase in sophistication of clausal texture. In other words, the students may have diverted their language competence to simply producing longer compositions, whilst maintaining the same levels of coordination and achieving only slightly enhanced levels of subordination. It is also possible that the lower-level experimental group showed changes which the higher group did not simply because there was more scope for change at such a low level. It may have been that the higher classes — both control and experimental — had already achieved an optimal balance of main to non-main clauses for this kind of low level simple narrative text. The ratio of approximately 55% main clauses to 45% non-main clauses, or just under one dependent subordinate or coordinate clause for each independent main clause, which *both* experimental classes displayed, may represent a sort of plateau in this kind of text production, and it may not have been possible for the higher-level experimental students to change this ratio without a very drastic change in proficiency level.

6.10.1 Relative clauses

Although all types of subordinate clauses were grouped together for the above analyses, relative clauses were nonetheless singled out for particular comparisons of frequency, as these have been cited as specific predictors of putative L2 English developmental stages (*e.g.* Izumi, 2002). These did not present a normal distribution, and a number of compositions did

not demonstrate any use of relative clauses. Because overall numbers were relatively low, and to increase the power of the statistic, defining and non-defining relative clauses were grouped together for the purpose of significance testing. A non-parametric significance test was used to compare control and experimental performance at the two levels. Table 6.34 gives overall numbers of defining and non-defining relative clauses for the four classes. Table 6.35 gives the results for Mann-Whitney U tests.

Table 6.34 ***Numbers of defining and non-defining relative clauses: school 4; four classes***

	lower level group		higher level group	
	control *N = 34*	experimental *N = 33*	control *N = 40*	experimental *N = 43*
N defining relative clauses	22	51	47	81
N non-defining relative clauses	0	5	1	7
total N relative clauses	22	56	48	88

Table 6.35 ***Z values for Mann-Whitney U tests comparing number of relative clauses (defining + non-defining) for control and experimental groups: school 4; two levels***

lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
Z value	significance	Z value	significance
2.288	.022	1.958	.050

Although results were significant for both the lower and the higher levels, the numbers of clauses involved were relatively small, representing, on average, only one or two instances of a relative clause per composition, and, in the light of such general sparseness of incidence, might best be seen as evidence of a trend which may have been in its beginning stages. As the numbers of relative clauses increased, however, there was a steady reduction in the number of compositions exhibiting zero instance of relative clauses, from lower to higher level — supporting the claim that the emergence of relative clauses is an indicator of

increasing language competence — and from non reading-scheme to reading-scheme students — suggesting that the reading scheme supported the emergence of relative clauses. Table 6.36 gives numbers and percentages of compositions which displayed zero use of relative clauses.

Table 6.36 ***Number of compositions which did not contain relative clauses: school 4; four classes***

	zero defining relative clauses		zero non-defining relative clauses		zero relative clauses (defining or non-defining)	
	number	percent	number	percent	number	percent
lower level control group; *N = 34*	18	52.9	34	100	18	52.9
lower level experimental group; *N = 33*	12	36.4	28	84.8	11	33.3
higher level control group; *N = 40*	15	37.5	39	97.5	14	35
higher level experimental group; *N = 43*	12	27.9	37	86	10	23.3

The tendency towards increased use of relative clauses by the higher-level experimental group suggested that there *was* some improvement in the level of clausal sophistication for this group, despite the non-significance of the observed slight increase in proportional use of subordination and the stability of proportional amount of coordination in their compositions. This, in turn, might offer some support for the proposition that overall increase in subordination may have been more difficult to achieve at this higher level than at the lower level, and that this prevented the higher-level experimental group demonstrating much progress in this area.

6.10.2 Full coordinating and reduced coordinating clauses

The classification of coordinate clauses into full and reduced coordinating clauses, as a measure of the possible emergence of the use of ellipsis, did not reveal any differences between proficiency levels, but suggested some difference between control and experimental students. Numbers for these clauses and percentages of coordinating clauses which were ellipted are given in Table 6.37.

Table 6.37 ***Numbers of full and reduced coordinating clauses and percentages of coordinate clauses which used ellipsis: school 4; four classes***

	N full coordinating clauses	N reduced coordinating clauses	total N coordinate clauses	percentage of coordinate clauses using ellipsis
lower level control group *N = 34*	88	56	144	38.9%
lower level experimental group: *N = 33*	139	102	241	42.3%
higher level control group *N = 40*	182	118	300	39.3%
higher level experimental group: *N = 43*	238	174	412	42.2%

Although ratio of coordinate clauses demonstrating the use of ellipsis was not tested for significance, the fact that the lower- and higher-level control groups demonstrated similar percentages, as did lower- and higher-level experimental groups, suggested two things. Firstly, the use of ellipsis was not a discriminating factor between the text produced by lower- and higher-level students who did not take part in the ERS. Thus any use of ellipsis did not appear to be a level effect. Secondly, since there was increased use of ellipsis in the writing of both lower- and higher-level experimental groups, the reading scheme may have encouraged a tendency towards this. Such a tendency, not being associated with proficiency level, may have been entirely an effect of the increased exposure to instances of ellipsis afforded by extensive reading.

6.10.3 Relationship of clause types to raters' evaluations of *grammatical complexity*

Table 6.38 ***Correlations between numbers and ratios of clause type and raters' evaluations of grammatical complexity: school 4***

	number of main clauses	number of coordinate clauses	number of subordinate clauses	number of main clauses per 100 words	number of coordinate clauses per 100 words	number of subordinate clauses per 100 words
grammatical complexity	.266*	.413**	.520**	-.376**	.223*	.270*

N = 150
** Significant at p < .01*
***Significant at p < .001*

The strongest correlation with raters' judgements of *grammatical complexity* was with number of subordinate clauses. Number of subordinate clauses per 100 words also showed the strongest positive relationship with *grammatical complexity* out of the three ratio measures — numbers of main, coordinate and subordinate clauses per 100 words of running text. This second correlation coefficient was, however, much lower than that between *grammatical complexity* judgements and overall production of subordinate clauses. Whilst the raters may have been influenced by *proportion* of subordinate clauses, they may have been even more influenced by sheer numbers of these. Number and ratio of coordinate clauses showed a similar pattern, although these did not correlate quite so highly with *grammatical complexity* judgements as did subordinate clauses, in terms of either relative frequency or overall production, and may have been less influential.

Although number of main clauses showed some relationship with *grammatical complexity* judgements, the proportional use of these had a negative correlation with these judgements. This negative correlation may be taken as duplicating evidence of the positive relationship between *grammatical complexity* judgements and relative frequency of coordinate and subordinate clauses. The higher frequency per 100 words of one type of clause depended not only on its own increase in numbers, but also on the lower frequency of the two other types. The more main clauses per 100 words, the fewer coordinate and subordinate clauses, and vice versa. A negative correlation between relative frequency of main clauses and *grammatical complexity* judgements implies that as the relative frequency of main clauses went up — and relative frequencies of the other clause types correspondingly fell — *grammatical complexity* judgements came down; as the relative frequency of main clauses went down — and relative frequencies of the other two clause types rose — *grammatical complexity* judgements went up.

This situation did not hold for absolute numbers of clauses, since increased production of one clause type did not automatically depress numbers of the others, and in an open-ended task a student may have simply increased production of all three types (although the limited time available for production may, in practice, have resulted in some trade-off effects between clause types). Number of main clauses must still hold some relevance for *grammatical complexity*, since subordinate and coordinate clauses cannot exist without main clauses. Number of main clauses necessarily holds a relationship with text length, which, in

turn, may hold some relationship with numbers of coordinate and subordinate clauses, which, in *their* turn, may influence *grammatical complexity* judgements. The weaker correlation coefficient of .266 between *grammatical complexity* judgements and number of main clauses (as compared to correlations between *grammatical complexity* judgements and numbers of coordinate and subordinate clauses) may have been a side-effect of these other relationships, and may have simply represented the minimum relationship possible in a data set of this kind, given the interdependence of the three variables of main, coordinate and subordinate clauses.

6.11 Measures of accuracy

6.11.1 Error-free T-units

Quantity of grammatically correct output was measured in number of error-free T-units and in number of words contained in error-free T-units (Table 6.39). Within that output, the consistency of grammatical accuracy was measured by mean number of error-free T-units per 100 words and mean number of words contained in error-free T-units per 100 words (Table 6.41). Distributions for all four of these variables are given in Appendix 16.

Table 6.39 ***Mean numbers of error-free T-units and words contained in error-free T-units per composition: school 4; four classes***

	lower level control class: N = 34		lower level experimental class: N = 33		higher level control class: N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
N error-free T-units	7.47	5.97	12.55	8.04	14.80	6.72	15.81	7.84
N words contained in error-free T-units	46.38	39.61	76.27	55.61	101.10	40.39	107.79	50.18

For the lower-level pair of classes, differences between control and experimental students' production of both number of error-free T-units and number of words contained within those units were quite dramatic, and statistically significant (Table 6.40). For the higher-level pair of classes, however, differences were slight.

Table 6.40 ***Results of independent-samples t-tests for number of error-free T-units and number of words contained in error-free T-units: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
N error-free T-units	2.939	.005	.630	.530
N words contained in error-free T-units	2.540	.013	.666	.507

Not only did the lower-level experimental group produce more error-free T-units and greater numbers of words contained within those units, but, whilst so doing, they also showed a higher level of relative accuracy, as measured by numbers of, and numbers of words contained in, error-free T-units per 100 words (Table 6.41). They produced more, and relatively more grammatically accurate, text. They did not, however, attain the same level as the higher-level control group.

The higher-level experimental group, whilst producing slightly more error-free T-units, with more words contained in these, showed a *decrease* in relative accuracy. In other words, although the students produced slightly more text which was grammatically accurate than their control counterparts, this was not enough to counteract the effect on overall relative accuracy of additional text produced which was not accurate.

Table 6.41 ***Mean numbers of error-free T-units and words contained in error-free T-units per 100 words of text: school 4; four classes***

	lower level control class N = 34		lower level experimental class: N = 33		higher level control class N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
error-free T-units per 100 words	3.04	2.21	4.52	2.33	5.57	2.06	4.81	2.15
N words contained in error-free T-units per 100 words	18.88	14.55	27.62	17.70	38.28	13.26	32.87	13.81

Whilst the increase in consistency of grammatical accuracy of the lower-level experimental group was statistically significant, the *de*crease in consistency of grammatical accuracy of the higher-level experimental group was not. T-values for differences between mean numbers of error-free T-units and words contained within these, per 100 words of text, are given in Table 6.42.

Table 6.42 ***Results of independent-samples t-tests for number of error-free T-units per 100 words and number of words contained in error-free T-units per 100 words: school 4; two levels***

	lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
	t-value	significance	t-value	significance
N error-free T-units per 100 words	2.655	.010	1.648	.103
N words contained in error-free T-units per 100 words	2.210	.031	1.816	.073

The observed differences between control and experimental compositions in consistency of grammatical accuracy mirrored the raters' judgements. Whilst raters did not find a significant difference for *grammatical accuracy* between the lower-level pair of classes, they did observe a gain in favour of the experimental class which approached significance (t = 1.916; p = .060: *cf.* Table 6.11). Similarly, for the higher-level pair of classes, raters observed a slight, non-significant *de*crease in *grammatical accuracy* for the experimental class (*cf.* Table 6.12). That results obtained from objective measurement of relative grammatical accuracy, via error-free T-units, were consistent with raters' *grammatical accuracy* judgements suggested that these judgements reflected *actual*, proportional levels of grammatical accuracy and that, at the higher level, lower rater scores on *grammatical accuracy* for the experimental group were not simply an effect of increased length which permitted increased display of error (*cf.* Section 6.8).

Although the increased output of the higher-level experimental students may have had a detrimental effect on *relative* levels of accuracy, this did not extend to *absolute* levels of accurate text production, since the total amount of error-free production was, in fact, very slightly increased. With reference to Table 6.43, below, one might say that a lower-level experimental student produced on average, compared to his control peers, three more T-units, all of which were error-free. In addition, two previously *with-error* T-units also

became error-free, giving a total of five more error-free T-units, three of which represented an increase in output. A higher-level experimental student produced on average seven more T-units than his control peers, only one of which was error-free. In terms of words, a lower-level experimental student produced approximately 40 words more than a student in the matching control class, 30 of which were contained in error-free T-units, whilst a higher-level experimental student produced nearly 70 words more than his control counterparts, only six or seven of which were contained in error-free T-units. (The apparent discrepancies between additional numbers of error-free T-units and increased amount of error-free text are caused by the fact that error-free T-units were typically much shorter than *with-error* T-units. It should also be noted that mean length of T-unit was different for each group.)

Table 6.43 ***Mean numbers of T-units, error-free T-units, words and words contained in error-free T-units: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
mean number of T-units	32.27	35.91	35.23	42.18
mean number of error-free T-units	7.47	12.55	14.80	15.81
mean number of words*	235.5	272.1	267.0	333.1
mean number of words contained in error-free T-units	46.38	76.27	101.10	107.79

**Mean numbers of words are derived from the transformation described in Section 6.8*

It has been recognised in the literature that there may often be some tension between accuracy, complexity and fluency, and that trade-offs between these might not be unexpected (Skehan and Foster, 1999). Whilst the figures in Table 6.43 show that increased output, or production fluency, effected no trade-off in absolute grammatical accuracy, the figures in Table 6.41 strongly suggest that there was some trade-off in relative accuracy for the higher-level group. Whereas the lower-level experimental students showed a more all-round improvement — in accuracy, and, as previously discussed, complexity, with nonetheless some increase in output — the higher-level experimental students showed a much more dramatic increase in output, maintaining relative complexity, but with a slight deterioration of overall relative accuracy.

It is possible that results, and observed improvements in accuracy and complexity, were subject to effects of the language learning curve. It is easier to make significant proportional gains at lower levels than at higher levels, since any absolute gain achieved at a lower level must be greater by comparison to the original small proficiency than the same gain would be by comparison to a much larger original proficiency. (Most teaching methods have more immediately visible effects at lower levels.) The non-significant improvement of the higher-level experimental students in complexity may have represented as much effort as the significant improvement of the lower-level experimental students. That absolute accuracy remained stable for the higher-level group may have been because accuracy required much more effort to produce a visible gain at the higher level than at the lower level. Fluency, on the other hand, may not be subject to the same type of learning curve. A minimum level of proficiency is necessary before fluency may even begin to develop. The more language a learner has at his command, the more potential there is to develop fluency. In fact, fluency derives not so much from *learning* as from *practice*. It may well, under ideal conditions, have a progress curve which is J-shaped, which is to say quite flat to begin with, but becoming gradually steeper as more and more practice is accumulated. This may be one reason why the higher-level experimental group exhibited a much greater gain in fluency than the lower-level experimental group.

6.11.2 Relationship of error-free T-unit measures to raters' evaluations

The four objective error-free T-unit measures correlated well with the raters' judgements of *grammatical accuracy* (Table 6.44), validating both kinds of measurement.

Table 6.44 ***Correlations between error-free T-unit measures and raters' judgements: school 4***

raters' judgements	N error-free T-units	N words contained in error-free T-units	N error-free T-units per 100 words	N words contained in error-free T-units per 100 words
grammatical accuracy	.651	.720	.620	.689
grammatical complexity	.505	.598	.343	.439
coherence and flow	.586	.640	.463	.519
overall quality	.667	.736	.489	.589

N = 150

Correlation coefficients obtained between *grammatical accuracy* judgements and the objective accuracy measures were considerably higher than those obtained between *grammatical complexity* judgements and objective complexity measures (*cf.* Tables 6.25, 6.29 and 6.38). Two reasons might be suggested for this. The raters, who were all classroom teachers, may have had a more immediate sense of, and been better judges of, textual accuracy than textual complexity, given that the traditional task of teachers is more often to point out error than to point out simplicity of structure. The choice of level of complexity of production is most often left to the student, and the teacher then works to help the student within that choice. Weighing against this possibility, however, was the relatively high inter-rater reliability obtained for *grammatical complexity*, which was very nearly the same as that obtained for *grammatical accuracy* (*cf.* Table 6.13). Had raters not been dependable judges of complexity, they would not have obtained this level of reliability.

The second, more likely, possibility was that grammatical accuracy, notwithstanding the countless possibilities of types and combinations of error, is, at its most basic level, a very simple, almost binary, construct. A verb tense is either appropriately used, or it is wrong; a sentence is either correct, or it is incorrect. This is not to say that at higher proficiency levels there may not be some disagreement amongst judges of "correctness". However, for low-level L2 texts such scope for disagreement is something of a luxury. Textual complexity, on the other hand, is not so easy to deconstruct, and may be the product of a variety of elements. Sentence length, clause ratios and clause types may all have an association with complexity, but so may sophistication of other grammatical structures, including verb tense and structures, noun phrases, amount of modification and of reference. Error-free T-units may, in fact, simply have been a much better, and much more complete, measure of grammatical accuracy than lengths of sentences and T-units and ratios of clause-types were of grammatical complexity.

Error-free T-unit measures also show a strong association with *overall quality* judgements, particularly number of words contained in error-free T-units. In fact, both number of error-free T-units and number of words contained in error-free T-units show a stronger association with *overall quality* judgements than with *grammatical accuracy* judgements. However, *relative* numbers of error-free T-units, and numbers of words contained within these (*i.e.* per 100 words) show a considerably stronger association with *grammatical accuracy* judgements

than with *overall quality* judgements. This may mean that the strength of association of *absolute* numbers of error-free T-units and words contained in these with *overall quality* judgements is enhanced by the known relationship between *overall quality* judgements and composition length (*cf.* Table 6.20). Error-free T-unit measures also have a relatively strong association with *grammatical complexity* and *coherence and flow* judgements. Again, the association is stronger for absolute numbers of these error-free T-unit measures than it is for relative consistency of these, suggesting, similarly, that composition length may have been an influential intervening factor.

6.12 Spelling mistakes

6.12.1 Number of spelling mistakes

Each composition had a *types* and a *tokens* count of misspelt words (*cf.* Section 4.3.6.4). Although raters' assessments of *spelling* did not produce normally distributed data, numbers of *types* of actual mistakes did. (Distributions for numbers of spelling mistakes *types* and *tokens* are to be found in Appendix 17.) At the lower level, the experimental students made slightly fewer mistakes than the control students, and at the higher level the experimental students made slightly *more* mistakes than their control counterparts (Table 6.45). These differences were not significant. As has already been discussed (*cf.* Section 5.4), spelling mistakes were relatively rare in the data.

Table 6.45 ***Mean numbers of spelling mistakes (types and tokens) per composition: school 4; four classes***

	lower level control class: N = 34		lower level experimental class: N = 33		higher level control class: N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
spelling mistakes *types*	3.32	2.23	2.70	1.81	2.50	2.01	3.07	2.33
spelling mistakes *tokens*	3.91	2.78	3.70	2.91	3.12	2.74	4.12	3.34

Since students produced compositions of unequal length, numbers of spelling mistakes per 100 words, for both types and tokens, were also calculated (Table 6.46). This calculation

showed that, although the higher-level experimental students made slightly more mistakes, in terms of raw numbers, than their control counterparts, when these were taken as numbers of mistakes per 100 words, there was no difference between the two groups. Higher numbers of spelling mistakes for the experimental group were simply an effect of greater output. The lower-level experimental students, however, not only produced fewer spelling mistakes overall than their control counterparts, but also produced relatively fewer per 100 words. Although these effects were not statistically significant, they tend to support Krashen's contention that reading helps to improve spelling (Krashen, 1989, 1993), at least at this low level. Again, the higher-level experimental students appear to have favoured increased levels of output over increased levels of accuracy. However, at both levels, the overall low numbers of spelling mistakes, possibly due to the students' L1 and L2 instructional backgrounds (*cf.* Section 5.4), may have inhibited the emergence of any effects.

Table 6.46 ***Mean numbers of spelling mistakes (types and tokens) per 100 words: school 4; four classes***

	lower level control class N = 34		lower level experimental class: N = 33		higher level control class N = 40		higher level experimental class: N = 43	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
types of spelling mistakes per 100 words	1.46	.99	1.07	.87	.97	.82	.97	.79
tokens of spelling mistakes per 100 words	1.69	1.16	1.40	1.18	1.20	1.09	1.29	1.12

6.12.2 Relationship of numbers of spelling mistakes to raters' evaluations of *spelling*

Although raters reported difficulties in awarding *spelling* scores (*cf.* Section 5.4), Spearman's rho correlation coefficients between these and actual numbers of mistakes were very high (Table 6.47). (A Pearson correlation was not used since neither *spelling* scores nor number of *tokens* of spelling mistakes were normally distributed.) Raters may have been better judges of *spelling* than they felt themselves to be. Alternatively, correlations may be very high because, as they reported during debriefing sessions, two of the raters, as a result of the comparative scarcity of spelling mistakes, resorted to covertly estimating their actual numbers, making an informed adjustment for length of composition, and using this as a basis for awarding *spelling* scores.

All correlations are negative, since high *spelling* scores co-occur with low numbers of spelling mistakes. Numbers of *types* of spelling mistake were more influential on raters' judgements than numbers of *tokens*. Relative frequency was more influential than absolute frequency. These effects, are, however, relatively slight.

Table 6.47 ***Spearman rank-order correlations between numbers of spelling mistakes and raters' spelling judgements: school 4***

	N spelling mistakes *types*	N spelling mistakes *tokens*	spelling mistakes (*types*) per 100 words	spelling mistakes (*tokens*) per 100 words
spelling scores	-.823	-.763	-.845	-.823

N = 150
All correlations are significant at p < .001

6.13 Past tense verb forms

Table 6.48 shows the total number of verb uses by each class. These include finite and infinite forms. The object of this part of the study was to investigate whether students' participation in the reading scheme had impacted on accuracy of simple past verb forms. If irregular verb forms are more subject to frequency effects than regular verb forms (Ullman, 2001a), then repeated exposure to instances of the simple past in the narrative texts of the graded readers used by the ERS might be expected to have more effect on accuracy of production of *irr*egular simple past forms.

Only verbs used within narrative text were investigated, for both practical and theoretical reasons. Directly reported dialogue changes the time- and discourse-frames of the text, inviting a much more varied use of verb forms. Reported dialogue was therefore not considered rich data for the analysis. Additionally, directly reporting speech, even if it is imagined speech, may be considered a different task-type from narrative story-telling and may engage the learner in different behaviours. Learner production in general is known to be inherently variable (Romaine, 2003; Ellis and Barkhuizen, 2005), and consistency of produced morphological forms has been found to be affected by task-type, task-difficulty, situational context and degree of spontaneity of production. Pienemann (1998), for example

found that correct marking of the plural *-s* for the same learner varied across six time-controlled tasks, from 100% correct on a "spot the difference" picture task, to only 29% correct on a picture sequencing task. Correct 3rd person singular *-s* marking varied across the same six tasks from 26% to 0% correct. Ellis (1987) found that production of simple past *-ed* varied under different planning conditions. Reproduction of simple dialogue might be easier for some students than explaining events in a story, or may provoke more, or less, spontaneous text. Changes in consistency of correct past tense morphology caused by this, whilst an interesting subject of study, was not the object of this research. Thus, to achieve maximum clarity of results, only narrative text was studied. As may be seen from Table 6.48, narrative text contained considerably greater numbers of verb uses than did reported dialogue.

Table 6.48 Numbers of verb uses: school 4; four classes

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
total verb uses	1,555	1,783	2,106	2,773
verb uses within directly reported dialogue	154	250	192	440
verb uses within narrative text	1,401	1,533	1,914	2,333

Verbs contained in narrative text were coded as one of eight primary types. Five are classed as past forms. These were: irregular simple pasts, regular simple pasts, simple past BE, "simple past" modal structures (*e.g. could see*) and simple past passives. The other three categories were: infinitives, other tenses, which included not only present tenses but other past forms such as past continuous and past perfect, and uncoded verbs. Infinitives were mainly used in complement clauses, (*e.g.* "I wanted *to find my friend*") and subordinate clauses of purpose (e.g. "We went in the shop *to buy a lunch*"). Verbs were classed as *uncoded* when it was unclear what form the student had attempted. Table 6.49 shows numbers and percentages of uses of each primary type by each class. Figures given include affirmatives, negatives and interrogatives.

Table 6.49 ***Verb use within narrative text: school 4; four classes***

verb classification	lower level control class: N = 34		lower level experimental class: N = 33		higher level control class: N = 40		higher level experimental class: N = 43	
	N	%	N	%	N	%	N	%
irregular simple past	526	37.54%	644	42.01%	697	36.42%	885	37.93%
regular simple past	258	18.41%	331	21.59%	411	21.47%	483	20.70%
simple past BE	273	19.49%	202	13.18%	344	17.97%	415	17.79%
simple past modal structures	103	7.35%	91	5.94%	112	5.85%	131	5.62%
simple past passives	4	0.29%	14	0.91%	23	1.20%	22	0.94%
infinitives	101	7.21%	131	8.55%	148	7.73%	174	7.46%
other tenses	45	3.21%	68	4.44%	118	6.17%	138	5.92%
uncoded	91	6.50%	52	3.39%	61	3.19%	85	3.64%
total verb uses within narrative text	1,401		1,533		1,914		2,333	

Note: percentage figures may not add up to exactly 100 because of rounding to two decimal points.

Irregular simple pasts were more frequent than regular simple pasts, matching Pinker's observation that "the 13 most frequent verbs in English — *be*, *have*, *do*, *say*, *make*, *go*, *take*, *come*, *see*, *get*, *know*, *give*, *find* — are all irregular" (1991: 532), although *be*, for this research, was not included in the irregular verb count, but was a category by itself. Altenberg and Granger (2001), using corpus-based information, estimated that, disregarding modals and auxiliaries, of the 15 high frequency verbs most commonly topping any frequency list, only two were regular.

Distribution patterns for the two higher classes were very similar. However, for the lower-level pair of classes there were some notable differences. The non reading-scheme class demonstrated a higher percentage of *be* use and modal use, and a correspondingly lower use of regular and irregular simple pasts. Since the use of regular and irregular verbs represents *lexical* choice, this suggests that the experimental group had been facilitated in a move away from overuse of the copula *be*, towards greater lexical variety, by participation in the reading scheme. Likewise, overuse of modals is often a characteristic of L1 Chinese speakers' written L2 English text (Hinkel, 2002), and the reading scheme may have helped students to

move away from this. This interpretation is given weight by the fact that the percentage of modal use for the lower-level experimental group was very similar to that for both the higher-level groups, suggesting an approximate optimal ratio for this kind of text at intermediate level. These two propositions, however, can only be tentative. The higher frequency of uncoded verbs in the text of the lower-level non reading-scheme students can be accounted for by the greater occurrence of ambiguous or unidentifiable verb uses.

Correct to incorrect ratios of regular and irregular simple past, past *be* and past modal forms are shown in Tables 6.50 to 6.53. Only declaratives were considered in the calculations for regular and irregular simple pasts, since negatives and interrogatives require the use of *did*, leaving the verb lemma unchanged, and, in any case, the formula is the same for both regular and irregular verbs. For these lexical verbs, declaratives were the most frequent form in the data, accounting for 94.7% and 94.1% of irregular and regular simple past use respectively, and so numbers were still sufficient to permit a clear comparison. To maintain numbers of modal and *be* forms, negatives and interrogatives were included. For these, the past form does not revert to root form for interrogatives and negatives, with the use of *did*, but is constant across the three types of use. The use of passives was too infrequent to afford any reliable or generalizable information concerning consistency of accuracy of form.

Table 6.50 ***Correct and incorrect uses of irregular simple past declaratives: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
correct uses	382	513	589	782
incorrect uses	117	84	72	66
percentage correct	76.6%	85.9%	89.1%	92.2%

Table 6.51 ***Correct and incorrect uses of regular simple past declaratives: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
correct uses	137	259	334	370
incorrect uses	104	56	49	86
percentage correct	56.8%	82.2%	87.2%	81.1%

Table 6.52 ***Correct and incorrect uses of past BE: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
correct uses	185	141	305	356
incorrect uses	88	61	39	59
percentage correct	67.8%	69.8%	88.7%	85.8%

Table 6.53 ***Correct and incorrect uses of past modal forms: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
correct uses	23	36	60	69
incorrect uses	80	55	52	62
percentage correct	22.3%	39.6%	53.6%	52.7%

This kind of data is not susceptible to significance testing, since this depends on scores per individual composition. Correct-to-incorrect ratio measures for individual cases, such as percentages of correct verb forms per composition, are unsafe measures of comparison when production is not obligatory. For example, one student might produce two instances of a past modal structure, and another might produce ten such instances. If the first student has produced two *correct* instances, and the second student has produced nine correct instances, technically, using a correct-to-incorrrect ratio measure, the first student has outperformed the second student, with 100% correct production compared to 90% correct production. This is not a safe comparison, and would seriously undermine the validity of any results obtained from between-groups significance testing.

However, from the figures in Tables 6.50 to 6.53, it is clear that several effects took place, between lower and higher proficiency levels and between non reading-scheme and reading-scheme students. All four classes showed a better command of irregular than of regular simple past. This learning pattern has been so commonly found in both text-based and oral-production L2 acquisition research that it has become part of a putative model of the

hierarchy of development of verb morphemes in English L2 acquisition (Ellis, 1994). Researchers have proposed various reasons for this, ranging from the simple impact of frequency to biological programming. One of the most convincing explanations is that forwarded by Goldschneider and DeKeyser (2001) who propose that the acquisition order results largely, though not entirely, from a combination of factors which include perceptual salience, semantic complexity, morphophonological regularity, or *irr*egularity, syntactic category, and frequency.

For the lower-level pair of classes, the experimental group showed greater accuracy than the control group for all four past forms, only marginally for past form *be*, but dramatically for the other three types, particularly the regular simple past, with 82.2% correct production, compared to 56.8%. Contrary to the hypothesis, the reading scheme, whilst, indeed, having the expected effect on irregular simple past, had an even greater effect on regular simple past.

The higher-level experimental class showed greater accuracy than its matched control class only in the irregular simple past. The other three past forms all showed *lower* percentages of accurate production. These lower accuracy ratios may be compared to the observed lower level of relative accuracy for this group using error-free T-unit measures (*cf.* Table 6.41), and may be partly a result of having produced much longer compositions. For these three past form types, experimental students produced, in raw numbers, more correct instances than control students. However, they also produced similar additional numbers of *in*correct instances. With both groups producing, overall, much greater numbers of correct than incorrect instances, the resulting *proportional* increase in numbers of correct instances for the experimental group was considerably smaller than the proportional increase in numbers of incorrect instances. In other words, additional numbers of incorrect instances had a more powerful effect on the correct-to-incorrect ratio than similar, or slightly higher, additional numbers of correct instances. That *irregular* simple past did *not* conform to this pattern, but demonstrated an increase in number of correct forms produced by the experimental group accompanied by a *de*crease in number of incorrect forms, suggests the reading scheme may also have had a beneficial effect on accuracy of irregular simple past at the higher level.

The decrease in proportional accuracy was more marked for regular simple pasts than for modals or *be*, suggesting other factors may also have been in play. Bardovi-Harlig (1995)

found that, for a group of intermediate learners, accurate simple past forms in narrative text were more prevalent in foreground than in background text. More generally, Bardovi-Harlig's findings may be taken as supporting evidence for the influential *aspect* hypothesis (Anderson, 1991), which proposes that the semantic function of a verb plays a major role in the acquisition of the simple past, and that accurate use of past-tense morphology emerges in verbs which have an inherent end point, or verbs of *achievement*, before it emerges in verbs denoting processes of no inherent end point, and finally in stative verbs.

Narrative accounts of what might loosely be described as adventure stories may be expected to contain high numbers of *achievement* actions. Shorter compositions may, in order to accommodate the complete story in fewer words, have a stricter focus on actual events, hence a more concentrated use of achievement verbs; longer compositions may have more diversity of semantic verb types, resulting in a lower proportion of those verb types which achieve the past tense more easily. In Bardovi-Harlig's terms, the higher-level experimental students, as part of producing longer stories, may have engaged in more backgrounding, producing more of the "supportive material that elaborates on or evaluates the events in the foreground" (Bardovi-Harlig, 1995: 266), providing, specifically, more orientation, evaluation or explanation for main narrative events.

Housen found that the aspect hypothesis, if valid, applies more verifiably to regular verbs, the emerging past tense of which shows "a significantly stronger link with inherent aspect than irregular [past tense]" (2002: 107). Salaberry has suggested that "the effect of inherent lexical aspect (classes of verb phrases) may be independent of the effect of the cognitive saliency of irregular morphology" (2000: 145). If the higher-level experimental students *did* engage in more backgrounding, and used a higher proportion of the type of verbs which take longer to achieve accuracy of past-tense use, this could explain why an effect was apparent for regular verbs, but not for irregular verbs, which may be impervious to aspectual conditions.

Other researchers have found that, in any case, the regular simple past is more prone to destabilization effects than irregular simple past. Ellis (1987) found that it was not uncommon for a regular simple past to be produced accurately on one occasion and subsequently produced *in*accurately on another. An illustration of this is given by a student's production of "It contained a big snake" in one task, and later production, under different

planning conditions, of "The basket contain a snake". Ellis concluded that style shifting (reverting incorrectly to the root form in obligatory past tense context), as a result of insufficient planning time, was "most clearly evident in regular past tense forms, less so in the past copula, and hardly at all in irregular past tense forms" (1987: 10). As has been observed by Goldschneider and DeKeyser (2001), a possible reason for this is that similarity of form between grammatical units, as, for example, in *like* and *liked*, obscures the one-to-one relationship between form and meaning, and production of the less safe form is destabilized under conditions of stress. The saliency of the differences between root and simple past forms inhibits this effect in *irr*egular verbs.

Since 78.8% of the inaccurate regular simple past forms produced by the higher-level experimental group consisted of a reversion to root form, what was manifested as "incorrect" regular past tense may have been, not an effect of use of the type of verb which, because of its semantic aspect, is more resistant to acquisition of the past form, but, more simply, style shifting of the kind noted by Ellis. This may have been caused by what was, in effect, a reduction in planning time, per language item, occasioned by the production of much longer compositions.

6.14 Vocabulary measures

6.14.1 Lexical sophistication

Table 6.54 gives means and standard deviations for numbers of types and tokens used by each class at four frequency levels identified by the computer programme Web VocabProfile (Cobb, *n.d.*). These frequency ratings are those of Laufer and Nation's Lexical Frequency Profile (1995), based, in turn, on West's (1953) General Service List. *Second 1000* refers to words which are placed on the frequency scale after the first thousand most frequently used words, up to the 2000th word. Words *beyond* the first 2000 are included in the *other* category. VocabProfile, in fact, makes a distinction between words which are beyond the first 2000 most frequently used and which, referring partly to the University Word List (Xue and Nation, 1984) and partly to the Academic Word List (Coxhead, 2000), may be considered *academic,* and words which are merely beyond the first 2000. Since there were very few instances of either of these in the data, accounting jointly for just over 1% of the total number of tokens produced, with a mean frequency of 2.36 types per composition

between them, numbers were combined and are presented here as *other*.

Table 6.54 ***VocabProfile analysis; mean numbers of types and tokens at different lexical frequency levels: school 4; four classes***

		lower level control class N = 34		lower level experimental class: N = 33		higher level control class: N = 40		higher level experimental class: N = 43	
		mean	s.d.	**mean**	s.d.	**mean**	s.d.	**mean**	s.d.
first 500	types	76.9	13.3	85.2	14.7	86.6	14.0	99.2	17.5
	tokens	198.3	47.7	231.2	59.5	221.4	47.3	277.5	75.3
second 500	types	14.6	5.3	16.3	5.3	18.4	5.3	20.8	6.1
	tokens	21.3	8.4	24.5	10.4	25.6	8.5	31.9	9.9
second 1000	types	9.5	4.5	10.0	3.6	12.2	5.1	14.5	5.8
	tokens	13.1	6.6	14.1	5.5	16.6	7.9	20.2	8.4
other	types	2.5	2.2	1.6	1.5	2.5	2.1	2.7	2.2
	tokens	2.7	2.5	2.4	2.4	3.3	2.7	3.5	3.1
total	types	103.5	19.3	113.3	21.1	119.9	20.8	137.3	25.5
	tokens	235.5	56.6	272.1	71.1	267.0	58.3	333.1	88.0

Note: numbers may not add up to exact totals reported because of rounding to one decimal point.

With two notable exceptions to the rule, there is a clear and consistent pattern of production, which is that, for each frequency band, experimental students produced greater numbers of both types and tokens than their counterpart control students, and higher-level control students produced more than lower-level experimental students. Thus there is a steady increase in production of all word categories across the four classes, suggesting that VocabProfile discriminates between proficiency levels and may also reveal differences caused by a treatment.

One exception to this pattern is that, for *first 500*, the lower-level experimental students produced more tokens, but not types, than the control students from the higher-proficiency pair of classes. As has already been noted (Section 6.8), the lower-level reading-scheme class wrote longer compositions than the higher-level non reading-scheme class. Figures from the VocabProfile analysis show that the overall higher output of the lower, experimental, group derived entirely from a proportionate, and also absolute, higher use of

words in this most-frequently-used category, since the higher level group produced more tokens than the lower-proficiency experimental class at every other VocabProfile level.

Some interesting questions may be raised here regarding the use of counts of high- and low-frequency words as a measure of vocabulary knowledge. There may not always be an unimpeded, direct relationship between these. Although the lower-level experimental group did not achieve a higher mean score for *grammatical complexity* than the higher-level control group (6.85 and 7.88 respectively), they did, however, produce longer sentences and T-units, with higher clauses-per-sentence and per-T-unit ratios, and more relative clauses. These objective measures may indicate a greater *syntactic* complexity. Since functors such as prepositions, conjunctions, subordinators, and relative and other pronouns are all very high frequency, a higher percentage of "low level" words in a text may be as much a result of a different syntactic structure as of limited vocabulary knowledge.

A second irregularity is that the lower-level control group produced more *other* vocabulary than its peer experimental group. Investigation into the distributions revealed that in the lower-level control class three outliers in the data produced 23 *other* types between them, out of a total of 84. Examination of the compositions in question further showed that five of these were contained in just one sentence: *My third wish was wanted to help the mentally and physically handicapped and donated plenty of money to the Community Chest.* Other sentences were: *In the forest reptiles and insects distribute everywhere*; *The house was very classical*; *Every room had their special style*. These sentences may have derived their rather specialist vocabulary from English-medium subject classes. The writer of the first may even have made slightly modified use of a memorised chunk from a history lesson on the life and times of some public figure, or from a school charity interest. Although this vocabulary cannot be discounted, this serves to illustrate how easily data which relies on such low overall counts of instances as was the case for *other* vocabulary can be skewed by a few unusual cases.

Table 6.55 gives the results of significance tests for differences in numbers of types and tokens produced by peer control and experimental classes, for each frequency band.

Table 6.55 ***Significance levels for differences in mean numbers of types and tokens at VocabProfile lexical frequency bands: school 4; two levels***

		lower level *N control = 34* *N experimental = 33*		higher level *N control = 40* *N experimental = 43*	
		t-value	significance	t-value	significance
first 500	types	2.419	.018	3.593	.001
	tokens	2.504	.015	4.028	.000
second 500	types	1.368	.176	1.898	.061
	tokens	1.370	.175	3.094	.003
second 1000	types	.561	.577	1.930	.057
	tokens	.638	.566	1.999	.049
other	types	1.822	.073	.321	.749
	tokens	.605	.547	.257	.798
totals	types	1.984	.051	3.403	.001
	tokens	2.323	.023	3.959	.000

For the lower-level pair of classes, the only significant difference was for the lowest band, with experimental students producing significantly more low-level vocabulary. Either the experimental students produced more high-frequency lexis or they used a wider variety of the functors which are tied to this level because of their syntactic importance to, hence frequency in, text. In terms of means, an experimental student used approximately eight more *first 500* types than his control counterparts, and it is conceivable that some of these, at least, may be functors. The difference seen in *totals* is simply a repeat of the same information.

For the higher-level pair of classes the difference in number of types was significant only at the lowest band. The differences in numbers of tokens were significant for all frequency bands except *other*. It is very difficult to interpret this data since one, or both, of two major constraining factors are likely to be in operation. Firstly, as Laufer (1998) points out, the LFP measures only *use*, which is to say what students *choose* to write. It does not measure vocabulary size. Investigating the relationship between passive, "controlled active" (elicited) and free active vocabulary knowledge, Laufer found that "in spite of an impressive increase in passive vocabulary and a good progress in controlled active vocabulary size, learners did not put this knowledge into use when left to their own choice of text" (1998: 266). In a more recent article, Laufer (2005) suggests that a small increase in vocabulary use may reflect a

large increase in vocabulary knowledge. If this is the case, then not only the higher-level but also the lower-level experimental students may have had considerably more vocabulary knowledge than their respective control peers, in all three of the GSL-based frequency bands. (The low numbers of *other* vocabulary make this band too disproportionately susceptible to chance fluctuations to be a reliable measure.)

Secondly, however, there is the fact that we are comparing compositions of unequal lengths. The higher numbers of total types produced by the experimental groups, in relation to their peer control groups, significant at the higher level but not at the lower level, *may* be strong evidence that the students have learned more words, or could just be a co-product of greater amount of text. In other words, the longer the composition, the more opportunity — indeed, need — to produce additional types, at *every* band level. In *this* case, the greater numbers of types may represent, not a difference in vocabulary knowledge, but greater production fluency.

Recasting raw numbers used at each band as percentages of total word use (Table 6.56) clarifies only a little. Here, the data suggests mainly a difference between proficiency levels, with the two higher-level classes showing a proportional decrease in use of *first 500* words, accompanied by a slight increase in use of *second 500* and *second 1000*. Between matched control and experimental classes any differences are less clear. At both levels, the experimental students used a slightly greater proportion of *first 500* words, similar proportions of *second 500*, and proportionately slightly fewer *second 1000* words. If this is a true reflection of proportional knowledge of higher and lower frequency lexis, it is strange that no effect was apparent at the middle level of *second 500*. Given that both experimental groups demonstrated greater syntactic complexity than their peer control groups (*cf.* Section 6.9), the higher percentages of *first 500* use for these two groups may be connected to syntactic structure differences. That any increase in one proportion must always be accompanied by a *de*crease in another, co-dependent, proportion may account for the proportional decrease in *second 1000* words.

Table 6.56 ***Percentage of total word use at VocabProfile frequency levels: school 4; four classes***

frequency level	lower level control class: N = 34	lower level experimental class: N = 33	higher level control class: N = 40	higher level experimental class: N = 43
1st 500	84.20%	84.97%	82.92%	83.31%
2nd 500	9.04%	9.0%	9.59%	9.58%
2nd 1000	5.56%	5.18%	6.22%	6.06%
other	1.15%	0.88%	1.24%	1.05%

Note: the category other is included only for information. It should not be regarded as a reliable indicator of group performance.

The second tool used to attempt to capture any changes in lexical sophistication was a 10-band Internal Frequency List constructed from the data itself (*cf.* Section 4.3.6.6). Table 6.57 gives the cumulative percentages of text coverage of the whole corpus for each frequency band and the number of types[1] used by each class within these bands.

Table 6.57 ***Numbers of types at each Internal Frequency band: school 4; four classes***

Band	0	1	2	3	4	5	6	7	8	9
	Cumulative % of text coverage for each band across the whole corpus of data									
	26.51%	44.49%	56.77%	65.35%	73.51%	83.33%	88.72%	92.34%	96.27%	100%
whole corpus	7	16	23	26	41	91	95	118	235	806
lower level control	7	16	23	26	41	91	89	90	141	215
lower level experimental	7	16	23	26	41	91	92	100	148	197
higher level control	7	16	23	26	41	90	95	114	170	310
higher level experimental	7	16	23	26	41	91	95	116	189	405

[1] Because of differences in the parameters of what constitutes a *type* (*cf.* Section 4.3.6.6), total numbers of types for each class using the Internal Frequency List will not match numbers obtained using VocabProfile.

What is striking about the information in Table 6.57 is, firstly, how so few types may account for such large proportions of text, and, secondly, differences in type use are only apparent in a very small percentage of the text. Between the two proficiency levels, a difference begins to emerge after approximately 83% of text coverage. For the lower-level matched control and experimental classes, a slight difference is seen in the last 10% of coverage. (It should be noted, however, that the 215 lowest-frequency types used by the control class include the 23 types found in the three outlier compositions, and so this figure may be slightly misrepresentative.) For the higher-level pair of classes, a difference only emerges in the last 8% of text coverage, although the only entirely clear difference is not observed until the last 4%.

This illustrates the difficulty of capturing differences in high and low-frequency vocabulary use using counts of words from frequency bands. A higher use of low-frequency words may occur only within a very small proportion of the text, since these words *are* low-frequency. A law of diminishing returns must apply, with increasingly infrequent words being found in an increasingly smaller and smaller portion of the text.

Table 6.58 sums up the information in Table 6.57, giving the total numbers of types for each class. As may be seen from Table 6.57, most of these types appear in Bands 5-9, with approximately half found in Bands 8 and 9.

Table 6.58 ***Total numbers of types within Internal Frequency List: school 4; four classes***

	lower level control class N = 34	lower level experimental class: N = 33	higher level control class N = 40	higher level experimental class: N = 43
N types	739	741	892	1009

It is clear from Table 6.58 that the higher-level control class produced more types than the lower-level experimental class, in spite of having produced slightly shorter compositions (Table 6.18). Although there are 40 students in the higher control class, compared to 33 in the lower experimental class, it is unlikely that seven students between them produced 151 new types. (In fact, as is seen in the next section, the average number of new types produced by a single student in that group was 3.7.) On this evidence, it is possible to conjecture that

the higher number of types produced by the upper-level *experimental* class may not simply be a corollary of the much longer compositions which *they* produced. In other words, the figures in Table 6.58 provide some evidence that increased number of types cannot be completely accounted for as no more than the unavoidable consequence of producing more text.

In terms of percentage of text coverage by *tokens* at each Internal Frequency band (Table 6.59), again there is no very clear pattern. At the lower level, there may be a slight tendency for the experimental group, compared to the control group, to use more vocabulary at bands 4 and 5, and slightly less at the bands below that. This may represent a slight shift in lexical sophistication for this group. At the higher level, there are no clear differences between control and experimental.

Table 6.59 ***Percentage of text coverage at each Internal Frequency band: school 4; four classes***

Band	0	1	2	3	4	5	6	7	8	9
cumulative % coverage of whole corpus	26.51%	44.49%	56.77%	65.35%	73.51%	83.33%	88.72%	92.34%	96.27%	100%
whole corpus	26.51%	17.98%	12.28%	8.58%	8.16%	9.82%	5.39%	3.62%	3.93%	3.74%
lower level control	26.05	19.47	13.43	9.33	7.22	8.84	5.0	3.0	4.04	3.61
lower level experimental	26.71	18.44	12.31	8.27	9.35	10.37	4.86	3.43	3.34	2.94
higher level control	26.34	18.01	11.55	8.53	8.02	9.73	5.53	3.9	4.29	4.11
higher level experimental	26.78	16.84	12.16	8.38	8.04	10.09	5.83	3.88	3.97	4.03

6.14.2 Lexical originality

Lexical originality was defined as the number of types unique to one group divided by the number of students in that group, and these figures are given in Table 6.60.

Table 6.60 ***Number of types unique to each group and lexical originality quotients: school 4; four classes***

	lower level control class: N = 34	lower level experimental class: N = 33	higher level control class: N = 40	higher level experimental class: N = 43
N types unique to group	96	80	148	240
Lexical originality quotient	2.8	2.4	3.7	5.6

Again, it is likely that results for the lower-level pair of classes do not represent an entirely accurate picture, since 23 words, produced by the three student outliers in the control class, and unique to that group, may have a disproportionate effect. For the higher-level pair of classes, however, the figures above represent the clearest evidence of a vocabulary effect for the reading scheme, since the difference between the lexical originality quotients of control and experimental groups, unlike numbers of types produced (*cf.* Table 6.58), is proportionately much greater than the difference in amount of text produced. The average length of a control composition at this level was 267 words. The average length of an experimental composition was 333 words, representing a 24.7% higher output. The lexical *originality* quotient of the experimental group, however, is just over 51% higher than that of the control group. This increased lexical originality is less likely than increased numbers of types to be simply an unavoidable, and directly proportional, result of having used more words.

6.14.3 Relationship between objective measures of lexical sophistication and raters' evaluations of vocabulary range

Raters' judgements of *vocabulary range* correlated moderately well with numbers of types and tokens at all VocabProfile bands (Table 6.61). At each band, apart from *other*, the correlation was slightly higher between rater judgements and number of types than number of tokens. Raters may have been marginally more sensitive to types than tokens, discounting, to some small extent, repetitions of words. What is odd, however, is that, as relative infrequency goes up, correlations with raters' *vocabulary range* judgements steadily decrease, and *first 500* appears to have had most effect on judgements, with words beyond the first 2000, the *other* category, having the least effect.

Table 6.61 ***Correlations between raters' vocabulary range judgements and N tokens and types at VocabProfile frequency levels: school 4***

	first 500		second 500		second 1000		other	
	N tokens	N types	N tokens	N types	N tokens	N types	N tokens	N types
vocabulary range judgements	.558	.609	.551	.557	.421	.439	.419	.386

N = 150
All correlations are significant at p < .001

Correlations between *vocabulary range* judgements and numbers of tokens at each Internal Frequency band are more consistent (Table 6.62). Correlations were not calculated between raters' judgements and numbers of *types* at each Internal Frequency band, since the number of types at each band was all but constant for all the compositions, across the four classes, for *bands 0* to *3*. Nor was there much variety, across individual compositions, in numbers of *bands 4* and *5* types.

Table 6.62 ***Pearson correlations between raters' vocabulary range judgements and N tokens at Internal Frequency bands: school 4***

	N tokens at band:									
Band	0	1	2	3	4	5	6	7	8	9
vocabulary range judgements	.444	.380	.386	.408	.462	.486	.416	.414	.389	.564

N = 150
All correlations are significant at p < .001

After an initial relatively high correlation with number of *band 0* words, the correlation between raters' judgements and number of words at a given band starts at .380 and steadily increases as relative infrequency of words also increases. After a slight dip, at *bands 6*, *7*, and *8*, the highest correlation is obtained with the lowest-frequency band. This is the highest of all the correlations obtained between rater judgements on *vocabulary range* and number of tokens at a frequency band for both this frequency list and VocabProfile. The Internal Frequency List may thus be seen to have provided the more reliable tool. The relatively high correlation it produced between rater judgements and numbers of words at its lowest band may be explained by an indirect relationship. The seven types which account for this band,

and for over 25% of all text in the corpus, include *the*, *a*, *we* and *I*. (Appendix 27 gives the full Internal Frequency List.) This monopoly of the two main articles, and the two first person subject pronouns, in a first person narrative, guarantees a strong relationship with numbers of nouns and lexical verbs. Thus the correlation between rater judgements and *band 0* words is not because of these words themselves but because their numbers are related to numbers of adjoining lexical items.

6.15 Summary of findings for school 4

Table 6.63, below and overleaf, sums up the main findings from school 4. Where a significance level (p value) is given, significance was in favour of the experimental group. Additionally, both experimental classes used more subordinate clause reduction (ellipsis) than their control counterparts, although this was not tested for significance.

Table 6.63 Summary of main findings for school 4: two levels

	Lower level: *N = 34 control; 33 experimental*	Higher level: *N = 40 control; 43 experimental*
Rater judgements		
Overall quality	p = . 019	n.s.
Grammatical accuracy	n.s.	n.s.
Grammatical complexity	p = . 002	n.s.
Vocabulary range	n.s.	n.s.
Coherence and flow	p = . 003	n.s.
Punctuation & paragraphing	n.s.	n.s.
Spelling	n.s.	n.s.
Objective measures		
Fluency measures		
number of words	p = . 023	p = . 000
number of T-units	n.s.	p = . 006
number of clauses	p = . 006	p = . 001

Syntactic complexity measures		
words per sentence (integral text)	p = . 040	n.s.
words per sentence (narrative text)	p = . 014	n.s.
words per T-unit	n.s.	n.s.
words per clause	n.s.	n.s.
clauses per T-unit	p = . 002	n.s.
clauses per sentence	p = . 000	n.s.
Clause types		
coordinate clauses per 100 words	p = . 001	n.s.
subordinate clauses per 100 words	p = . 018	n.s.
number of relative clauses	p = . 022	p = . 050
Accuracy measures		
number of error-free T-units	p = . 005	n.s.
words in error-free T-units	p = . 013	n.s.
error-free T-units per 100 words	p = . 010	n.s.
words in error-free T-units per 100 words	p = . 031	n.s.
number of spelling mistakes	n.s.	n.s.

% Correct simple past verb forms

	control group	experimental group	control group	experimental group
irregular declarative	76.6%	85.9%	89.1%	92.2%
regular declarative	56.8%	82.2.%	87.2%	81.1%
past BE	67.8%	69.8%	88.7%	85.8%
"past" modals	22.3%	39.6%	53.6%	52.7%
	not significance tested		not significance tested	

Vocabulary measures

VocabProfile	unclear results	unclear results
Internal Frequency List	slight shift towards relatively less frequent vocabulary for experimental group	unclear results: experimental produced more types but wrote longer compositions
Lexical originality (words unique to group per student)	control 2.8 / experimental 2.4 not significance tested	control 3.7 / experimental 5.6 not significance tested

n.s. = not significant
all reported significance is in favour of the experimental group

7. DISCUSSION

7.1 Reliability of findings

A potential weakness of the present study at the outset was the absence of pre-test scores. The research design of the larger ERS evaluation project precluded experimental pre-testing, since reading-scheme and non reading-scheme cohorts were separated in the education system by one year. The admission of a school into the ERS meant that, at the same time as Secondary 1 students became experimental students, the students of the previous year's Secondary 1 became control students with one year of the control teaching method (*i.e.* normal English classes) already completed. Comparability of control and experimental groups in this study, as in the large-scale evaluation project, thus relies on the homogeneity of a school's intake and the strictness of its admission and streaming procedures.

As a general rule, the student population of any Hong Kong secondary school, from one intake to the next, is homogeneous in the extreme, and the research design was approved by senior officials and academics at the Hong Kong Institute of Language in Education with many years of field experience. One reason for the homogeneity may be the large number of secondary schools (approximately 500, of which 80% are government-funded), which permits a very fine demarcation of ability levels amongst schools. The differences between schools are created and maintained incidentally by the schools themselves, as a by-product of their selection procedures. Because of the highly competitive ethos of the education system, schools are anxious to accept the highest-performing students they can. Thus, year after year, the highest-ranking schools unfailingly enrol the highest-ability students, the next best schools the next best students, and so on down the line. Geographical location is not a factor in parents' choice of secondary school. A school's intake might only differ significantly from that of the previous year if the whole Hong Kong student intake for that year was significantly different — a highly unlikely situation with such a large population.

Results from the evaluation of the compositions confirm the comparability of control and experimental groups in this study. Most of the measures used, including raters' judgements, objective measures and vocabulary profiles, clearly differentiate between the lower- and higher-level control classes of school 4, which is to say two groups of different ability but receiving the same instructional treatment. Appendix 28 presents a comparison of the groups, giving also the results of independent-samples t-tests where significance testing is

appropriate. Apart from *spelling* — which did not prove to be a particularly appropriate variable for the data set, perhaps due to the specific instructional setting; *cf.* Section 5.4 — the only variables which do not show a significant difference in performance between ability levels are those representing objectively measured syntactic complexity, namely total number of subordinate clauses, number of subordinate clauses per 100 words, sentence length (for narrative text only) and T-unit length.

We might expect, then, that a higher-ability student would do better than a lower-ability student on *every* measure except objective syntactic complexity measures, and had experimental students uniformly outperformed their control peers across *all* these other measures, we may have had grounds for wondering whether this was due to participation in the reading scheme or whether both reading-scheme classes were simply — even if improbably — of initial higher ability than the corresponding control group. However, those measures which differentiated between ability levels only did so between same-level control and experimental classes for *certain* constructs. At the lower level, raters perceived significant differences between control and experimental groups for *coherence* and *complexity* but not for *accuracy* or *vocabulary range*. At the higher level, raters perceived no significant differences between groups, but the experimental class strikingly outperformed the control class in fluency of production. Moreover, although objective syntactic complexity measures showed a significant difference between control and experimental groups at the lower level, these same measures were the only variables which did *not* discriminate between lower- and higher-level control groups. That the pattern of differences between same-level reading-scheme and non reading-scheme students does not resemble the pattern of differences between lower- and higher-ability groups shows that the former does not represent normal level or proficiency differences (within the particular educational setting), and that something other than normal classroom-led progress has taken place. In fact, as we shall shortly see, the present findings sit very well with current SLA theory.

7.2 Interaction between raters' judgements and objective measures

As Wolfe-Quintero *et al.* point out, objective measures used in the evaluation of L2 writing "consist of intuitive rather than theoretical operationalizations of fluency, accuracy and complexity" (1998: 4). In other words, we do not have a valid theory-driven developmental index which would allow us to hypothesize what changes to expect, as writing proficiency

improves, in objective measurements such as sentence length or clause ratios. Although many studies have used objective measures as evaluation tools, findings have not always been consistent, and text genre, purpose and personal writing style, as well as proficiency, may all contribute to observed measurements. Further, increases in unit length and unit ratios may not be linear, and there may be level effects, such that what distinguishes between, for example, intermediate and upper-intermediate learners does not *ipso facto* also distinguish between upper-intermediate and advanced. There may also be ceiling effects. Findings must therefore be considered in context, and in the light of other available information.

In the present study, the triangulation of objective measures and subjective judgements proved very successful, and was particularly helpful in clarifying the issue of raters' lower *grammatical accuracy* judgements for the higher-level experimental class, as compared to the matched control class, where the question of the former group's much longer compositions was a complicating factor. Correlations between raters' *vocabulary range* judgements and numbers of tokens at each level of the Internal Frequency List also demonstrated the validity of this previously untested approach to vocabulary measurement.

The strongest relationship between raters' judgements of performance in any individual construct and an associated objective measure was that between *grammatical accuracy* and number of words contained in error-free T-units, which showed a correlation coefficient of $r = .720$. The use of error-free T-units as an objective measure of accuracy is one of the most consistently reliable and widely used measures, although Wolfe-Quintero *et al.* classify number of words in error-free T-units as a fluency measure rather than an accuracy measure. In the present study, number of words in error-free T-units did not correlate so well with *coherence and flow* judgements as with *accuracy* judgements. The relatively strong relationship between *complexity* judgements and number of subordinate clauses — by comparison to the weaker relationships between *complexity* judgements and mean length of sentence and T-unit — reaffirms the validity and reliability of amount of subordination as a measure of syntactic complexity.

Referring to Table 6.63, the only apparent discrepancy between results obtained from objective measures and those obtained from raters' judgements is between error-free T-unit measures and *accuracy* judgements at the lower level. Raters did not perceive significantly higher levels of accuracy for the experimental compositions at this level, although objective measures suggest that levels of accuracy *were* significantly higher. However, it should be

recalled that raters did perceive a difference in *accuracy* in favour of the experimental compositions which narrowly failed to reach significance ($p = .06$).

Increased complexity, as indicated by both objective measures and raters' perceptions, may have led to errors considered by the raters to be of greater gravity, or to more errors per *with-error* T-unit, and this, combined with increased length, may have masked the increase in accuracy to a certain extent. Moreover, although number of error-free T-units had a stronger overall relationship with *grammatical accuracy* judgements than with *coherence and flow* judgements, when the data set was split into higher and lower levels the correlation between *coherence and flow* judgements and number of error-free T-units was markedly higher for the lower-level group ($r = .684$; $N = 67$) than for the higher-level group ($r = .394$; $N = 83$). This suggests that the presence or absence of error may have much more impact on perceptions of fluency at very low proficiency levels than in general, and that Wolfe-Quintero *et al.*'s contention may be justifiable at these low levels, if less so at higher levels. If this is the case, then the apparent discrepancy is less obvious, since raters may have factored the objectively verifiable higher levels of accuracy in lower-level experimental compositions into their *coherence and flow* judgements, which did show a significant difference in favour of the experimental students, at the same time making *grammatical accuracy* judgements which tended towards a significant difference.

Other objective measures mirrored raters' perceptions. At the higher level in particular the unexpected, though non-significant, lower *accuracy* ratings for the experimental group matched the equally unexpected and non-significant lower numbers of, and numbers of words in, error-free T-units per 100 words produced by this group. The same slightly lower level of relative accuracy is also seen in the percentages of correct regular simple past, past BE and "past" modal verb forms. At both higher and lower levels, the failure of raters to observe significant differences between control and experimental groups for *vocabulary range* was replicated by the failure of VocabProfile to demonstrate any differences. As can be seen from Table 6.56, the percentage of vocabulary drawn from each of the VocabProfile frequency bands was nearly identical for the two experimentally matched classes at each of the upper and lower levels. However, a difference in frequency band use *can* be seen between lower and higher levels — and a difference between levels is also observed by raters in their *vocabulary range* judgements. Indeed, it is only when vocabulary types are pooled for each class that any difference between control and experimental, at the higher level, is objectively observable. Raters only worked with individual compositions. Even

were they aware of it — which they could not be, since they did not know from which class any composition came — they could not factor in group lexical originality to an individual composition score.

For the lower-level pair of classes, raters' perceptions of higher levels of complexity for the experimental group concur well with objective measurements. Of particular note is the significant increase in mean sentence length, despite the problems noted with this measurement (*cf.* Section 6.9). Given that mean sentence length was found by Ortega (2003) to discriminate between previously-recognised proficiency levels in only six out of 14 between-level comparisons, including non-adjacent and even maximally different proficiency levels, this represents a high level of improvement. It might even be taken to show that participation in the reading scheme enabled the lower-level experimental students to produce text of a syntactic complexity equal to that normally produced by students of at least one proficiency level higher. (Proficiency level was established in the studies which were examined by Ortega either by programme level or independent holistic ratings.)

The lower-level experimental students did, in fact, consistently either match or outperform the control students from the higher-proficiency pair of classes in every objective measure of syntactic complexity except mean length of clause, for which, as we have seen, increasing length may not, in any case, be symptomatic of increasing sophistication, at least at this level. They did not, however, outperform the higher-level control class with regard to raters' *grammatical complexity* judgements, which raises some interesting points. As we saw in Sections 6.9 and 6.10, *grammatical complexity* judgements showed significant correlations with objective syntactic complexity measurements. These, however, tended to be relatively low, ranging mainly between r = .235 (for *grammatical complexity* and mean length of T-unit) and r = .297 (for *grammatical complexity* and number of clauses per sentence). Only two objective syntactic complexity measures obtained a correlation coefficient higher than this, the highest being r = .520 (for *grammatical complexity* and total number of subordinate clauses). These figures point to a considerably weaker relationship between *grammatical complexity* judgements and associated objective measurements than between *grammatical accuracy* judgements and error-free T-unit measurements. (Mean length of clause is not considered here to be a valid indicator of syntactic complexity: *cf.* Section 6.9.)

From this existing, but comparatively weak, relationship between *grammatical complexity* judgements and syntactic complexity measurements, we may conclude that, although aware

of syntactic complexity, raters based *grammatical complexity* judgements on more than just this. There are, of course, many other language parameters which may be included in any definition of grammatical complexity, such as amount of modification, sophistication of verb forms, including use or non-use of passives and gerunds and the use, or otherwise, of complex nominal verb phrases, to name but a few. The lower-proficiency experimental students may have matched the higher-proficiency control students in *syntactic* complexity, but their failure to match the higher class in overall, rater-judged *grammatical complexity* shows that the reading scheme did not improve these other aspects of grammatical complexity to the same extent. Here, we have evidence that certain aspects of language development are more susceptible than others to an effect of extensive reading. It will be argued in Section 7.4 that the amenability of any aspect of L2 language development to improvement through the activity of extensive reading may depend largely on its tangibility.

7.3 From reading to writing

To date, very little research has attempted to explore the precise nature of the reading-writing relationship. In L2, only two directly relevant studies have been published in the major applied linguistics journals in the past 15 years. Both were correlational studies which used holistic assessments of L2 reading and/or writing abilities in attempts to quantify the relationship between these, operationalizing each skill very broadly, in terms of a single component. Carson, Carrell, Silberstein, Kroll and Kuehn (1990) found that scores on an L2 English cloze test showed a correlation with holistic scores, ranging from 1 to 6, on an L2 English writing task of $r = .494$ ($N = 48$) for Chinese L1 speakers and $r = .271$ ($N = 57$) for Japanese L1 speakers. Although this points to *some* kind of relationship, particularly in the case of the L1 Chinese speakers, as with most correlational studies of this nature there can be no ascription of cause and effect. Nor is it possible, because of the single-component operationalization of each variable, to locate, or even to hypothesize the nature of, the overlap between the two skills. Hedgecock and Atkinson (1993), in a study involving 115 subjects of mixed L1 background, found no statistically significant relationship between English L2 writing ability and self-reported L1 and L2 reading practice. This second study, however, was seriously methodologically flawed, in the first instance by overloading the regression procedure with no fewer than 24 variables, and in the second instance, by using a number of theoretically invalid and statistically redundant variables. (For example, L2 students enrolled in a language course at an American university were asked how much reading of English newspapers they did whilst in elementary school. The "mean" response on

a 4-point Likert scale where the score of 1 represented "none at all" was 1.2, probably because most of the students went to elementary school in their home countries. It is not surprising that an activity which never took place did not have a statistically significant relationship with skills in a subsequently-learned L2.)

In L1, the reading-writing relationship has been researched in more detail by Shanahan (1984) and Shanahan and Lomax (1986). The second of these studies investigated *directional* relationships, using path analysis, between a variety of reading and writing subskills at the levels of second- and fifth-grade. Of particular interest to the present study is the finding that reading comprehension contributed more to writing syntax at the lower level. This coincides with the present finding that reading-scheme students exhibited significantly higher levels of syntactic complexity than control students at the lower level, but at the upper level there was little difference between reading-scheme and non reading-scheme students. Shanahan and Lomax also found a greater level of mutual interaction between reading and writing skills at the lower level, whilst at the higher level the relationship was more unidirectional, with reading competence impacting more on writing competence.

Shanahan (1984), investigating the reading-writing relationship for beginning readers, middle-level readers and proficient readers, found that at the highest level vocabulary diversity was the greatest contributor to the reading-writing relationship. Vocabulary diversity was also more important at the high middle level than at the beginning level, suggesting that this factor increases in importance as the reader becomes more and more proficient. Again, this coincides with the finding that, although the lower-level reading-scheme students were better overall writers than their control peers (as shown by *overall quality* ratings), better vocabulary knowledge did not seem to be a part of this, since neither holistic *vocabulary range* judgements nor objective vocabulary measurements showed the former group to have performed better. At the higher level, however, the reading-scheme class displayed a higher degree of *lexical originality* than the non reading-scheme class. At the lowest reading level in Shanahan's study, spelling was found to contribute more to the reading-writing relationship than other variables.

There may, however, be limited validity to comparisons drawn between L1 reading-writing relationships and L2 reading-writing relationships. In the particular instance of the 1986 directional study, when examining the *interactive* relationship between the two skills, Shanahan and Lomax found that (reading specific) vocabulary knowledge exerted a strong

influence on reading comprehension at the level of the fifth-grade but did not explore the reverse possibility — sometimes hypothesized in L2 — that skill in reading comprehension may have an incremental effect on reading vocabulary knowledge or on written vocabulary production. (In path analysis, causal relationships are extracted from multiple regression models; however, the independent variables used in the first instance are selected by the researcher, and are those believed by that researcher to influence the particular dependent variable.) Further, being a poor reader in an L1 may not involve the same weaknesses as being a poor reader in an L2, where an almost unavoidable lack of at least some necessary grammar and vocabulary knowledge (if the text is not simplified) is far more likely to cause disruption to the comprehending process, and the lack of *verbal efficiency* (Perfetti, 1985; 1988), which is to say automatization of reading subcomponents, including instant, effortless, *automatic* meaning-identification of words once they have been decoded, will in many cases slow down the reading process. The L2 writer, similarly, may face a somewhat different task from the L1 writer, and may even deploy L1-based strategies, including translation, in the production of L2 text, resulting in quite different composition and language retrieval processes.

Whilst the writing skills of students in the present data set may be divided into component linguistic constituents such as, for example, syntactic complexity, coherence and vocabulary use, the reading of the ERS students can only be considered in terms of *reading practice.* It is a particular strength of the research design that the extra reading undertaken by the experimental students was not accompanied by additional writing practice. If anything, the experimental students may have undertaken *less* writing practice than their control peers, since the latter group had two extra classes of integrated-skills coursebook-based instruction. Thus, observed effects may be attributed to increased time spent on the *activity* of reading, and to exposure to written *input* (*cf.* Krashen, 1982, 1985), quite independently of the deliberate and sustained focus on grammatical, lexical and rhetorical form necessitated by producing written *output*, which many believe to be just as influential, or more so, as regards the development of L2 proficiency (*cf.* Swain, 1985).

It is quite clear from the data that reading practice did affect writing skills, although the reading practice cannot be considered as *uniquely* causative of observed effects, since concurrent writing and grammar instruction continued to take place during other English classes, and different effects might have been produced by the reading practice had this not been the case. What we are comparing is the writing of two treatment groups, one of which

received eight to ten 40-minute periods of standard classroom L2 instruction weekly over a period of three years while the other received two periods less per week of standard classroom instruction but with at least one and a half hours of additional silent reading practice. Whilst the reading scheme afforded students the opportunity of reading English outside the classroom, not all of them, in fact, took advantage of this. Interviews with students and teachers revealed that, although many students did read outside school, perhaps finishing a graded reader between one ERS class and the next, for others their only sustained reading was done during the two reading-scheme classes, under the watchful eye of the teacher. For the reading-scheme students, levels of grammar and vocabulary knowledge were maintained to the same standards as those of the non reading-scheme students, despite less time spent on formal instruction — either because there was no qualitative difference between, for example, six classes and eight classes of instruction per week, or, more probably, because the two reading-scheme classes, acting in tandem with concurrent formal classroom instruction, were just as effective in this as two more classes of instruction. Moreover, levels of coherence, complexity and fluency were considerably enhanced by the reading treatment.

Reading and writing processes are invisible and as yet little understood. The models which we have are theoretical (*e.g.* Goodman, 1976; Stanovich, 1980; Hayes, 1996; Hayes and Flower, 1980), consisting of hypothetical mechanisms such as mental lexicons, filters, monitors and retrieval systems, which, whilst providing helpful analogies, may, in fact, have no neurological correlates (Jacobs, 2004). It has not been possible to establish in what way, if any, the two processes utilize common, or linked, cerebral pathways. This may be one reason why the research bias in L2 has not favoured investigation into the more precise reading-writing relationship, but centred, instead, on the larger, visible, paradigms of input and output, which subsume, to some extent, those of reading and writing, but do not clarify in any fine detail.

There can be no question that the two processes are served by the same underlying linguistic knowledge, and in this must lie part of any observed relationship. Not even staunch Universal Grammar theorists have suggested that output can produce itself all on its own without corresponding previous input. Writing must draw on the written forms of words acquired from reading, starting at the outset with written input which might consist of no more than a single isolated word, which the novice writer attempts to copy. It is true that, once the system of sound-to-letter matching and the skill of graphic reproduction are both

mastered, a writer may produce a word he has never before seen written — although this may be easier in some languages than others. This, too, however, cannot come from nowhere, representing as it does an informed best-guess, computed from various stored sources of information. In this case, the aural/oral form of the word is retrieved from memory and converted to graphic form, possibly referring to other words deemed similar for which a known written form is already available. Here, indeed, we have one of the major differences between L1 and L2 writing: in normal L2 instruction, the written form is generally presented more or less concurrently with the aural form, and, in fact, the written form serves as a very important pedagogical tool in the L2 teaching and learning process, which it does not in the L1 learning process. *Input* in general, then, and — more particularly in instructed L2 — *written* input, must exert considerable influence on writing, and may reappear in *output*. However, accepting that L2 reading and writing share, in large part, the common resource of knowledge of written forms, and also knowledge of the underlying language system in general, in what way, if any, can we say that the *process* of reading affects the *process* of writing?

Reading practice is commonly supposed to enhance, within that same modality, speed and levels of *automaticity* of lexical access (Segalowitz, Poulsen and Komoda, 1991; Paran, 1996; Walczyk, 2000). (It should be noted that speed and automaticity are not necessarily the same thing.) The more often we are exposed to a word, the more quickly we recognise it on future occasions. This may also extend to *chunks* of language, which might be formulaic expressions such as "Pleased to meet you", common collocations, or even grammatical structures such as "is going to see" or "has been found". So reading practice can not only speed up vocabulary recognition, but apprehension of grammar, resulting in enhanced reading *fluency*. DeKeyser (1997) has suggested that practice is skill-specific, and does not transfer from one modality to another. If such were the case, reading practice could have no beneficial effects on writing fluency. Our present data, however, contradicts this hypothesis. Since the reading-scheme classes did not benefit from additional writing practice, and, hence, could not simply have been faster and more skilled in the activity of graphic reproduction, their much greater output within a restricted time must be interpreted as indicative of increased ease, hence speed, of language retrieval. So speed of access caused by reading practice translated into speed of retrieval for writing. In this way, the *process* of reading can be seen to have impacted on the *process* of writing.

One reason for the lack of congruence between the present findings and those of DeKeyser may be the respective research designs — in particular, the duration of the treatments. Whilst DeKeyser's students had only15 practice sessions, over five or six weeks, the reading-scheme students had approximately 150-300 hours of reading practice. Practice in skills may take a long time to transfer. In addition to actual amount of time spent on-task, there may be maturational requirements. DeKeyser's students had only just learnt the grammatical rules in question, as part of the experiment, so, although the practice may have speeded up recognition and apprehension, the recency of learning may have affected the students' ability to automatize the rules. Automatization is not a gradual process, but involves a "switch" to an entirely different memory system. In fact, the consistency of the regression lines shows that automatization probably did not take place in DeKeyser's experiment. Even apart from any maturational constraints, the very fact of learning the rules specifically for the experiment, then practicing *only* these rules, in controlled, so-called *laboratory* conditions, could have meant that the students retained at least some conscious awareness of the rules throughout. Conscious awareness is an inhibiting, possibly *pro*hibitive, condition for automatized memory function. If the speeding up of recognition and apprehension, or *access*, caused by comprehension practice in DeKeyser's study did *not* come from automatization of language, and had no impact on subsequent production fluency, then here we have evidence that the increased speed of language production shown by the reading-scheme students *could* have been the result of automatization, since we may conclude from the first study that the mere speeding up of language access would not have produced this result.

It is also likely that enhanced speed of retrieval, whether or not this comes from automatization, is conducive to the production of longer syntactic units and more syntactically complex text. In reading, speed of access to lexis and grammar becomes more and more important in the comprehending process as sentences, or other meaning-units, get longer and longer, since all the necessary components must be apprehended and integrated before the first part of the sentence is forgotten. So, too, in writing, speed of production enables the skilled writer to hold in working memory longer strings of text and relationships between sub-units such as clauses. Effortful retrieval may result in short sentences which, additionally, may have a low level of inter-sentential cohesive fluidity, since each sentence will be produced as an independent task, and its structure — although not its meaning — may be forgotten before the next sentence is well under way. The enhanced levels of syntactic complexity and of coherence observed in the writing of the lower-level reading-scheme students may be, not so much the effect of additional exposure to input exhibiting

such text features, as the result of faster and less effortful language retrieval. Again, it is the activity or *process* of reading, and — as we will later see, extensive reading in particular — which may have benefited the writing process.

Although the most immediately apparent benefits of the reading-scheme were syntactic (as opposed to formal) complexity, coherence and fluency, all of which, as we have just seen, may have been the result of increased speed of access to, hence retrieval of, the L2, there are nonetheless also a number of exposure-to-input effects. Despite the fact that the raters did not discern a significant difference in *grammatical accuracy* between the control and experimental students at the lower level (although this may, to some extent, have been included in their *coherence and flow* judgements), error-free T-unit measures show clearly that the reading-scheme group made significant progress in grammatical accuracy. This was probably the result of repeated exposure to the correct forms. Moreover, this lower-level group may even have acquired the past morphology of regular verbs simply from exposure. (The reading-scheme class produced 82.2% correct forms of regular declarative simple past compared to 56.8% for the control class.) Again, the data contradicts the findings of DeKeyser (1997), who found that not only increased fluency, but also accuracy, resulting from repeated instances of comprehending particular structures via reading (using picture and text matching techniques) did not have an effect on the fluency and accuracy of subsequent written production of those same structures. In our case, grammar information deriving from reading did successfully cross over to the modality of writing. This may be evidence that some knowledge of grammar structures was also successfully automatized.

At the higher level, the reading-scheme class demonstrated considerably greater lexical originality than any of the other three classes, showing that this group used vocabulary which none of the other students did. The scale of this originality (*cf.* Tables 6.57 and 6.60) strongly suggests that this was not a mere co-product of the longer compositions which the students from this group produced, but may have been the result of their reading, which either gave them access to vocabulary outside that of their coursebook and other class materials, or gave them the assurance to use words which the other students may have known but were less confident in using. As we shall see in the next section, other possible exposure-to-input effects for this level may have been severely disrupted by extensive reading practice effects.

7.4 Implicit and explicit learning

Recent advances in neurobiology, using noninvasive imaging techniques such as fMRI (functional magnetic resonance imaging) and PET (positron emission tomography) to study the brain activity patterns of normal individuals, have established that the human brain has two distinct long-term memory systems, subserved by different neural mechanisms and located in quite separate areas (Aglioti, 1999; Milner, 1999). *Declarative* memory, also called *explicit* memory, or memory with knowledge, is located in the hippocampal region and the neocortex (the most recently evolved part of the human brain, which other species do not have). *Nondeclarative* memory, also known as *implicit* memory, or memory without knowledge, is largely supported by the basal ganglia, at the base, or more primitive part, of the brain.

These two types of memory differ in function, in type of information collected and stored, and in stability. In broad terms, "declarative memories are memories for facts and events, and nondeclarative memories are memories for habits, motor and perceptual skills, and emotional learning" (Schumann, 2004: 5). When we consciously think about something, we are using declarative memory, the contents of which can be "consciously recalled, represented, or verbalized" (*ibid*). When we know something but have no conscious awareness of that knowledge, we are using nondeclarative memory, the contents of which "cannot be accessed through conscious effort" (*ibid*). Declarative memory collects its information explicitly and consciously. Nondeclarative memory collects information implicitly, by way of "a nonconscious and automatic abstraction of the structural nature of the material arrived at from experience of instances" (N.C. Ellis, 1994: 1). In terms of both evolutionary and individual anatomical development, nondeclarative memory precedes declarative memory (Paradis, 1994; Aglioti, 1999), such that infants initially use the former during the earliest stages. Finally, nondeclarative memory, although it may take longer to acquire, is considerably more robust than declarative memory, does not deteriorate over time in the same way and is more resistant to change.

These memory systems are not domain-specific, which is to say that they subserve every aspect of learning, including language acquisition. In fact, in the light of what is currently known about the brain, it now seems probable that there is no such thing as a domain-specific language-acquisition device, but that language is learned using the same mechanisms as any other skill, and may be learned implicitly and nondeclaratively, as in the

early, formative stages of L1 (when, developmentally, there is no other option), explicitly and declaratively, as in a later-stage, instructed L2, or in some combination of the two methods. Language competence which is stored in nondeclarative memory is durable and executes automatically. Language competence stored in declarative memory requires conscious execution and needs some level of maintenance in order to persist. It is also, however, relatively much easier to modify. (The phenomenon of L2 fossilization is likely to be the result of partially mastered language structures being prematurely taken into nondeclarative memory, either as a result of persistent, unconscious use of an incorrect form or some maturational effect — hence their great resistance to modification. Pronunciation is proceduralized — procedural memory being a specific form of nondeclarative memory — very early on in the L2 acquisition process, and is thus particularly susceptible to fossilization. This may be because, after the initial stages, learners do not, on the whole, consciously think about their pronunciation.)

In SLA, many of the apparent inconsistencies and ongoing disagreements about optimal teaching methods might be elucidated somewhat if viewed within the framework of these recent findings in neurobiology. Different, apparently contradictory, teaching methods may serve one, but not the other, of the memory systems. Focus on form, for example, must serve declarative memory, since it *does* focus on form. Focus on meaning, without conscious attention to form, may result in input to the *non*declarative memory system. Both, however, contribute to the language system, since it uses both systems of memory. The second type of learning may take longer to attain, but will also be more durable. Krashen's *comprehensible input* hypothesis can be vindicated, as serving long-term, implicit language acquisition, stored in nondeclarative memory, whilst specific *output hypotheses*, such as *output plus correction* and *comprehensible output* (*cf.* N.C. Ellis and Laporte, 1997), which may be seen to require explicit attention to language, will favour the kind of language development which takes place in declarative memory. Extensive reading, as we shall shortly see, may provide a very valuable source of implicitly acquired input.

It is also the case that language information (or any other type of information) which is held in nondeclarative memory, and which, consequently, executes *automatically* when required, is impervious to short-term outside influence, such that top-down reading strategies will not be useful in the word identification process. Whilst top-down reading comprehension strategies may be seen to help declarative memory to make word-meaning connections, it is bottom-up reading skills, or extensive practice of same-word recognition, irrespective of

context, which will aid proceduralization and entry into the nondeclarative memory system. Thus the apparently contradictory positions held by the respective schools of thought may be theoretically reconciled. Both are correct.

Although the separateness of the two memory systems is as yet little recognised in SLA research, there has nonetheless been much recent debate concerning implicit and explicit rule learning, two main issues being, firstly, whether language rules *can* be learned implicitly, and, secondly, whether explicit knowledge may ultimately "convert" to implicit knowledge. This is the same, theoretically, as asking whether language rules can be learned by nondeclarative memory and whether the content of declarative memory can be subsumed, or absorbed, through practice, into nondeclarative memory.

As regards the second question, there are three main positions; the strong interface, the weak interface and no interface (Segalowitz and Hulstijn, 2005). Proponents of the strong interface position claim that explicit, declarative knowledge can "convert" to implicit knowledge as a result of practice (*cf.* Anderson, 1993; Anderson and Lebiere, 1998). The weak interface position proposes that explicit, declarative knowledge somehow facilitates the acquisition of implicit knowledge, although exactly *how* this comes about is unclear (*cf.* Paradis, 1994), and the no interface position, of which Krashen might be regarded as the best-known proponent, that the two do not impact on each other. It is probably the second of these positions which best corresponds to that most recently published by the Neurobiology of Language Research Group of the University of California at Los Angeles, which is that:

> knowledge that is stored declaratively is not *converted* into nondeclarative knowledge. Instead, learners acquire and store information in both declarative (hippocampus/cortex) loops and nondeclarative (basal ganglia/cortex) loops. Thus, what would appear on the behavioural level to be a "conversion" is, in actuality, probably a strengthening of connections in the nondeclarative loop that is sometimes accompanied by a weakening of connections in the declarative loop (Crowell, 2004: 101; emphasis in original).

In other words, the two memory systems acquire their knowledge separately. However, as Paradis points out, "this does not mean that metalinguistic knowledge cannot be useful in the process of learning another language .... But it is the practice, not the metalinguistic knowledge, which improves automatic performance (and by implication, linguistic competence)" (1994: 405). So explicit instruction can involve practice, which in turn may provide input material for the nondeclarative memory system. Moreover, as Crowell notes, acquisition of the same knowledge by the two systems may result in the situation whereby,

as nondeclarative memory acquires and consolidates that knowledge, declarative memory begins to forget it, since declarative memory need no longer be used for that particular knowledge — and it may be this that leads to the false hypothesis that the knowledge is transferring from one memory system to the other.

The first question — whether language rules can be learned implicitly — has been the subject of much debate in Applied Linguistics over the past 15 years, and is a central theme of the *focus on form* and *focus on forms* (*e.g.* Long, 1983, 2000; Doughty and Williams, 1998) and the *noticing* (*e.g.* Schmidt, 1990, 2001) research strands. In a research synthesis article, Norris and Ortega gathered together the findings from 49 relevant L2 research studies published between 1980 and 1998 and concluded from these that "explicit types of instruction are more effective than implicit types, and that Focus on Form and Focus on Forms interventions result in equivalent and large effects" (2000: 417). There are many shortcomings, however, to Norris and Ortega's critical examination of these studies and their reported conclusion must be viewed with some reservation.

Firstly, the research synthesis included only studies which "experimentally" investigated the effectiveness of L2 instructional treatments — thereby building in a research bias from the beginning, since implicit learning does not lend itself to controlled, experimental research designs. For one thing, such learning takes longer to become apparent, and for another, if what is learned is to be *implicitly* learned, it must not be the subject of explicit focus, hence cannot be a point of instruction. If it is not a point of focus or instruction, then a researcher will not know *what* to test, since he cannot know what has, or has not been, accessed by the learner. In short, implicit learning does not lend itself either to short-term instruction or to discrete-point testing.

Another, very serious, weakness of the focus on form and focus on forms studies is that evaluation of the learning outcome is nearly always undertaken just after the instruction sessions, with little investigation of long-term, durational effects. Rather tellingly, of the 49 studies which were looked at by Norris and Ortega, fewer than half provided data from a single delayed post-test and only six from two delayed post-tests. Norris and Ortega conclude that, although effects can be seen to diminish over time, they are nonetheless "relatively durable". However, they either rather carefully or rather negligently (and extremely frustratingly) fail to quantify either the "delay" or the "durability", and it is impossible from their report to identify and follow up the studies in question. In practice,

such "delayed" post-tests rarely take place after more than a few weeks, and frequently after only a few days. Moreover, Norris and Ortega refer to "maturation effects" only incidentally and almost as if these might be extraneous variables which obscure the true picture. More generally, as Doughty puts it, "the case for explicit instruction has been overstated" (2003: 274) and the apparent advantage of explicit instruction should be "more properly interpreted .... as an artifact of cumulative bias" (*ibid*). What the focus on form/forms studies *have* been able to show is that instruction *is* attended to by the learner, which is itself a valuable finding, and does result in intake, but, in terms of the larger picture of real-life language learning, there are nonetheless a range of attendant issues such as maintenance, integration, maturation and forgetting, which cannot simply be ignored.

Theoretically, explicit instruction may better suit some aspects of L2 learning than others. MacWhinney, for example, argues that "explicit instruction works best for clear, simple structures" (1997: 278), and it has been the case that focus on form/forms experiments have focused almost exclusively on well-defined grammatical structures. A rare exception is DeKeyser 's (1995) investigation of "fuzzy" rules (prototypicality patterns) which he, consistently with MacWhinney's claim, found to be better learned via implicit teaching methods, although this was not confirmed statistically. Not all language is rule-based, however (*cf.* Skehan, 1998), nor do grammatical structures represent a whole language system. Other components, such as complexity, coherence, fluency and various aspects of vocabulary development are just as important a part of language competence. Vocabulary has received a similar kind of attention as that given to grammar, within the contexts of the *incidental learning* and *depth-of-processing* research agendas (*e.g.* Laufer and Hulstijn, 2001), but syntactic complexity, coherence and fluency have not been researched within the *focusing* and *noticing* frameworks.

In our present data, it is precisely these three constructs which can be seen to have most clearly benefited from extensive reading practice. For the lower-level pair of classes, whilst no difference was observed by the raters between control and experimental groups for *vocabulary range* and *grammatical accuracy*, *grammatical complexity* and *coherence and flow* judgements were significantly higher for the reading-scheme students than for the non reading-scheme students. Objective syntactic complexity measures confirmed the raters' judgements. For both lower- and higher-ability groups there was a significant increase in fluency for the reading-scheme students, particularly at the higher level, where reading-scheme students produced a quarter as much text again as the non reading-scheme students.

These differences in output are by no means trivial. Wolfe-Quintero *et al.* (1998), investigating the use of number of words as an objective fluency measure in timed compositions, found that of ten studies where a significant difference was found between groups, nine were between previously recognised different proficiency levels. One other between-proficiency study showed a trend. Of seven studies where no significant differences were found, all had investigated differences between learners at the same proficiency level (although subjected to different treatments). These differences in output may thus be seen as characteristic of differences between proficiency levels, and output levels may be seen also as relatively hard to improve upon. That reading practice effected such an improvement is worthy of note, and, in fact, this is an effect which has been consistently observed in investigations into the impact of extensive reading programmes (*e.g.* Hafiz and Tudor, 1990; Lai, 1993).

That the most clearly observable benefits of the reading scheme were in complexity, coherence and fluency may be partly because, unlike grammar structures and vocabulary, these constructs do not, in any case, lend themselves very readily to explicit classroom instruction. Although specific grammatical structures may be taught, techniques for enriching the overall complexity of a text, for example, by the use of increased levels of coordination and subordination, or by deliberate selection of more complex structures, generally are not. Instead, increasing complexity may develop naturally, as a result of increased experience and confidence, and a widening range of language competence on which to draw. As regards sentence length in particular, learners do not on the whole deliberately set out to produce longer sentences, and are not categorically taught to do so. In L2, as in L1, we produce longer and longer sentences simply because we *can*. Coherence is all but impossible to teach using any set of generalizable rules, and is usually allowed to develop on its own, as a result of increasing experience and confidence. Unless poor coherence actually violates rules of grammar or usage, it is little attended to by teachers generally more concerned with the more concrete building blocks of vocabulary and grammar. Fluency, too, will develop from practice, not from explicit teaching. So the practice of extensive reading may have benefited these constructs in a way that coursebook-based instruction did not partly because these do not form part of any ordinary EFL syllabus.

There is also, however, the fact that these constructs are in any case more amenable to the kind of *implicit* learning that extensive reading may engender (since it is focused on meaning, and not on form). Complexity and coherence do not correspond to a finite set of

examples; it is difficult to extract from them any workable *forms* on which to focus; they are not governed by well-defined — hence noticeable — rules. In short, what makes these so hard to teach also makes them good candidates for acquisition by the nondeclarative memory system, which is associationist, inductive and relies on instances rather than explicit rules (N.C. Ellis, 1994). It has also been argued by VanPatten (1996) that, when processing input, we focus on lexical items. If such is the case, the diminished, more fleeting, attention paid to what is *in between* the lexical items — the "glue" which holds these together, including patterns of sentence structure and coherence — is again more likely to be routed to nondeclarative memory, given that deliberate focus will claim the input for declarative memory processes. This focus on lexical items may be, as VanPatten argues, the result of a natural preference for content rather than form. It may also, however, be partly the result of the respective amenabilities to each memory system of well-defined and indefinite language items. In other words, we may focus more on lexical items partly because these are more amenable to focus, and, if we need to extract *meaning*, we must focus on what we can.

We have seen that the students in our study who engaged in the practice of extensive reading attained levels of syntactic complexity (for the lower class) and fluency consistent with a higher proficiency level. The lower-level reading class also attained a significantly higher level of rater-judged coherence than the matched control class. It has been argued, in Section 7.3, that these effects are the result of increased levels of automaticity, which is to say an increased amount of the L2 system being held in the nondeclarative memory. Since these effects pertain to students who differed from their peers only in that they took part in an extensive reading scheme, we must conclude that reading practice, and in particular *extensive* reading practice, has facilitated entry of L2 language into the more robust and durable nondeclarative memory system. In Krashen's terminology, this language has been *acquired*, as compared to simply *learned*.

We know that the implicitly processed contents of nondeclarative memory take longer to acquire than the explicitly processed contents of declarative memory. The simplicity of extensive reading materials allows for faster reading, hence more opportunities to encounter language structures and patterns, and this wealth of exposure to instances is likely to facilitate such implicit learning. There is, however, another way in which the simplicity of extensive reading materials might encourage this kind of language acquisition. Ease of comprehension allows focus on meaning without attending to language form, and this, too, may engender implicit learning. Text which requires deliberate attention to form will not

only reduce the amount of accessed input, by slowing reading speed quite considerably (if this may still be called reading) but may also result in the input being stored in declarative memory, as it seems that input acquired *consciously* is stored in *conscious* memory. Thus it is the *lack* of focus on form which may lead to storage in nondeclarative memory.

As regards vocabulary, it has been postulated for some time that, in L1, syntax and vocabulary are not stored in the same part of the brain (Ullman, Corkin, Coppola, Hickok, Growdon, Koroshetz and Pinker, 1997). This theory originally derived from the psychological study of individuals suffering from brain-damage or degenerative brain disease who had lost the use of, or suffered very serious disruption to, either syntax or vocabulary, whilst retaining full competence in the other. Conclusions from such studies have since been corroborated by neuroimaging studies (Ullman, 2001b). There have also been cases where all language was lost following insult to the brain, such as a stroke, and where either syntax or vocabulary, but not the other, was recovered. (Interestingly, there have also been cases where a speaker of two languages lost all language use, but subsequently recovered the lesser established language without recovering the L1.) It now seems very likely that in L1 syntax is stored in nondeclarative memory and the kind of vocabulary which may be more properly considered as *lexical* or *semantic* items, as compared to functional vocabulary, tends to be stored in declarative memory. Hence we are more susceptible to vocabulary loss as a result of normal ageing than to loss of syntax; declarative memory is far less stable than nondeclarative memory, and degenerates more quickly.

As regards L2, fMRI scans have shown that, in the first stages of learning, brain activity when using the L2 is very different to that when using the L1 (Abutalebi, Cappa and Perani, 2005), and is spread throughout the area associated with declarative memory. As the learner becomes more and more proficient, brain activity resembles more and more that of L1 brain activity, and in early bilinguals or near bilinguals there is very little difference in brain activity when speaking the L1 or the L2. It is likely that, with increasing L2 proficiency, the same components of language will be proceduralized as in an L1. If the L2, however, remains effortful — hence consciously accessed — for the learner, there will be more activity in declarative memory. (In the case of bilinguals recovering their less proficient language, but not their L1, after insult to the brain, it is supposed that this is because there have been more traces of the L2 left in declarative memory, whereas the L1 has been more

comprehensively proceduralized and could not be recovered when procedural memory was lost.)

If even in L1 lexical items are more resistant to proceduralization than syntax, this is likely to also be the case in L2, and this may be why there was no difference in vocabulary use between reading-scheme and non reading-scheme students at the lower level, and why, in general, it has been hard to find evidence of vocabulary benefits brought about by extensive reading (*e.g.* Tsang, 1996; Yu, 2000). Both well-defined grammar structures and lexis may benefit as much, or more, from the type of explicit, form-focused instruction which goes on in the normal L2 classroom. (It should be recalled that the reading-scheme students in the present study benefited also from normal classroom instruction.) Whilst morphology may, under optimal conditions, eventually proceduralize, this is less likely to happen with lexis, particularly lexis which is encountered erratically and in specific contexts. The much improved performance of the lower-level reading-scheme class (as compared to the non reading-scheme class) in production of regular past *-ed*, may be a good illustration of morphology *acquired* from frequency of exposure. As N.C. Ellis and Schmidt have pointed out, regularity in morphology "is frequency by another name" (1998: 307). It is likely, too, in this particular case, that the past *-ed* did not elicit much attention, far less deliberate focus, from the students, since the clear, *past*, time frame of a past narrative (by far the dominant genre in the graded readers of the Hong Kong ERS) makes the past marker all but redundant. Students did not need to attend to *-ed* in order to understand the story they were reading.

Some vocabulary, however, may proceduralize. The same students who outperformed their control peers in correct production of *-ed* forms also did considerably better in production of irregular past forms. It is often noted that the most frequently occurring and communicatively useful verbs are also irregular (*e.g.* Pinker, 1991; Altenberg and Granger, 2001). These irregular past tenses may have proceduralized because of their frequency, but as an item of vocabulary rather than as a frequently occurring morphological procedure such as the addition of *-ed* for past tense. Other vocabulary which may proceduralize relatively easily is the kind of very frequent, non context-specific words which might almost be called "auxiliary" vocabulary, including functors such as *this*, *on*, and *my*, as well as very common and useful lexis such as *today* and *want*. These words are more likely to enter implicit memory than more specific lexis, firstly, because they are very much more frequent and, secondly, because they are less likely to be consciously attended to than specific lexis. If frequently encountered function words did proceduralize as a result of reading practice, this

may be one explanation why the lower-level reading-scheme class performed better than the matched control class not only in coherence and complexity, but in numbers of error-free T-units produced. These functors are, of course, heavily implicated in basic syntax.

The kinds of language development apparent in the writing of the lower-level reading-scheme students, as we have seen, fit very well into the declarative/nondeclarative model. We might also describe this development in terms of explicit and implicit, or conscious and unconscious learning, leading to language use which is either attentional or automatic. The higher-level reading-scheme students, however, performed rather differently from the lower level, and we need to ask ourselves why this was so. The most striking effect of reading practice at this level was a very considerable gain in fluency — more so than at the lower level. Although syntactic complexity as measured by length of sentence and T-unit did not improve significantly, there was nevertheless a clear trend in this direction (*cf.* Tables 6.22, 6.23 and 6.24). However, raters, whilst perceiving a small, non-significant improvement in *complexity*, perceived almost no improvement in *coherence* and a slight slackening off of *accuracy* (*cf.* Table 6.12).

There are three possible explanations for the different progress patterns of the two proficiency levels. Firstly, as has been suggested in Section 6.11.1, *accuracy* and *fluency* may in any case benefit differently from input and practice at lower and higher levels. It is easier to show relative progress in accuracy at very low levels where even small gains may appear significant by comparison to the original low level of proficiency. In contrast, fluency may be harder to develop at low levels, since the fledgling language system may not yet consist of either enough language or the right *kinds* of language items (*i.e.* the language needed to formulate relationships between lexical items, such as prepositions, demonstratives, relatives, subordinators *etc.*) necessary for any sort of fluency practice.

A second explanation may be that the input, in the form of graded readers, suited the two proficiency levels differently. Although the reading scheme allowed for students to read at different levels, and made provision for students to move up a level as they improved, providing materials at all levels, up to unsimplified (though relatively simple) text, one, or both, of the groups may have found themselves with a selection of reading materials at not quite their optimal level. If the higher-level students were reading at a level which was relatively easy for them in terms of grammar, there may not have been much for them to notice, or ultimately to acquire (*i.e.* from the reading materials), in that area. Consequently,

the benefits of reading may have been limited to increased speed of access to the L2, or reading *fluency*, which, in turn, had a beneficial impact on speed, or fluency, of written output. Conversely, if the lower-level students were reading at a level which was too difficult for them to permit fluent reading for meaning, they may have had to engage in a different style of reading, closer to *intensive* reading, which caused more focus on grammar and text structure. This focus on grammar and text structure may have led to gains in grammatical complexity and accuracy which the higher-level group, focusing on meaning, did not achieve.

So the differing effects of extensive reading at lower and higher levels may have been a direct result of the input, or of normal language development curves. A third explanation proposes a possible tension between fluency and accuracy as part of a developmental language consolidation process. Skehan and Foster (1999, 2001) have proposed a limited attentional capacity model, which holds that complexity, fluency and accuracy compete with each other for limited cognitive resources, and that priority given to any one of these may detract from at least one of the others. On first consideration, this provides a simple answer. The higher-level reading class may have given the main priority to fluency, some priority to complexity, and this to the disadvantage of accuracy. However, on further consideration, this explanation is *too* simple, and leaves several questions unanswered. Particularly, *why* would one group of students prioritize fluency any more than a peer group (*i.e.* the control group) performing exactly the same language task? (Skehan and Foster suggest that the task itself may influence the learner's priorities.) Secondly, fluency and complexity did *not* compete with each other, but, rather, increased fluency seems to have supported complexity (*cf.* Tables 6.21 and 6.26).

Automatic processes do not compete with attentional processes (Hulstijn, 2001). Instead, automatic processes "run on their own resources, *i.e.* they do not share processing capacity with other processes" (Levelt, 1989: 20-21; cited in Hulstijn, 2001: 264). If increased fluency is a result of automatization, this should not compete with accuracy for attentional resources, but simply happen, quite independently. We do not make a conscious effort to execute a process or activity more quickly than we have the ability to do. If we *do* try to execute an activity at a speed beyond our ability, then the quality of our performance probably *will* deteriorate, but this is a conscious effort, and is not a normal situation. When we learn *any* skill, increasing (non-conscious) speed of execution, as a result of proceduralization, does not normally impact on accuracy of execution. The higher-level reading-scheme students

produced text much more quickly quite simply because they *could* do so, not because they were *trying* to do so. This increased speed of access, as we have seen (*cf.* Sections 6.8 and 6.9), appears to have facilitated increased syntactic complexity, probably because it allowed more language to be held in working memory at a time. It should not, however, have entered into competition with accuracy for attentional resources, since it was not a conscious effort.

Skehan and Foster have noted that "complexity is seen to operate in tandem with the attendant language learning process of restructuring" (2001: 191). Doughty has referred to "the mysterious process of restructuring" (2001: 221). Restructuring, although widely accepted as part of the L2 acquisition process, and often put forward as an explanation for variability in language performance, is very little understood. The new insights into language learning afforded by neurolinguistics may, in the course of the next ten or twenty years, finally shed some light on this process. Skehan and Foster further characterize learners who are restructuring as being "in the process of realising that their IL systems are limited and require modification" (2001: 191). There can surely be no greater "modification" than a switch from one memory system (declarative) to another (nondeclarative) for the retrieval of a language item — although it is very doubtful whether the learner "realises" anything at all. In practice, it seems likely that this is just what the higher-level reading-scheme students were doing — drawing more on the nondeclarative memory system. (We may deduce this from their very considerable improvement in production fluency.) In this case, any increase in syntactic complexity — occasioned by the increase in fluency — may be a naturally occurring correlate of this suggested restructuring process. In other words, increasing complexity may be a *consequence* of restructuring, and not, as Skehan and Foster suggest, learners engaging in an "attempt to pressure their own language systems" (*ibid*) by using more complex language, which in turn will *lead to* restructuring. So the slight fluctuation in relative — but not in absolute — accuracy observed in the performance of the higher-level reading-scheme students may have been, in some way we cannot yet explain, connected to what may be a major form of restructuring.

Of the six constructs evaluated by raters, one stands out as having been completely unaffected by participation or non-participation in the extensive reading scheme. This is *punctuation and paragraphing*, which has not been considered in the discussion until now, partly because it was the least reliable construct, and partly because it is, in any case, not part of the language system *per se*, but is a set of superficial conventions of presentation, which is unlikely to interact developmentally with constructs representing language competence.

Raters perceived no differences between reading-scheme and non reading-scheme students in *punctuation and paragraphing* at either lower or upper levels. This is slightly surprising if we consider that punctuation might also be subject to a frequency effect. *Punctuation and paragraphing* was also the only rater-judged construct which did not produce lower inter-rater correlations for the experimental than the control data (*cf.* Tables 5.12 and 6.14). If we see this turbulence in rater reliability as evidence that there may be something in the experimental compositions which raters did not respond to consistently, and which must be connected to participation in the reading scheme, even if we cannot explain it (*cf.* Section 5.3), then the absence of this effect for *punctuation and paragraphing* further confirms the reading scheme's lack of impact on this construct.

A possible reason for this is the very fact that punctuation and paragraphing are *not* part of the language system. Students involved in the practice of extensive reading, which is to say sustained periods of reading for *meaning*, whilst necessarily engaging (consciously or subconsciously) with the syntax at sentence level, and with the broader relationships and connections amongst units of text at the higher levels of overall text coherence and meaning, may not have engaged with conventions of presentation for long enough to have used long-term memory. Fuster (1995) identifies two types of working memory, one which has some measure of "future perspective", and one which has as its sole purpose the execution of a small local task. The contents of this second type of working memory are immediately discarded upon task completion. Whilst punctuation may have assisted the comprehension process by showing where lower-level, or local, units of meaning started and stopped, and signposting useful breaks in the process of meaning reconstruction, it did not in itself carry actual meaning. A student may have briefly registered the presence of a comma or a full-stop, used this as a break marker for meaning processing, and immediately discarded it, since it had no interactive function in the parsing of grammar and vocabulary, but a simple, categorical task which, once done, was done and needed no revisiting. Thus the lack of effect of reading on *punctuation and paragraphing* may have been due to the very different role of this construct from those of the language-related constructs. Punctuation and paragraphing, in effect, were mere technical facilitators within the process of meaning extraction.

An interesting exception to this non-linguistic (and non meaning-carrying) role of punctuation was the use of inverted commas to indicate direct speech. These may have greater meaning-related impact than commas or full-stops, since, in order to comprehend

otherwise confusing changes in, for example, subject pronouns (*e.g.* the sudden switch from "he" to "I") and verb tenses, direct-speech inverted commas must be noticed and attended to in the extraction of meaning. They might be said to be part of the text's plot, and may act as replacements for phrases such as "he said". Although this was not investigated quantitatively, the correct use of inverted commas for direct speech was more in evidence in the experimental compositions than in the control compositions, where students often neglected to use them.

Spelling is also part of only the written code, and, similarly, is unlikely to interact developmentally with fluency, complexity or grammatical accuracy. Although rater-judged *spelling* scores were not normally distributed, showing that, with this particular data, this was not a very suitable construct for holistic judgements, there was nonetheless some slight evidence of possible frequency effects at the lower level. As with accuracy, it may be easier to show relative improvement in spelling at lower levels, when even high-frequency vocabulary may still be subject to spelling inconsistencies. Counts of numbers of mistakes showed the same pattern as rater judgements, confirming a slight effect at the lower level and no effect at the higher level. However, as noted in Sections 5.4 and 6.12.2, raters' judgements may have been directly related to numbers of mistakes. In practice, spelling mistakes were so relatively infrequent in the data (showing mean numbers which ranged from 2.5 to 3.3 spelling mistake *types* per composition, across the four classes) that it is not possible to draw firm conclusions as to any effects of extensive reading on spelling.

7.5 Limitations and future research

As with much other research into L2 teaching method effects, the findings from this study pertain to a particular group of learners. The age of the learners, the L1, and the instructional setting may all have some bearing on the observed findings. Although present evidence from neurolinguistic research suggests there may be no such thing as a language-specific acquisition device, or LAD, which "shuts down" some time during adolescence, the human brain may nonetheless lose plasticity with ageing. What has been hypothesized by psycholinguists to be the closing down of the LAD is not specific to language acquisition. All skills may become harder to proceduralize as we grow older. Research is therefore needed to investigate whether extensive reading can have the same effects on fluency with

older learners, and if exposure to instances can result in the same kind of frequency effects as were observed for secondary schoolchildren.

The L1 may also be an important factor. Chinese is not a phonetic language, such that the written form of a word does not systematically relate in any way to its spoken sound. The former must simply be memorized. Chinese is also primarily an analytic language, which depends heavily on lexis and word order, and uses fewer morphological procedures than do more synthetic languages. For example, "He went yesterday", "He is going today" and "He will go tomorrow" differ from each other in Chinese only in the selection of the time adverbial. The verb does not change. It would be interesting to investigate whether similarity between the basic structural systems of an L1 and L2 (such as exists between Spanish and English) accelerates progress in fluency associated with extensive reading, leads to a less pronounced effect, or is irrelevant. It also remains to be seen whether extensive reading effects can impact on expository writing of the kind that is required in EAP programmes, and if reading relatively easy *narrative* text can do as much to help expository writing skills as narrative writing skills.

I believe that this study is the first which has attempted to distinguish the distinct cognitive processes engaged by L2 extensive reading (as compared to other types of text-based language task), suggesting, alongside this, particular ways in which the L2 might be expected to benefit from this kind of reading practice. It may be fruitful to continue in this direction. In the absence of such a rich source of field data as was provided by the Hong Kong ERS, we might investigate more specific hypotheses. For example, we could attempt to chart the relationship between L2 extensive reading practice and reading speed; research might be undertaken to confirm the relationship between reading speed and writing speed; we might investigate, also, the connection between (natural) writing speed and syntactic complexity, and text fluency and coherence.

Speed of written language access and production has received little attention as a marker of differences between L2 proficiency levels or developmental stages. However, it may be of much greater importance than is generally allowed by L2 evaluation procedures, which, other than imposing a single time limit to particular tasks, rarely attempt to capture differences in speed of execution. Indeed, unlike speaking and listening evaluation, reading and writing evaluation tends to focus on language *knowledge* (of grammar and vocabulary in

particular) with little enquiry into *competence* (as in fluency and ease of execution). This is another potential area for future investigation.

As regards the present study, error was operationalized only as being either present or absent. It could prove informative to investigate exactly what *kinds* of errors were more, or less, prevalent in the compositions of the reading-scheme and non reading-scheme students, and, from there, to explore more precisely what kinds of error may be amenable to "auto-correction" from implicit exposure, and which are more resilient to implicit learning processes, hence require more intensive classroom focus.

So far, there has been little interface between research into the effects of extensive reading programmes and theory deriving from other Applied Linguistics disciplines, in particular cognitive SLA. Extensive reading research has also tended to focus on questions concerning the input, rather than on the *process* of extensive reading. In fact, the exact nature of the input, so long as it is relatively easily understood and engaging for the learner, may be of less importance than we think. It is to be hoped that future research into extensive reading will continue to be informed by relevant findings in cognitive SLA, psycholinguistics and neurolinguistics.

7.6 Conclusion

Extensive reading in an L2 promotes writing fluency. It is likely that reading practice speeds up language recognition, or *access*, which, in turn, affects speed of *retrieval* for language production. It has been argued here that this increase in speed is directly related to increased levels of L2 *automaticity*, resulting from the proceduralization of the L2 caused by reading practice. It is the relative ease of extensive reading materials which makes this possible, by enabling the learner to access greater quantities of input than is the case when reading is effortful, thereby activating frequency effects. The learner is also relieved of the need to consciously focus on linguistic *form*, and may focus instead on meaning. This, too, facilitates proceduralization.

The benefits of extensive reading extend beyond enhanced speed, or fluency, of written language production. Fluency underpins other aspects of L2 development, notably favouring the concurrent development of syntactic complexity. It may also favour the development of coherence, grammatical accuracy and the acquisition of common morphological features.

Increased levels of complexity and coherence may be facilitated by the more fluent L2 user's ability to retrieve more of the language within a single working-memory span; grammar improvement may be the result of frequency effects enabled initially by reading simple, hence more, text, but enhanced also by improved reading fluency which permits yet more text to be accessed. Frequent exposure may, in turn, lead to a reduction of conscious focusing on form, thereby facilitating proceduralization.

The clearest benefits of extensive reading — enhanced fluency, complexity and coherence — are in constructs which are not, on the whole, formally addressed in the normal classroom. These constructs are difficult to operationalize, both for purposes of explicit instruction and for subsequent classroom testing. Levels of fluency and complexity cannot be "right" or "wrong" in the way that use of grammar and vocabulary may be, and the normal, supportive teacher does not criticise these, nor does current educational ethos require teachers to draw attention to them, but, instead, to encourage students each at their own working level. In any case, it is unlikely that explicit instruction, if such could be devised, could have much effect on fluency and complexity, and these are more likely to develop naturally as a result of practice. Coherence differs from the other two in that attempts at coherence can result in language that is formally wrong and so may be corrected by a teacher, but coherence is far too complex and instance-specific to be routinely taught.

Extensive reading, then, may be seen as a very useful complement to formal classroom instruction, since it aids the development of language features not normally addressed by formal instruction. Indeed, it may be this that lies behind the common teacher's intuition that reading is beneficial to the learner's L2 although it is very hard to pinpoint — and to "prove" — just exactly *how* this may be so. It may be no coincidence that, as is frequently remarked by L2 professionals, longer compositions are generally rated more highly than shorter ones. It may be not so much the length of a composition which influences raters as the enhanced levels of complexity and coherence — and possibly other "invisible" features — with which increased length tends to co-occur, since these are all correlates of the same developmental progress. In other words, of the complete range of fluency effects, increased length is the most immediately and categorically visible. (It might be recalled here that, in the case of school 4, "length", independently of any interaction with other variables, was uniquely responsible for over 7% of the unexplained variance in raters' *overall quality* judgements.)

Extensive reading has not been without its detractors. Bruton (2002) has claimed extensive reading to be no different from "reading extensively". If "reading extensively" allows for slow and effortful reading, with conscious deciphering of vocabulary, morphology and text structure, then the two are very different, and will engender different types of learning. A defining feature of *extensive reading* is the type of ("authentic") reading process it engenders, allowing the L2 learner to focus on meaning, and to hold longer stretches of text in working memory, thereby engaging more nearly in the same kind of processes as a reader with an L1 text. Green has criticised extensive reading as being less useful than a programme of task-based reading assignments, likening the typical Hong Kong ERS class to "a particularly monastic detention session" (2005: 308). Whilst Green, in fact, offers no evidence to support his opinion, but bases his claim entirely on theoretical arguments, he nonetheless pinpoints what may be the greatest weakness of extensive reading as a pedagogical practice — the apparent impotence of the teacher to intervene in the learning process.

To a certain extent, extensive reading lacks face validity. A teacher may find it difficult to sit back and allow students to take responsibility for their own learning, particularly in the prevalent L2 teaching climate of communicative and task-based methodologies. In response to Green's claim that a task-based reading programme may be more advantageous than an extensive reading programme, one might argue, again, that the two are entirely different and will result in different types of learning. Task-based learning — theoretically — is likely to result in incidental learning, probably of vocabulary and grammar structures. Extensive reading is more likely to result in proceduralization, leading to enhanced fluency, coherence and complexity, and the potential automatization, or acquisition, of morphology. A major advantage of extensive reading over task-based teaching practices is that extensive reading does not permit the possible input of incorrect language, thus helping, ultimately, to prevent proceduralization — or fossilization — of incorrect language items. Additionally, language which once has been proceduralized may stand the learner in good stead years later, with relatively little ongoing maintenance. Language which has been learned via declarative memory is more likely to be forgotten if the use of the L2 is allowed to lapse for any length of time.

Apart from teacher resistance, probably the biggest challenge faced by educators wishing to deploy extensive reading as a teaching method is that of student motivation. As Schumann points out, "motivation is not independent of cognition (as it is frequently treated in SLA

research), but, instead, it is part of cognition" (2004: 3). All input to the human brain is subject to stimulus appraisal, and without a positive or negative stimulus appraisal (*i.e.* cognitive motivation; positive stimulus appraisal involves a release of dopamine) no learning of any sort will ever take place. In the L2 classroom, positive stimulus may take on a variety of forms. If there is no social compatibility or self-realisation motivation (*i.e.* if the L2 is redundant for the purposes of everyday life), other, minor, motivational drives come into play. These may take the form of a teacher's approval, a smile, group acceptance, the achievement of being understood by a fellow student, expressing a view or opinion, getting the "right answer". All these can be stage-managed by a teacher. However, the student who is involved in individual silent reading is largely beyond the teacher's help. Motivation must be either intrinsic to the student (who may actively wish to learn the L2, or who may feel very positively about reading or even about a particular reading programme) or must come from the input. Levels of interest, novelty and pleasantness need to be maintained, and these are probably more important in a graded reader than even the exact linguistic structure of the input text.

To sum up, this study has shown that a programme of extensive reading in conjunction with formal classroom instruction is a very effective method of L2 learning. Students who spent two classes a week engaged in such a programme showed enhanced levels of coherence, complexity and fluency, as compared to peer groups who engaged, instead, in two classes of normal coursebook-based instruction. Additionally, grammar and vocabulary knowledge was maintained to the same levels as those of the students who had two more classes of normal instruction per week, showing that the practice of extensive reading may be just as effective in reinforcing formal language knowledge, if this is presented first during normal class, as any other classroom practice.

Extensive reading is not suggested here as a replacement for L2 instruction, but as a complement. MacWhinney has suggested that "students who receive explicit instruction, as well as implicit exposure to forms, would seem to have the best of both worlds" (1997: 278). I would like to extend this beyond "forms" and propose that students who receive explicit instruction, implicit exposure to forms *and* extensive implicit (meaning-focused) exposure to the indefinable something that holds forms together have an even better world. It is, however, probably the *practice* in accessing the L2, even more than any such exposure, leading to faster and faster processing and in turn enabling other developments, which constitutes the greatest benefit of extensive reading.

REFERENCES

Abutalebi, J., S.F. Cappa and D. Perani. 2005. What can functional neuroimaging tell us about the bilingual brain? In Kroll, J. F. and A. M. B. De Groot (eds). 2005. *Handbook of bilingualism: psycholinguistic approaches*. Oxford. Oxford University Press: 497-515

Aglioti, S. 1999. Language and memory systems. In Fabbro, F. and R.E. Asher (eds). 1999. *Concise encyclopedia of language pathology*. Oxford. Pergamon Press: 371-377

Alderson, J. C., C. Clapham and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge. Cambridge University Press

Altenberg, B. and S. Granger. 2001. The grammatical and lexical patterning of *make* in native and non-native student writing. *Applied Linguistics*, **22(2)**: 173-194

Anderson, J.R. 1993. *Rules of the mind*. Hillsdale, NJ. Lawrence Erlbaum

Anderson, J.R. and C. Lebiere. 1998. *The atomic components of thought*. Mahwah, NJ. Lawrence Erlbaum

Anderson, R. 1991. Developmental sequences: the emergence of aspect marking in second language acquisition. In Huebner, T. and C. Ferguson (eds). 1991. *Cross-currents in second language acquisition and linguistics theories*. Amsterdam. John Benjamins: 305-324

Bardovi-Harlig, K. 1995. A narrative perspective on the development of the tense/aspect system in second language acquisition. *Studies in Second Language Acquisition*, **17**: 263-291

Bauman, J. 1996. Vocabulary Resources for Material Writers. *The Materials Writers Newsletter: the newsletter of the materials writers' national special interest group of the Japan Association of Language Teachers*, **4(3)**. Retrieved March 2003 from http://jbauman.com/mat.art.html

Bell, T. 2001. Extensive reading: speed and comprehension. *The Reading Matrix*, **1(1)**. Retrieved December 2002 from www.readingmatrix.com/articles/bell/index.html

Blau, E. K. 1982. The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly*, **16(4)**: 517-528

Brown, J.D. 1990. The use of multiple t-tests in language research. *TESOL Quarterly*, **24(4)**: 770-773

Brown, J.D., T. Hilgers and J. Marsella. 1991. Essay prompts and topics: minimizing the effect of mean differences. *Written Communication*, **8**: 533-556

Bruton, A. 2002. Extensive reading is reading extensively, surely? *The Language Teacher*, **26/11**: 23-25

Carson, J., P. Carrell, S. Silberstein, B. Kroll and P. Kuehn. 1990. Reading-writing relationships in first and second language. *TESOL Quarterly*, **24(2)**: 245-266

Casanave, C. 1994. Language development in students' journals. *Journal of Second Language Writing*, **3**: 17-34

Claridge, G. 2005. Simplification in graded readers: measuring the authenticity of graded texts. *Reading in a foreign language*, **17(2)**: 144-158

Coady, J. 1997. L2 vocabulary acquisition through extensive reading. In Coady, J. and T. Huckin (eds). 1997. *Second language vocabulary acquisition.* Cambridge. Cambridge University Press: 225-237

Cobb, T. *n.d.* Web VocabProfile (Version 1.5). Retrieved April /May 2003 from http://www.er.uqam.ca/nobel/r21270/cgi-bin/webfreqs/web_vp.cgi

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ. Erlbaum

Constantino, R., S.Y. Lee, K.S. Cho and S. Krashen. 1997. Free voluntary reading as a predictor of TOEFL scores. *Applied Language Learning*, **8**: 111-118

Cooper, M. 1984. Linguistic competence of practiced and unpracticed non-native readers of English. In Alderson, C. and A.H. Urquhart (eds). 1984. *Reading in a foreign language*. Harlow. Longman: 122-135

Corder, S. P. 1974. Error analysis. In Allen, J. and S. P. Corder (eds). 1974. *The Edinburgh Course in Applied Linguistics, Vol. 3*. Oxford. Oxford University Press: 122-154

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* **34(2)**: 213-238

Crowell, S.E. 2004: The neurobiology of declarative memory. In Schumann, J., S.E. Crowell, N. Jones, N. Lee, S.A. Schuchert and L.A. Wood. 2004. *The neurobiology of learning: perspectives from second language acquisition.* Mahwah, NJ. Lawrence Erlbaum: 75-109

Crystal, D. 1988. *Rediscover grammar*. Harlow. Longman

Cushing Weigle, S. 2002. *Assessing Writing*. Cambridge. Cambridge University Press

Dabrowska, E. 2004. *Language, mind and brain: some psychological and neurological constraints on theories of grammar*. Edinburgh. Edinburgh University Press

Davidson, P. and E. Williams. 2005. Outcome evaluation of extensive reading programmes in Africa. *IATEFL 2005 Cardiff Conference Selections*: 76-77

Davies, A. 1984. Simple, simplified and simplification: what is authentic? In Alderson, C. and A.H. Urquhart (eds). 1984. *Reading in a foreign language*. Harlow. Longman: 181-195

Davies, A. and A. Irvine. 1992a. The EPER test of extensive reading. Edinburgh. University of Edinburgh

Davies, A. and A. Irvine. 1992b. The EPER test of vocabulary. Edinburgh. University of Edinburgh

Davies, A. and A. Irvine. 1992c. *The EPER Hong Kong Test Project Final Report.* Edinburgh. University of Edinburgh

Day, R., C. Omura and M. Hiramatsu. 1991. Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, **7(2)**: 541-551

Day, R.R. and J. Bamford. 1998. *Extensive reading in the second language classroom.* Cambridge. Cambridge University Press

DeKeyser, R. 1995. Learning second language grammar rules: an experiment with a miniature linguistic system. *Studies in Second Language Acquisition*, **17**: 379-410

DeKeyser, R. 1997. Beyond explicit rule learning: automatizing second language morphosyntax. *Studies in Second Language Acquisition*, **19**: 195-221

Del Gobbo, F. 2005. Chinese relative clauses: restrictive, descriptive or appositive? In Bruge, L., G. Giusti, N. Munaro, W. Schweikert and G. Turano (eds). 2005. *Contributions to the 30th Incontro di Grammatica Generativa.* Venice. Cafoscarina: 287-305

Doughty, C.J. 2001. Cognitive underpinnings of focus on form. In Robinson P. (ed). 2001. *Cognition and second language instruction.* Cambridge. Cambridge University Press: 206-257

Doughty, C.J. 2003. Instructed SLA: Constraints, compensation, and enhancement. In Doughty, C.J. and M.H. Long (eds). 2003. *The handbook of second language acquisition.* Oxford. Blackwell: 256-310

Doughty, C.J. and J. Williams (eds). 1998. *Focus on form in classroom second language acquisition.* Cambridge. Cambridge University Press

Dupuy, B. and S. Krashen. 1993. Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning*, **4(1)**: 55-63

Elley, W.B. 1991. Acquiring literacy in a second language: the effect of book-based programs. *Language Learning*, **41(3)**: 375-411

Elley, W.B. and F. Mangubhai. 1981. *The impact of a book flood in Fiji primary schools: Studies in South Pacific Education, No 1.* Wellington. New Zealand Council for Educational Research

Elley, W.B. and F. Mangubhai. 1983. The impact of reading on second language learning. *Reading Research Quarterly*, **19(1)**: 53-67

Elley, W.B., B. Cutting, F. Mangubhai and C. Hugo. 1996. Lifting literacy levels with story books: evidence from the South Pacific, Singapore, Sri Lanka and South Africa. In *Proceedings of the 1996 World Conference on Literacy.* Retrieved January 2002 from http://litserver.literacy.upenn.edu/products/ili/webdocs/ilproc/ilprocwe.htm

Ellis, N.C. 1994. Implicit and explicit language learning — An overview. In Ellis, N.C.(ed). 1994. *Implicit and explicit learning of languages.* London. Academic Press: 1-31

Ellis, N.C. and N. Laporte. 1997. Contexts of acquisition: effects of formal instruction and naturalistic exposure on second language acquisition. In de Groot, A.M.B. and J.F. Kroll (eds). *Tutorials in bilingualism: psycholinguistic perspectives*. Mahwah, NJ. Lawrence Erlbaum: 53-83

Ellis, N.C. and R. Schmidt. 1997. Morphology and longer distance dependencies: laboratory research illuminating the *A* in *SLA*. *Studies in Second Language Acquisition*, **19**: 145-171

Ellis, N.C. and R. Schmidt. 1998. Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax. In Plunkett, K. (ed). 1998. *Language acquisition and connectionism: a special issue of Language and Cognitive Processes*. Hove. Psychology Press: 307-336

Ellis, R. 1987. Interlanguage variability in narrative discourse: style shifting in the use of the past tense. *Studies in Second Language Acquisition*, **9(1)**: 1-20

Ellis, R. 1994. *The study of second language acquisition*. Oxford. Oxford University Press.

Ellis, R. and G. Barkhuizen. 2005. *Analysing learner language*. Oxford. Oxford University Press

Engber, C.A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, **4(2)**: 139-155

Fuster, J.M. 1995. *Memory in the cerebral cortex*. Cambridge, MA. MIT Press.

Goldschneider, J.M. and R. DeKeyser. 2001. Explaining the "natural order of L2 morpheme acquisition" in English: a meta-analysis of multiple determinants. *Language Learning*, **51(1)**: 1-50

Goodman, K.S. 1976. Behind the eye: what happens in reading. In Singer, H. and R.B. Ruddell (eds). 1976. *Theoretical models and processes of reading: 2nd edition*. Newark, DE. International Reading Association: 470-96

Green, C. 2005. Integrating extensive reading in the task-based curriculum. *ELT Journal*, **59(4)**: 306-311

Hafiz, F.M. and I.Tudor. 1989. Extensive reading and the development of language skills. *ELT Journal*, **43(1)**: 4-13

Hafiz, F.M. and I.Tudor. 1990. Graded readers as an input medium in L2 learning. *System*, **18(1)**: 31-42

Hamp-Lyons, L. 1991. *Holistic writing assessment of LEP students: CRAL research report*. Denver. University of Colorado

Harley, T.A. 1995. *The psychology of language: from data to theory*. Hove. Psychology Press Ltd.

Hayashi, K. 1999. Reading strategies and extensive reading in EFL classes. *RELC Journal*, **30(2)**: 114-132

Hayes, J.R. 1996. A new framework for understanding cognition and affect in writing. In Levy, C.M. and S. Ransdell. 1996. *The science of writing: theories, methods, individual differences and applications*. Mahwah, NJ. Lawrence Erlbaum: 1-27

Hayes, J.R. and L.S. Flower. 1980. Identifying the organization of writing processes. In Gregg, L. and E.R. Steinberg (eds.) 1980. *Cognitive processes in writing*. Hillsdale, NJ. Lawrence Erlbaum: 3-30

Hedge, T. 2000. *Teaching and learning in the language classroom*. Oxford. Oxford University Press

Hedgecock, J. and D. Atkinson. 1993. Differing reading-writing relationships in L1 and L2 literacy development? *TESOL Quarterly*, **27(2)**: 329-333

Henning, G. 1987. *A guide to language testing*. Cambridge, Massachusetts. Newbury House

Hill, D. R. 1992. *The EPER guide to organising programmes of extensive reading*. Edinburgh. Institute for Applied Language Studies, University of Edinburgh

Hill, D.R. 1997. Survey review: graded readers. *ELT Journal*, **51(1)**: 57-81

Hill. D.R. 2001. Survey of graded readers. *ELT Journal*, **55(3)**: 300-324

Hill, D.R. and H. Reid-Thomas. 1988. Survey review: graded readers (part 1). *ELT Journal*, **42(1)**: 44-52

Hinkel, E. 2002. *Second language writers' text: linguistic and rhetorical features*. Mahwah, NJ. Lawrence Erlbaum

Hirano, K. 1991. The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan*, **2**: 21-30

Homburg, T.J. 1984. Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly*, **18(1)**: 87-107

Honeyfield, J. 1977. Simplification. *TESOL Quarterly*, **11(4)**: 431-440

Hong Kong Reading Association. *n.d.* HKRA Standardized Test of Reading. Hong Kong. HKRA

Horst, M., T. Cobb and P. Meara. 1998. Beyond a clockwork orange: acquiring second language vocabulary through reading. *Reading in a Foreign Language*, **11(2)**: 207-223

Housen, A. 2002. A corpus-based study of the L2-acquisition of the English verb system. In Granger, S., J. Hung and S. Petch-Tyson (eds). 2002. *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam. John Benjamins: 77-116

Hu, M. and P. Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, **13(10)**: 403-430

Hulstijn, J. 2001. Intentional and incidental second language vocabulary learning: a reappraisal of elaboration, rehearsal and automaticity. In Robinson, P. (ed). 2001. *Cognition and second language instruction*. Cambridge. Cambridge University Press: 258-286

Hunt, K.W. 1970. Do sentences in the second language grow like those in the first? *TESOL Quarterly*, **4(3)**: 195-202

Hurford, J. R. 1994. *Grammar: a student's guide*. Cambridge. Cambridge University Press

Ishikawa, S. 1995. Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, **4(1)**: 51-70

Izumi, S. 2002. Output, input enhancement, and the noticing hypothesis: an experimental study on ESL relativization. *Studies in Second Language Acquisition*, **24**: 541-577

Jacobs, B. 2004. Foreword to Schumann, J., S. Crowell, N. Jones, N. Lee, S.A. Schuchert and L.A. Wood. 2004. *The neurobiology of learning: perspectives from second language acquisition.* Mahwah, NJ. Lawrence Erlbaum: ix-x

Jacobs, H.L., S.A. Zinkgraf, D.R. Wormuth, V.F. Hartfiel and J.B. Hughey. 1981. *Testing ESL composition: a practical approach*. Rowley, MA. Newbury House

Janopoulos, M. 1986. The relationship of pleasure reading and second language writing proficiency. *TESOL Quarterly*, **20(4)**: 763-768

Johnson, R.K. and P.L.M. Lee. 1987. Modes of instruction: teaching strategies and student responses. In Lord, R. and H.N.L. Cheng. 1987. *Language education in Hong Kong*. Hong Kong. Chinese University Press: 99-121

Krashen, S. 1982. *Principles and practice in second language acquisition*. Oxford. Pergamon

Krashen, S. 1985. *The input hypothesis: issues and implications*. London. Longman.

Krashen, S. 1989. We acquire vocabulary and spelling by reading: additional evidence for the input hypothesis. *The Modern Language Journal*, **73(4)**: 440-464

Krashen, S. 1993. *The power of reading: insights from the research*. Englewood, CO. Libraries Unlimited, Inc.

Lai, F.K. 1993. The effect of a summer reading course on reading and writing skills. *System*, **21(1)**: 87-100

Laufer, B. 1992. How much lexis is necessary for reading comprehension? In Arnaud, P.J.L. and H. Bejoint (eds). 1992. *Vocabulary and applied linguistics*. London. Macmillan: 126-132

Laufer, B. 1998. The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, **19(2)**: 255-271

Laufer, B. 2005. Lexical frequency profiles: from Monte Carlo to the real world. *Applied Linguistics*, **26(4)**: 582-588

Laufer, B. and J. Hulstijn. 2001. Incidental vocabulary acquisition in a second language: the construct of task-induced involvement. *Applied Linguistics*, **22(1)**: 1-26

Laufer, B. and P. Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, **16(3)**: 307-322

Lee, Y.O., S. Krashen and B. Gribbons. 1995. The effect of reading on the acquisition of English relative clauses. *I.T.L.* (Review of Applied Linguistics, Université catholique de Louvain, Belgium), **113-114**: 263-273

Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning*, **40(3)**: 387-417

Leow, R.P. 1993. To simplify or not to simplify: a look at intake. *Studies in Second Language Acquisition*, **15**: 333-355

Levelt, W. 1989. *Speaking: from intention to articulation.* Cambridge, MA. MIT Press

Liu, N and I.S.P. Nation. 1985. Factors affecting guessing vocabulary in context. *RELC Journal*, **16(1)**: 33-42

Long, M.H. 1983. Does second language instruction make a difference? A review of research. *TESOL Quarterly*, **17(3)**: 359-382

Long, M.H. 2000. Focus on form in task-based language teaching. In Lambert, R.L. and E. Shohamy (eds). 2000. *Language policy and pedagogy*. Amsterdam. John Benjamins: 179-192

Lumley, T. 2002. Assessment criteria in a large-scale writing-test: what do they really mean to the raters? *Language Testing*, **19(3)**: 246-276

Lynch, T. 1996. *Communication in the language classroom*. Oxford. Oxford University Press

MacWhinney, B. 1997. Implicit and explicit processes: commentary. *Studies in Second Language Acquisition*, **19**: 277-281

Malvern, D. and B. Richards. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, **19(1)**: 85-104

Mason, M. and S. Krashen. 1997. Extensive reading in English as a foreign language. *System*, **25(1)**: 91-102

Milner, P. 1999. *The autonomous brain: a neural theory of attention and learning*. Mahwah, NJ. Lawrence Erlbaum

Morris, L. and T. Cobb. 2004. Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, **32**: 75-87

Nagy, W. 1997. On the role of context in first- and second-language vocabulary learning. In Schmitt, N. and M. McCarthy (eds). 1997. *Vocabulary: description, acquisition and pedagogy*. Cambridge. Cambridge University Press: 64-83

Nagy, W.E. and P.A. Herman. 1985. Incidental vs. instructional approaches to increasing reading vocabulary. *Educational Perspectives*, **23**: 16-21

Nagy, W.E., P.A. Herman and R.C. Anderson. 1985. Learning words from context. *Reading Research Quarterly*, **20(2)**: 233-253

Nation, I. S. P. 2001. *Learning vocabulary in another language*. Cambridge. Cambridge University Press

Nation, P. and R. Waring. 1997. Vocabulary size, text coverage and word lists. In Schmitt, N. and M. McCarthy (eds). 1997. *Vocabulary: description, acquisition and pedagogy*. Cambridge. Cambridge University Press: 6-19

Norris, J.M. and L. Ortega. 2000. Effectiveness of L2 instruction: a research synthesis and quantitative meta-analysis. *Language Learning*, **50(3)**: 432-528

Nuttall, C. 1996. *Teaching reading skills in a foreign language: new edition*. Oxford. Heinemann

Oakhill, J. and A. Garnham. 1988. *Becoming a skilled reader*. Oxford. Blackwell

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

Pallant, J. 2001. *SPSS survival manual*. Maidenhead. Open University Press

Paradis, M. 1994. Neurolinguistic aspects of implicit and explicit memory: implications for bilingualism and SLA. In Ellis, N. C. (ed). 1994. *Implicit and explicit learning of languages*. London. Academic Press: 393-419

Paran, A. 1996. Reading in EFL: facts and fictions. *ELT Journal*, **50(1)**: 25-34

Perfetti, C.A. 1985. *Reading ability*. Oxford. Oxford University Press

Perfetti, C.A. 1988. Verbal efficiency theory in reading ability. In Daneman, M., G.E. MacKinnon and T.G. Waller (eds). 1988. *Reading research: advances in theory and practice*. New York. Academic Press: 109-143

Perfetti, C.A. and A.M. Lesgold. 1977. Discourse comprehension and sources of individual differences. In Just, M. and P.A. Carpenter (eds). 1977. *Cognitive processes in comprehension*. Hillsdale, NJ. Lawrence Erlbaum: 141-183

Pienemann, M. 1998. *Language Processing and Second Language Development*. John Benjamins. Amsterdam

Pinker, S. 1991. Rules of language. *Science*, **253**: 530-535

Pinker, S. and M. T. Ullman. 2002. The past and future of the past tense. *Trends in Cognitive Sciences*, **6(11)**: 456-463

Pitts, M., H. White and S. Krashen. 1989. Acquiring second language vocabulary through reading: a replication of the clockwork orange study using second language acquirers. *Reading in a Foreign Language*, **5(2)**: 271-275

Polio, C. 1997. Measures of linguistic accuracy in second language writing research. *Language Learning*, **47(1)**: 101-143

Polio, C. 2001. Research methodology in second language writing research: the case of text-based studies. In Silva, T. and P. K. Matsuda (eds). 2001. *On second language writing.* Mahwah, NJ. Lawrence Erlbaum: 91-115

Polio, C., C. Fleck and N. Leder. 1998. "If only I had more time": ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, **7**: 43-68

Pollitt, A. 1991. Response to Charles Alderson's paper 'Bands and Scores'. *Language Testing in the 1990s*, **1(1)**: 87-94

Quirk, R. and S. Greenbaum. 1979. *A university grammar of English (ninth impression).* London. Longman

Renandya, W.A., B.R.S. Rajan and G.M. Jacobs. 1999. Extensive reading with adult learners of English as a second language. *RELC Journal*, **30(1)**: 39-61

Richards, J.C. and W.A. Renandya. 2002. *Methodology in language teaching: an anthology of current practice.* Cambridge. Cambridge University Press

Robb, T., S. Ross and I. Shortreed. 1986. Salience of feedback on error and its effect on EFL writing quality. *TESOL Quarterly*, **20(1)**: 83-95

Robb, T.N. and B. Susser. 1989. Extensive reading vs. skills building in an EFL context. *Reading in a Foreign Language*, **5(2)**: 239-251

Romaine, S. 2003. Variation. In Doughty, C. and M. Long (eds). 2003. *The Handbook of Second Language Acquisition.* Blackwells. Oxford: 409- 435

Rott, S. 1999. The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, **21**: 589-619

Salaberry, M. R. 2000. The acquisition of English past tense in an instructional setting. *System*, **28(1)**: 135-152

Saragi, T., P. Nation and G. Meister. 1978. Vocabulary learning and reading. *System*, **6**: 70-78

Schmidt, R. 1990. The role of consciousness in second language learning. *Applied Linguistics*, **11**: 129-158

Schmidt, R. 2001. Attention. In Robinson, P. (ed). 2001. *Cognition and second language instruction.* Cambridge. Cambridge University Press: 3-32

Schmitt. N. 1998. Tracking the incremental acquisition of second language vocabulary: a longitudinal study. *Language Learning*, **48(2)**: 281-317

Schumann, J.H. 2004. Introduction to Schumann, J., S. Crowell, N. Jones, N. Lee, S.A. Schuchert and L.A. Wood. 2004. *The neurobiology of learning: perspectives from second language acquisition.* Mahwah, NJ. Lawrence Erlbaum: 1-6

Science Research Associates. 1959. SRA Reading Laboratory. Chicago, Ill. SRA

Scott, M. 1998. WordSmith Tools (Version 3.0). Oxford University Press

Segalowitz, N. and J. Hulstijn. 2005. Automaticity in bilingualism and second language learning. In Kroll, J. F. and A. M. B. De Groot (eds). 2005. *Handbook of bilingualism: psycholinguistic approaches*. Oxford. Oxford University Press: 371-388

Segalowitz, N., C. Poulsen and M. Komoda. 1991. Lower level components of reading skill in higher level bilinguals: implications for reading instruction. *AILA Review*, **8** ("Reading in two languages"): 15-30

Shanahan, T. 1984. Nature of the reading-writing relation: an explanatory multivariate analysis. *Journal of Educational Psychology*, **76**: 466-477

Shanahan, T. and R. Lomax. 1986. An analysis and comparison of theoretical models of the reading-writing relationship. *Journal of Educational Psychology*, **78(2)**: 116-123

Shelton, S. *n.d.* Encouraging extensive reading. *DevelopingTeachers.com*. Retrieved October 2005 from www.developingteachers.com/articles_tchtraining/extread1_scott.htm

Shlayer, J. 1996. Extensive reading. *English Teachers' Journal*, **49**: 32-33

Skehan, P. 1998. *A cognitive approach to language learning.* Oxford. Oxford University Press

Skehan, P. and P. Foster. 1999. The influence of task structure and processing conditions on narrative retellings. *Language Learning*, **49(1)**: 93-120

Skehan, P. and P. Foster. 2001. Cognition and tasks. In Robinson, P. (ed). 2001. *Cognition and second language instruction.* Cambridge. Cambridge University Press: 183-205

Stanovich, K.E. 1980. Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, **16**: 32-71

Swain, M. 1985. Communicative competence: some roles of comprehensible input and comprehensible output in its development. In Gass, S.M. and C.G. Madden (eds). 1985. *Input in second language acquisition.* Rowley, MA. Newbury House: 235-256

Swain, M. 1993. The output hypothesis: just speaking and writing aren't enough. *Canadian Modern Language Review*, **50(1)**: 158-64

Tabachnick, B.G. and L.S. Fidell. 1996. *Using multivariate statistics* (3rd edition). New York. Harper Collins

Tedick, D. 1990. ESL writing assessment: subject matter knowledge and its impact on performance. *English for Specific Purposes*, **9**: 123-143

Tommola, J. 1979. Some parameters of simplification. Communication studies paper, MSc in Applied Linguistics. Edinburgh. Dept. of Linguistics, University of Edinburgh, mimeo

Treffers-Daller, J. 2004. Report on BAAL/CUP Seminar: Vocabulary knowledge and use: measurement and applications. *BAAL News*, **77**: 24

Tsang, W.K. 1996. Comparing the effects of reading and writing on writing performance. *Applied Linguistics*, **17(2)**: 210-233

Tudor, I and F. Hafiz. 1989. Extensive reading as a means of input to L2 learning. *Journal of Research in Reading*, **12(2)**: 164-178

Tweissi, A.I. 1998. The effects of the amount and type of simplification on foreign language reading comprehension. *Reading in a Foreign Language*, **11(2)**: 191-206

Ullman, M. 2001a. A neurocognitive perspective on language: the declarative/procedural model. *Nature Reviews Neuroscience*, **2**: 717-726

Ullman, M. 2001b. The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and Cognition*, **4(1)**: 105-122

Ullman, M., S. Corkin, M. Coppola, G. Hickok, J. Growdon, W. Koroshetz and S. Pinker. 1997. A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, **9**: 266-276

VanPatten, B. 1996. *Input processing and grammar instruction: theory and research.* Norwood, NJ. Ablex

Vicary, T. 2006. *The hitch-hiker*. Oxford. Oxford University Press

Walczyk, J. 2000. The interplay between automatic and control processes in reading. *Reading Research Quarterly*, **35(4)**: 554-566

Waring, R. and M. Takaki. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, **15(2)**: 130-163

West, M. 1953. *A general service list of English words*. London. Longman

White, E. M. 1984. Holisticism. *College Composition and Communication*, **35(4)**: 400-409

Widdowson, H.G. 1978. *Teaching language as communication*. Oxford. Oxford University Press

Wolfe-Quintero, K., S. Inagaki and H. Kim. 1998. *Second language development in writing: measures of fluency, accuracy and complexity: Technical Report 17*. Honolulu. University of Hawai'i

Xue G. and I. S. P. Nation. 1984. A university word list. *Language Learning and Communication*, **3(2)**: 215-229

Yang, A. 2001. Reading and the non-academic learner: a mystery solved. *System*, **29(4)**: 451-466

Yano, Y., M.H. Long and S. Ross. 1994. The effects of simplified and elaborated texts on foreign language reading comprehension. *Language Learning*, **44(2)**: 189-219

Yau, M. 1991. The role of language factors in second language writing. In Malave, L. and G. Duquette (eds). 1991. *Language, culture and cognition: a collection of studies in first and second language acquisition.* Clevedon, England. Multilingual Matters: 266-283

Young, R. 1995. Discontinuous interlanguage development and its implications for oral proficiency rating scales. *Applied Language Learning*, **6 (1-2)**: 13-26

Yu, V.W. 2000. *An extensive reading scheme for secondary schools in Hong Kong: its roles in second language development and curriculum renewal.* Unpublished doctoral thesis. University of Wales, Cardiff

Appendices

APPENDIX 1

Day and Bamford's characteristics of extensive reading programmes:

1. *Students read as much as possible*, perhaps in and definitely out of the classroom.

2. *A variety of materials on a wide range of topics is available* so as to encourage reading for different reasons and in different ways.

3. *Students select what they want to read* and have the freedom to stop reading material that fails to interest them.

4. *The purposes of reading are usually related to pleasure, information and general understanding.* These purposes are determined by the nature of the material and the interests of the student.

5. *Reading is its own reward.* There are few or no follow-up exercises after reading.

6. *Reading materials are well within the linguistic competence of the students* in terms of vocabulary and grammar. Dictionaries are rarely used while reading because the constant stopping to look up words makes fluent reading difficult.

7. *Reading is individual and silent,* at the students' own pace, and, outside the class, done when and where the student chooses.

8. *Reading speed is usually faster rather than slower* as students read books and other material they find easily understandable.

9. *Teachers orient students to the goals of the program,* explain the methodology, keep track of what each student reads, and guide students in getting the most out of the program.

10. *The teacher is a role model of a reader for students* — an active member of the classroom reading community, demonstrating what it means to be a reader and the rewards of being a reader.

From Day and Bamford, 1998: 8 (emphases in original)

APPENDIX 2

Sample pages from a graded reader

'Well, I wasn't driving as fast as that. I was driving quite slowly, the road was clear, and I hadn't been drinking. In fact, I *never* drink and drive, *never* – I'm sure the doctors tested that, didn't they, Officer?'

I looked at my notes. 'Yes, I think so. It doesn't say anything here about alcohol, Mr Jackson.'

'So it couldn't be my fault, could it? I mean, I was going at a normal speed, slowly even, on a sunny day, and I hadn't been drinking, and then there he was! Right in front of me!'

I felt confused. 'I'm sorry, Mr Jackson, I don't understand. *Who* was in front of you?'

He stared at me strangely. 'Well, the man, of course. The man who was killed.'

11

From: *The Hitch-hiker* by Tim Vicary. *Storylines* Level 4. Oxford University Press, 1996. (Previously published in the Streamline Graded Readers series.)

APPENDIX 2 (continued)

Sample pages from a graded reader

Now I was really confused. 'What? I'm sorry, Mr Jackson, but no one was killed. The people in the other car were hurt, but they're both alive. One has a broken leg, I think, and the little girl hurt her face, but that's all.'

'Which car?' asked Mr Jackson.

'The car that hit you from behind.'

Now Mr Jackson looked confused. He looked even more unhappy, and his face went white, as white as the bandage round his neck.

'Do you mean . . . are you telling me that a car hit me from behind, Officer?'

'Yes.'

He put his hands to his neck. He had been trying to shake his head but he couldn't.

'I don't remember that,' he said.

'You don't remember anything about a car that hit you from behind?'

'No, nothing. And you say there were two people in it? A man and a girl? Oh, how terrible.' He began to cry, and took a handkerchief from the table beside his bed.

I thought carefully. 'Let's start from the beginning, Mr Jackson. You say you were driving along slowly, about forty-five miles an hour. Is that right? Then what happened? Think carefully, and tell me slowly.'

He put the handkerchief down, and stared at me, his eyes big and wide.

'Well, then I saw him, that's all. A man. He was

From: *The Hitch-hiker* by Tim Vicary. *Storylines* Level 4. Oxford University Press, 1996. (Previously published in the Streamline Graded Readers series.)

APPENDIX 3

EPER pre-reading card

THE HITCH-HIKER	2240 D

BEFORE READING

This is a modern ghost story. It is strange and quite sad, and you will probably go on puzzling over it after you have finished the book.

Sue is a policewoman, and so she is trained to be observant and pick up as much information as possible from people's appearance and speech.

THE HITCH-HIKER	2240 D

Author:

Time and Place of the Story:

The Characters:

a) .... was a police officer.

b) Her boyfriend .... was a reporter.

c) The hitch-hiker's name was ....

The Story:

1) What happened when Mr Jackson stopped his car suddenly?

2) Why did Mr Jackson stop his car so suddenly?

3) What was David Holland trying to do when he was killed in July 1970?

4) What did Simon notice about David Holland's grave?

5) Where did Sue see the hitch-hiker again, and with whom?

6) Why was Sue surprised by the way the hitch-hiker got out of her car?

Comment:

a) Did you enjoy the book? Give two reasons for your answer.

b) Where do you think David Holland went when he got out of Sue's car at the end of the story?

THE HITCH-HIKER 2240 D

ANSWERS

Author: Tim Vicary

Time and Place of the Story: the present, in England

The Characters:

a) ***Sue*** was a police officer.

b) Her boyfriend ***Simon*** was a reporter.

c) The hitch-hiker's name was ***David Holland*** .

The Story:

1) When Mr Jackson stopped his car suddenly, ***another car went into the back of it*** . (p5)

2) Mr Jackson stopped his car so suddenly ***because he saw a man running across the road in front of him*** . (p12)

3) David Holland was trying ***to save his son*** when he was killed in July 1970. (p19)

4) Simon noticed that David Holland's grave ***always had fresh flowers on it*** . (p20)

5) Sue saw the hitch-hiker again ***at the University with his son*** . (p24)

6) Sue was surprised by the way the hitch-hiker got out of her car ***because he went through the closed door*** . (p27)

APPENDIX 4

Overall quality rating instrument and instructions to raters for rating procedure:

Holistic Scoring Instrument

6 High/Excellent
5 Good
4 High Average
3 Low Average
2 Weak
1 Low/Very Weak

The holistic mark is an impressionistic mark. Compositions should be assessed *rapidly*, and in relation to the others in the group.

Procedure

1. From the first 40 compositions, individually try to choose the strongest and the weakest. Discuss within the group. When agreement is reached as to which two compositions are the strongest and the weakest, these will represent the end-points of the range of scores.

2. Individually make judgements about the first 10 compositions in relation to the end-point compositions and to each other. Allocate these within the range of six grade scores. Discuss as a group. Agreement should be reached, resulting in sample scripts for each grade.

3. Repeat this process with batches of 10 compositions. When a high level of inter-rater consistency is achieved, each rater will individually rate 40 as yet un-rated compositions.

4. After a break, each rater will rate the same 40 compositions again to check intra-rater reliability.

5. This process may be repeated until a high degree of intra-rater reliability is achieved.

6. Raters may now proceed to rate all the compositions, in random order. (Compositions will need to be re-shuffled.)

7. Intra-rater reliability checks will be carried out from time to time throughout, by re-inserting a random selection of five or six compositions back into the pile of un-rated scripts, to be rated twice by the same person.

APPENDIX 5

Benchmark scripts

The benchmark compositions below were assigned to their given level by all three raters. They originate from the complete data set of 392 compositions.

Level 1

Composition 336

One day, I go to Ocean Park. Sea one old man come with me. He wearing a jacket a long and big. He knocked my houses in the evening. I feering surprising, he asked give me three wishes come true for you on one special condition. And he said him is magician. I want something he can come true.

I doesn't beling him but I want to try. One true, I hope I can have a big house and with my family live in there. Two true, all people is always hearthy. Three true, many money.

I had just have 3 wishes, so I will help the old man to clean the house. Now, I had to do my work.

I am very happy with 3 wishes. I will thank you to the old man.

Cog 3C Control

APPENDIX 5 (continued)

Benchmark scripts

Level 2

Composition 311

At the last holiday I and my classmate went to camping. We went to Lantu Island. The first we went to the beach after our dinner, because we though that go out at night is very excited.

We play, for a long time, we all go tided, we sat on the beach to have a talk. All of my friend wanted to went back to the house for sleep, except I and my very good friend Amy. We talk until twelve o'clock, we wanted to back home again but when we want to leave, we found that I and Amy did not know the way back to the house.

We fell cold and tired, we could only found one house, the ghost's house.

We went in the ghosts, I was very afraid but Amy was brave, she told me not to afraid of the ghosts, because god is always with us. When I heard god, I did not afraid any more, because I belive him. I could not went sleep, and Amy too. We only sat on the floor to talk, I heard some strange sound. I and Amy were afraid, the sound became louder and louder. we decided to found sound is it. That sound seem like ghost but we could only found some cat, dog And a lot of another animals. The sun rises we could found the way back to the house, we call all of our friends to went to that house to see the cute animals.

Cog 3C exp

APPENDIX 5 (continued)

Benchmark scripts

Level 3

Composition 110

Yesterday, my friend whose name was Tom visited me from England. He told me about his camp in last year.

He said, "About September in the last year. I had camp with my friend, John, Jack, Mary and Ann. We want to pass a great forest, but we got lost. Mary and Ann was very frightened. They cried, "We can do what now, the sky will be dark quickly. we want to go home." John said to them, "Don't be frightened! Now, we must find a place to sleep. I hope we can find it before the sky has got dark." Luckily, they found a big house when the sky had got dark. The boys were so happy, but the girls were still frightened because the house was horror.

John was the happiest one, so he went into the house first. Then, we followed him go to the house. Inside the house, we felt the house which was more bigger, but it was more horror. In the house, there was not any people and any sound except we made. Ann said, "I thing it is a ghost house because I saw many film about ghost house like this." John said, "Don't worry, there is not ghost in the world." Although John had said that, we still afraid. We still stayed in the house because in the outside would be more teror.

In the mid-night we heard some sound, but we didn't want to check it except John. He went next to the door and looked out the outside. There was nothing near the house. Then He turned back and said to us, "It's only wind, don't worry!"

Next day morning, we were alright. John said, "Now, are you believe me." We were thankful him.

Cog 3H exp

APPENDIX 5 (continued)

Benchmark scripts

Level 4

Composition 205

Last week, my friends and I went camping on a mountain. One day, we wanted some water for putting off the fire, which has just used for cooking a meal. So Sue and I went to find water together.

We went down the mountain and found water. At first, we couldn't find any river. So that we walked for a long distance and we both didn't mentioned it until we has found the water. When we turned back. We were so afraid because there were only trees around us. We tried to find the way back to our friends, but we failed. We walked and walked, and the result had no change that we couldn't find the way. Suddenly, a 'Ko' sound came behind me that Sue fell down. She was hurt and couldn't walk far, so I tried to some places where we could stay. Luckily, there was an old house behind a big tree. We walked to the house slowly and entered into it when there was no response. Unexpected, that was so many strange things happened.

When we entered the house, it was so beautiful that we couldn't believe it was an old house. At first, I didn't manage the beautiful things because I had to take care of Sue. I tried to find the medical box for Sue. After this, Sue was in asleep. so I tried to look everywhere. After looking the whole house, I tried to find if there was any telephone inside the house. At last, I found a telephone was in the room. So I tried to call up my parents. But when I dialled the telephone number, the light suddenly closed. But the fan was still working. So that I tried to find the switch of the light. When I found it. I felt there was someone holding my hand. And I couldn't move. At last, I could move and I wanted to find Sue saying about this. But when I went back to Sue, she was playing with the air happily. So that I tried to slap her and she waked up. I told her what has happened and we decided to leave. Unluckily the door was shut and we was so afraid that closed ourselves in the room the whole night.

When we waked up, our parents were next to us and told us what had happened. And we found that the house was the people who said that was full of ghosts.

BLMC Control

APPENDIX 5 (continued)

Benchmark scripts

Level 5

Composition 24

There are many stories and movies about going back to the past and future. The stories are very attractive and I usually read these books until midnight.

Today, I had borrowed a book called 'How to go back to future?' from the library. After doing my homework, I read this read and put my mind into the story....

I was awaken by a beam of light. I opened my eye and looked around. Around me was many animals including rabbits, deers and . They looked at me and seemed to be very afraid. But in fact, I was much more afraid than them. Where am I? I only found that I was in a cave, a large cave with a hole on the top, allowing a beam of light to come in. I stood up and went out of the cave. Outside the cave is a big forest. There was a lot of tall trees and shrubs. There was no people living nearby. The only thing I could do is running, However, the forest was too large and just like a puzzle. I continued to run and rest until the sun set in the west. I couldn't find the exits. The condition was much worse than before. There was some snakes on the trees and there was no cave and water. I found that I was lost. I walked slowly without destination. I knew if I couldn't find any water or food, I would die.

Suddenly, I found that there was a water pool at the front. I ran to it without thinking if there was dangerous. I drank the water and washed my face. At this moment, an animal like dinosaur come out from the pool. It looked at me and thought that I was its dinner. I quickly ran away but my stomach was painful. The water I'd drunk was poisonous. I took a wood as a weapon. My tears came out and I shouted....

"Wake up, girl! It's time to go to school!" I wake up. I found that my face is wet and my hand was holding a wood. This was not a dream, was it? If I did not wake up, would I die in the dream? I didn't know. I always remember this. This was not a dream, it was a part of my life. It was really terrible but also very exciting.

Bel Control

APPENDIX 5 (continued)

Benchmark scripts

Level 6

Composition 373

Last year, during the summar holiday, I went shopping with my parents. After we had bought a few things, my father said that he had to go to a Hongkong Bank nearby to did some bussiness-related affairs. Suddenly, I felt dizzy. Because my parents were friends of the branch manager and my parents still had some shopping to do, the branch manager allowed my parents left me in a resting room. My parents said that they would pick me up after one and a half hour.

I slept peacefully without any disturbance. Suddenly, I was woken up by a very loud sound followed by a series of screaming. I got up and went to find out what's happening. When I was in the corridor, I saw a men pointing a gun at the manager, with his back facing me. All the other clerks and customers crouched down and put their hands on their head. Then I realized that the man holding a gun was a robber. I looked around to see whether the robber had any partner, fortunately, he was alone. I heard that the robber was forcing the manager to tell the code of the safe. So, I quickly crept back to the manager room and found a racket. I also pressed the burglar alarm switch. I went out to the corridor again with the racket in my hand. The robber heard the burglar alarm and was very angry. While the robber was searching for the person who pressed the alarm switch, I went behind the robber and hit him with the racket with all my strength. The robber fell to the floor lying unconsciously. The clerks tighted the robber up with a rope. I became a hero. After a while, many policemen came, they were surprised that the robber was caught by a young girl. At this time, my parents came back and they accompanied me to the police station.

How exciting that day was!

APPENDIX 6

Constructs evaluation instrument and guidelines for raters:

Rating profile instrument

ID _____	1	2	3	4	5	6
Grammatical complexity						
Grammatical accuracy						
Vocabulary range						
Spelling						
Punctuation & paragraphing						
Story reads well						

6 = Very High/Excellent
5 = Good
4 = High Average
3 = Low Average
2 = Weak
1 = Very Weak

You should write the identification number of the composition in the **ID** box. Put a cross in the grid under one of the grade points 1 to 6 for each category (*i.e.* "grammatical complexity" *etc.*). You need not decide the marks in the same order as given in the grid. The mark for each category is an impressionistic mark. Compositions should be assessed *rapidly*, and in relation to the others in the group.

APPENDIX 6 (continued)

Notes:

Grammatical complexity: high scoring compositions will demonstrate a wider range of grammatical structures (*e.g.* passives, greater range of verb tenses, conditionals, gerunds, two-verb structures *etc.*) than low scoring compositions. High scoring compositions will exhibit more frequent use of complex, multi-clause sentences. The lowest scoring compositions will consist almost entirely of simple one-clause sentences.

Grammatical accuracy: high scoring compositions will have relatively few grammatical inaccuracies. The lowest scoring compositions will have many major grammatical inaccuracies (proportional to the length of the composition). Disregard spelling mistakes.

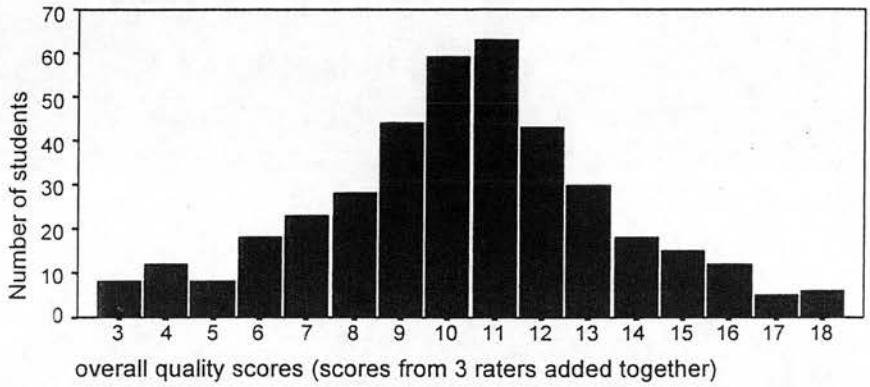
Vocabulary range: high scoring compositions will exhibit a better range of vocabulary, with credit given for more sophisticated or less commonly encountered words. Low scoring compositions will use common, low-level vocabulary and may use the same words repeatedly.

Story reads well: this is less to do with actual content, or events within the story, than with **the way the story is carried along by the language**. Is it easy to follow? Do the sentences "flow"? Can the reader engage easily with the sequence of events, or is reading fluency interrupted at times by a lack of textual coherence?

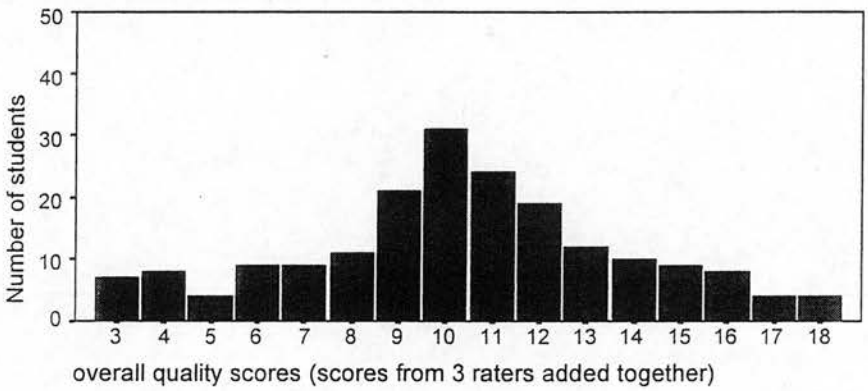
APPENDIX 7

Overall quality scores distributions: four schools

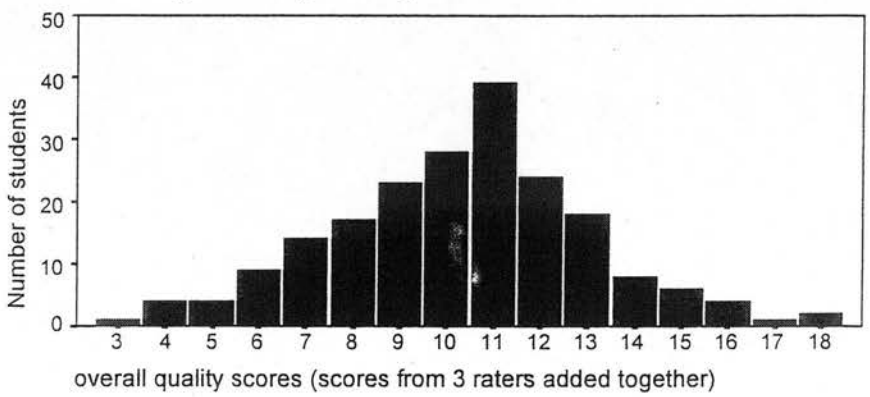
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

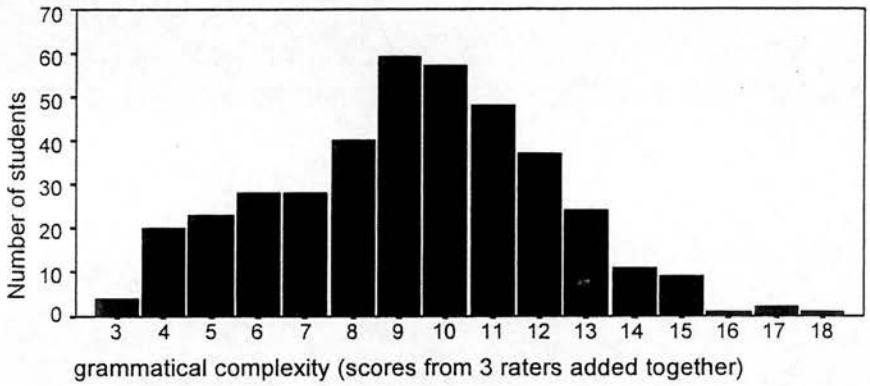


APPENDIX 8

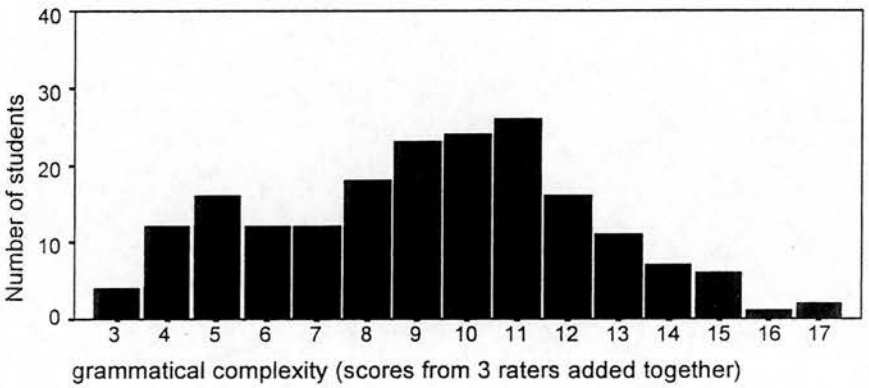
Distributions for scores on rater-judged constructs: four schools

GRAMMATICAL COMPLEXITY

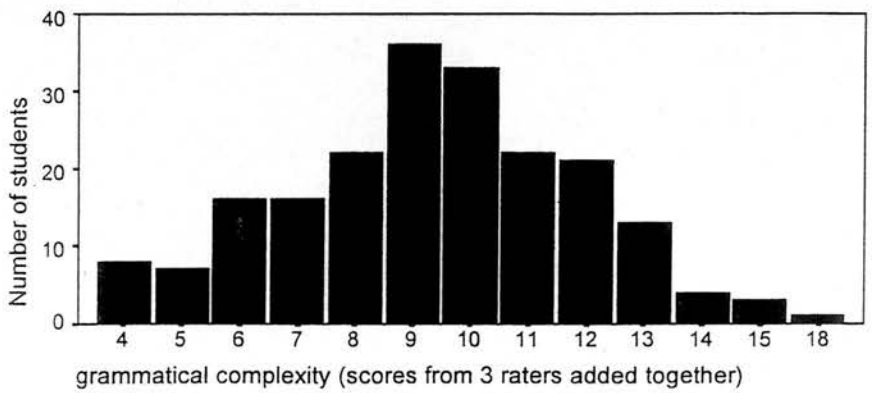
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

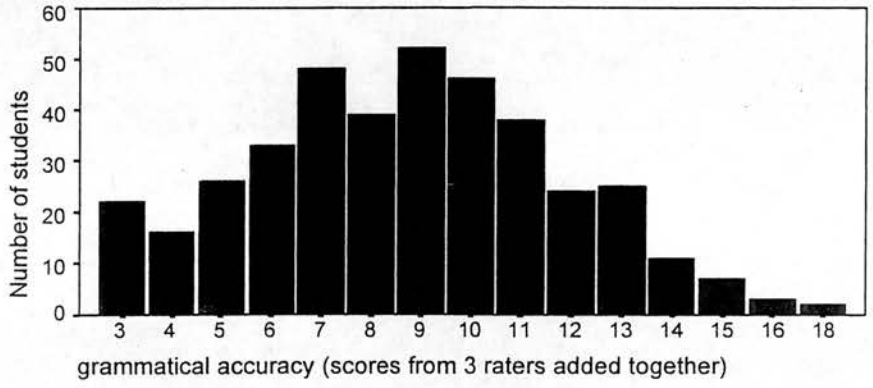


APPENDIX 8 (continued)

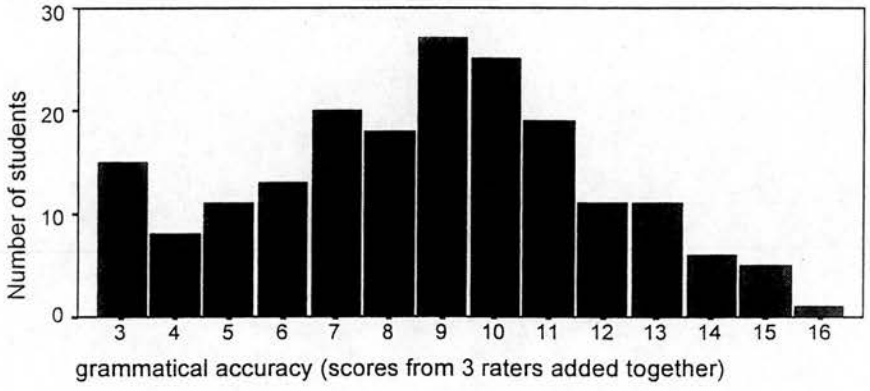
Distributions for scores on rater-judged constructs: four schools

GRAMMATICAL ACCURACY

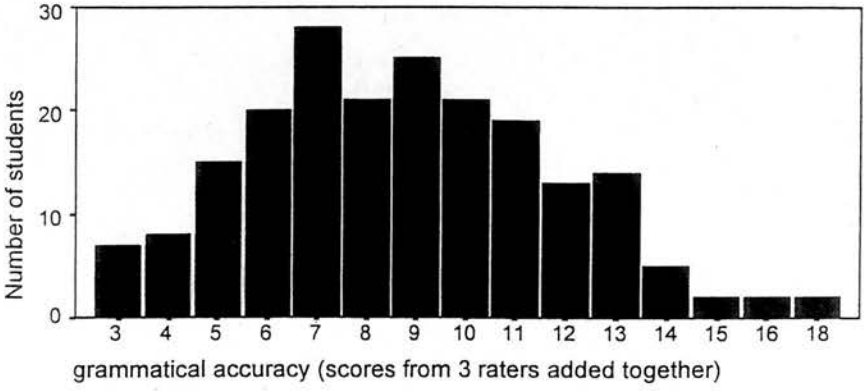
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

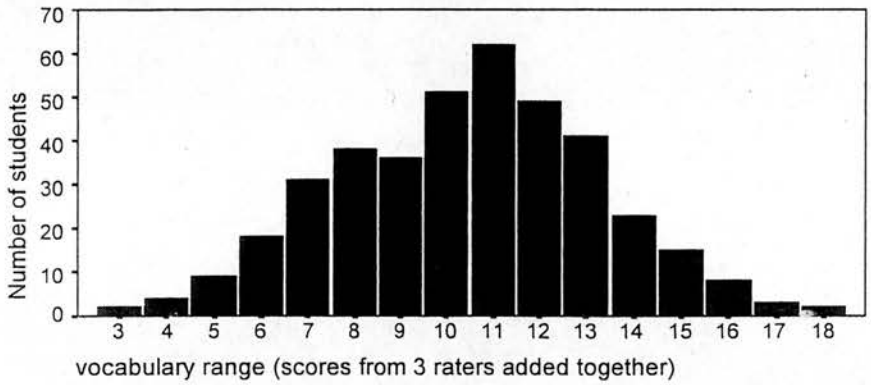


APPENDIX 8 (continued)

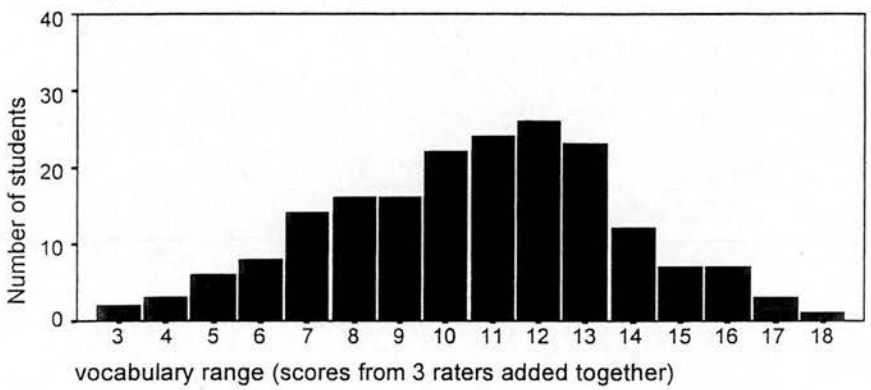
Distributions for scores on rater-judged constructs: four schools

VOCABULARY RANGE

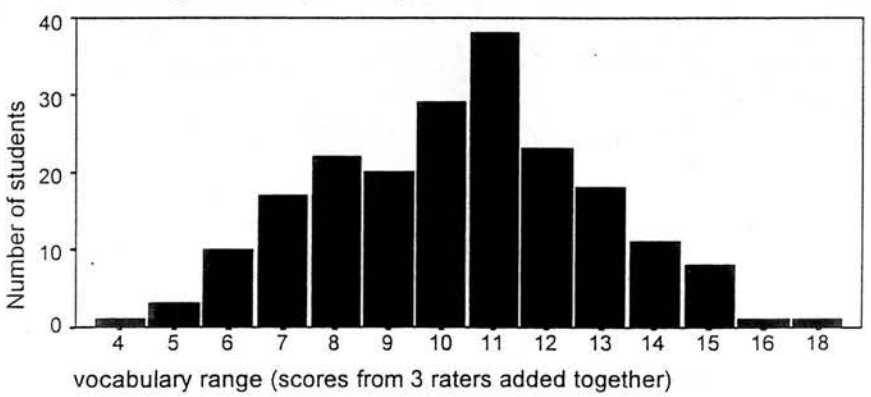
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

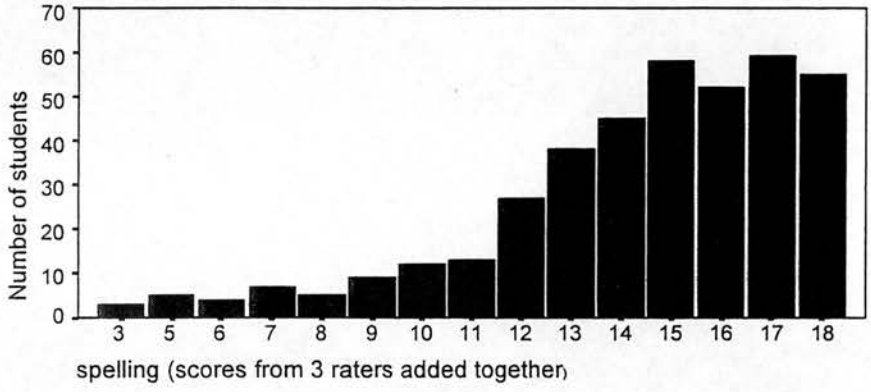


APPENDIX 8 (continued)

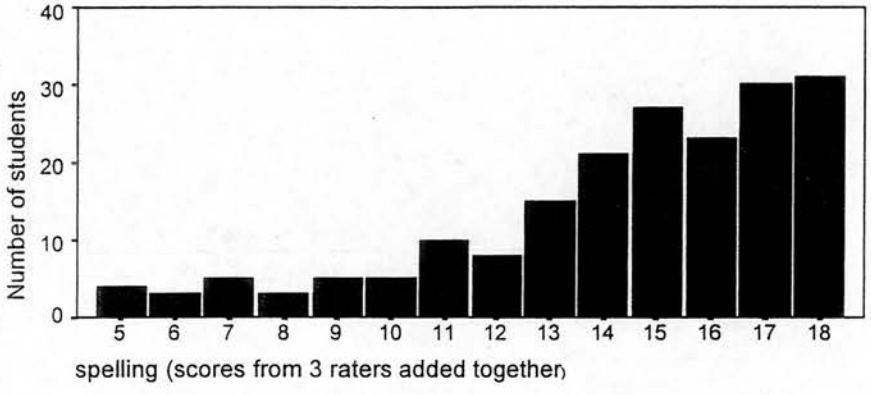
Distributions for scores on rater-judged constructs: four schools

SPELLING SCORES

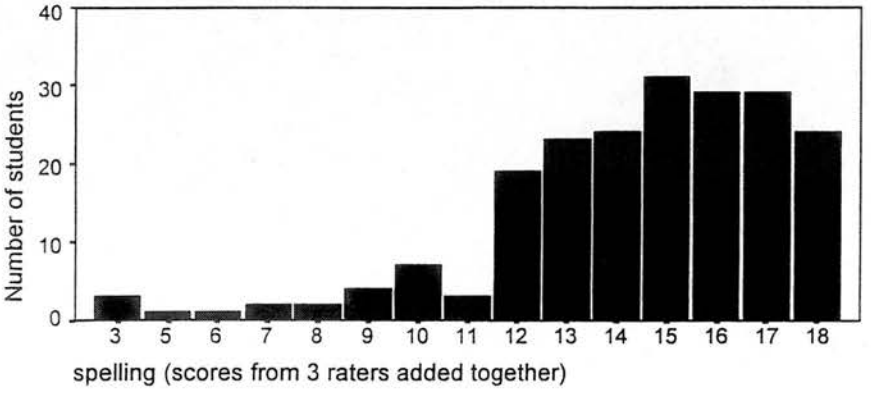
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

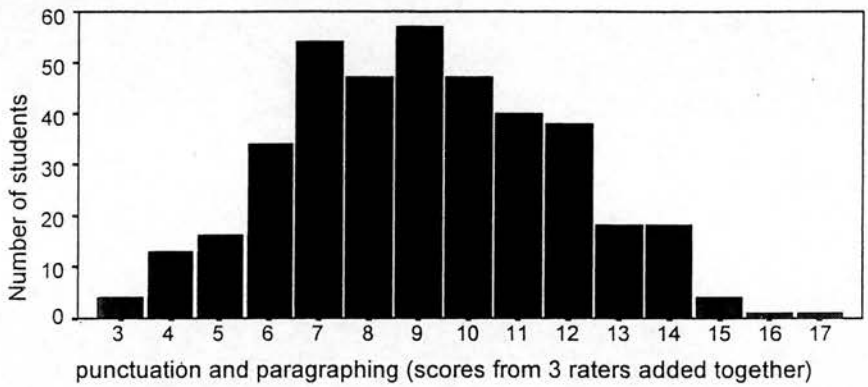


APPENDIX 8 (continued)

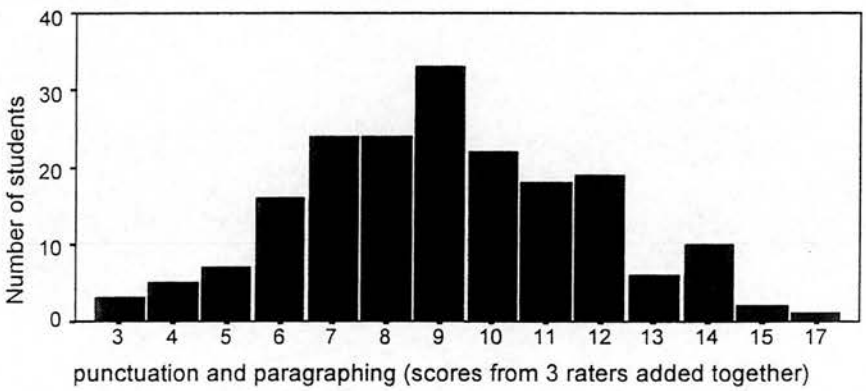
Distributions for scores on rater-judged constructs: four schools

PUNCTUATION AND PARAGRAPHING

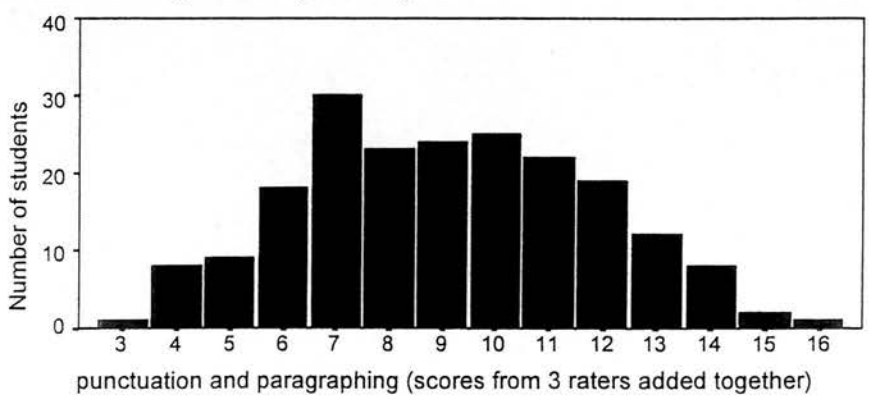
Complete data set (N = 392)



Control compositions (N = 190)



Experimental compositions (N = 202)

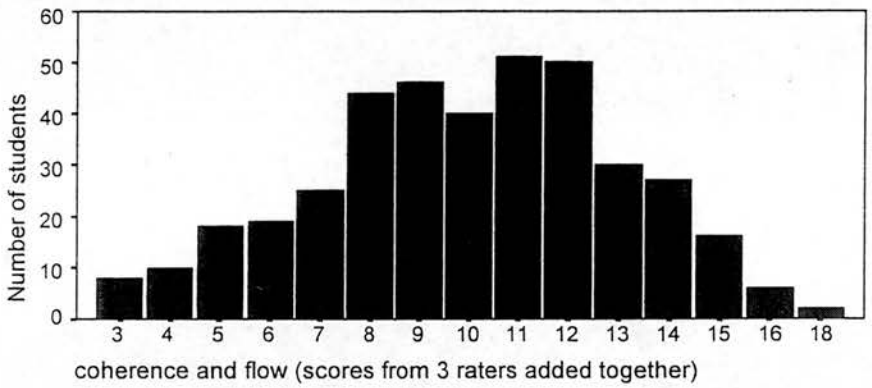


APPENDIX 8 (continued)

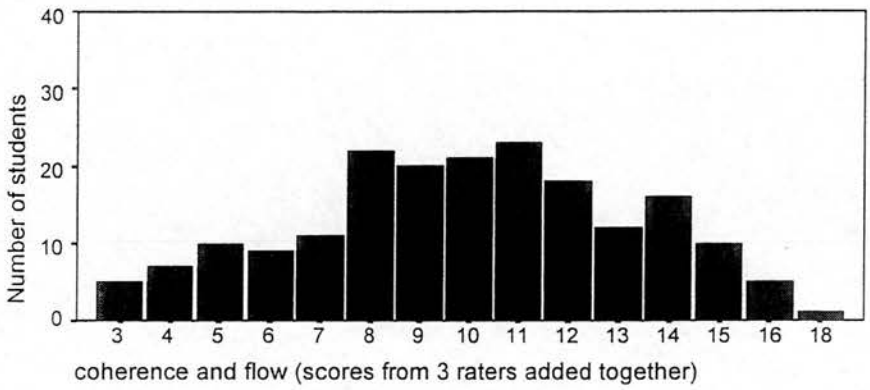
Distributions for scores on rater-judged constructs: four schools

COHERENCE AND FLOW

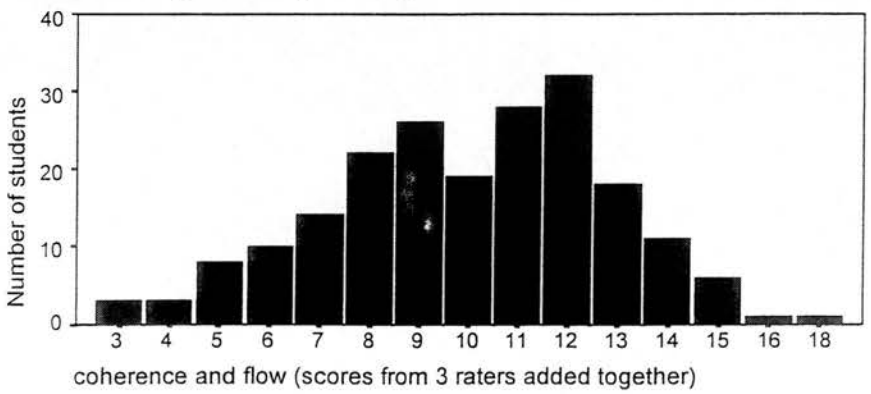
Complete data set (N = 392)



Control compositions (N = 190)



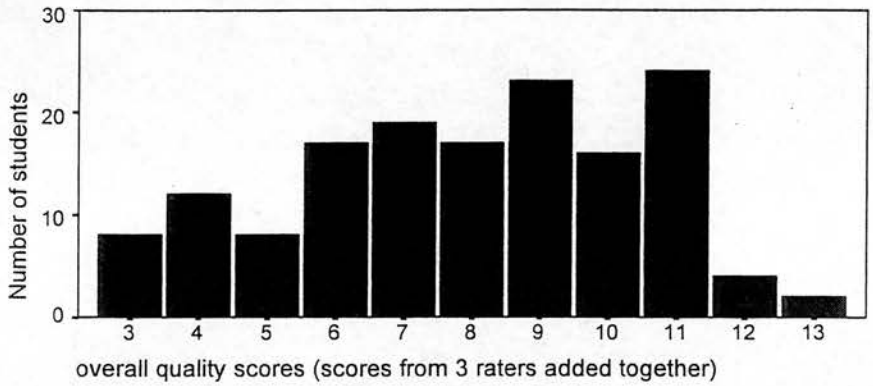
Experimental compositions (N = 202)



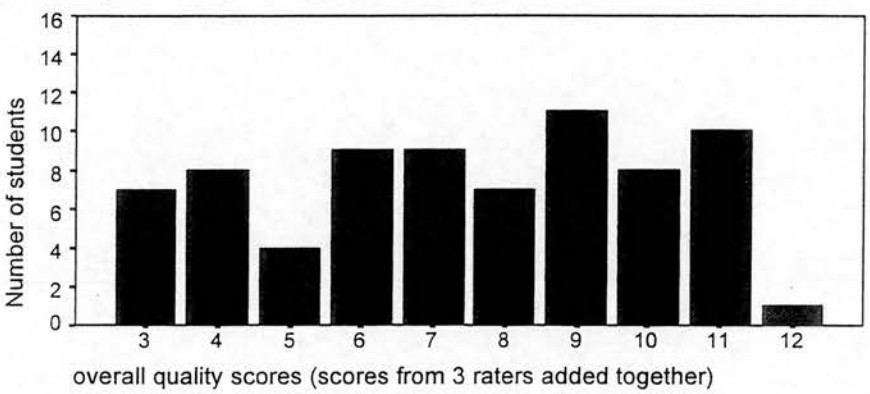
APPENDIX 9

Overall quality scores distributions: school 4

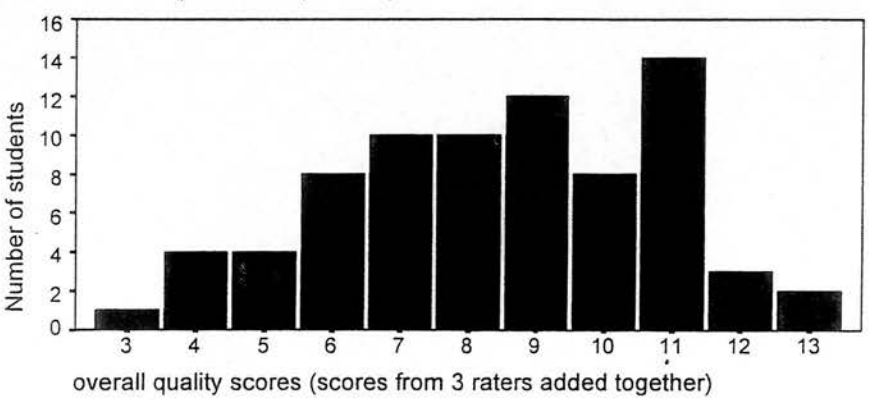
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

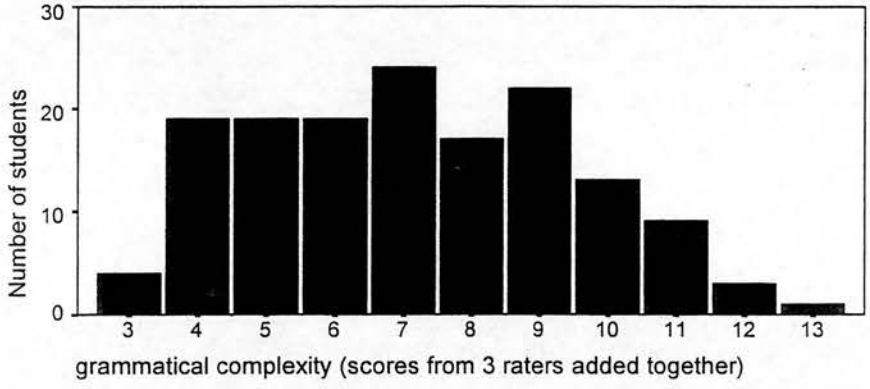


APPENDIX 10

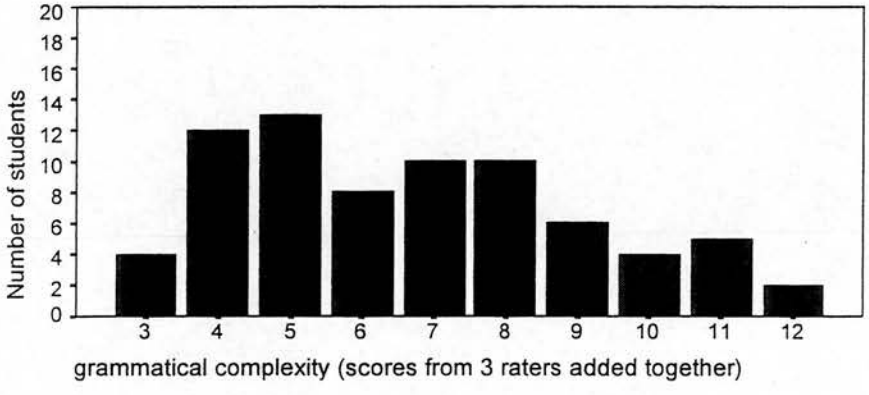
Distributions for scores on rater-judged constructs: school 4

GRAMMATICAL COMPLEXITY

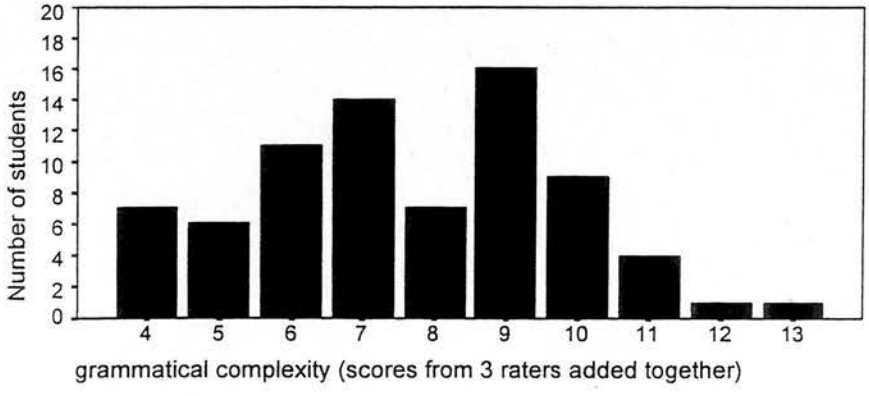
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

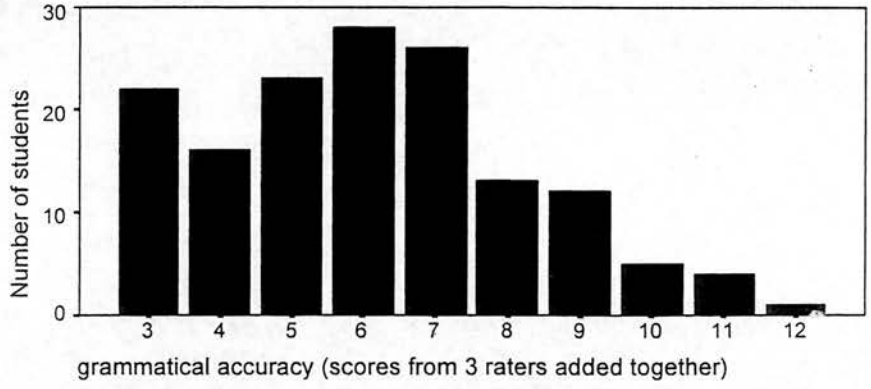


APPENDIX 10 (continued)

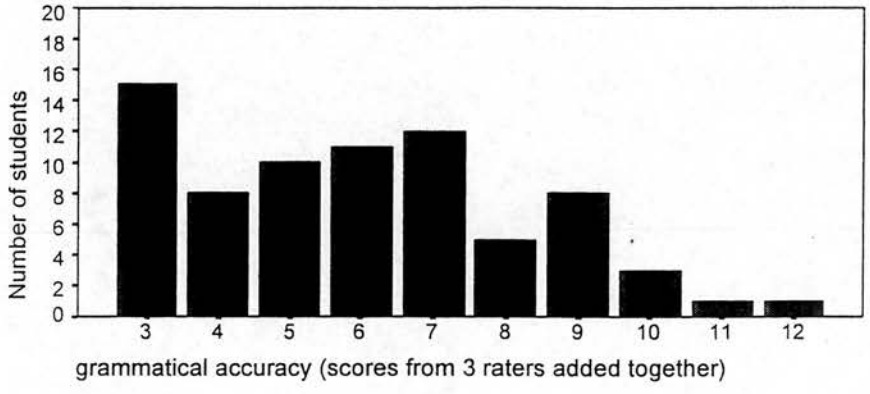
Distributions for scores on rater-judged constructs: school 4

GRAMMATICAL ACCURACY

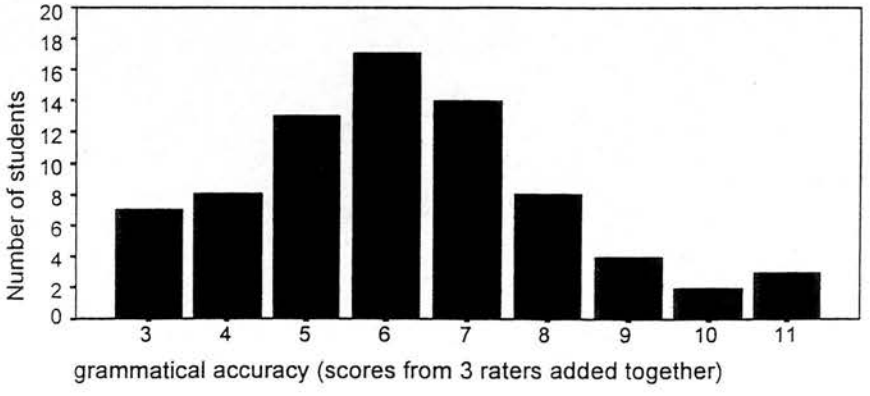
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

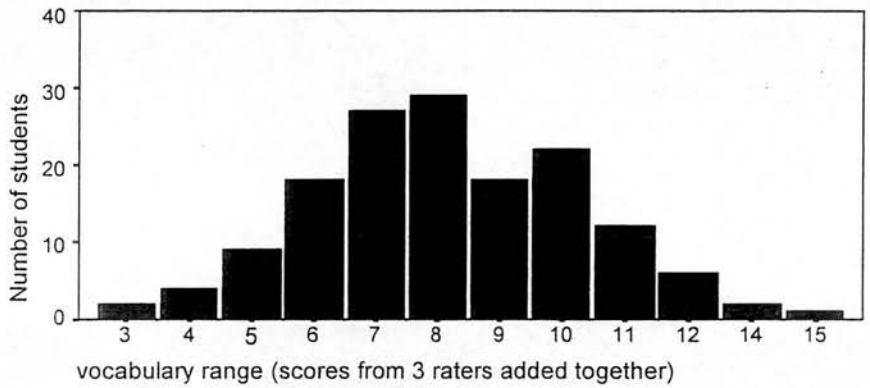


APPENDIX 10 (continued)

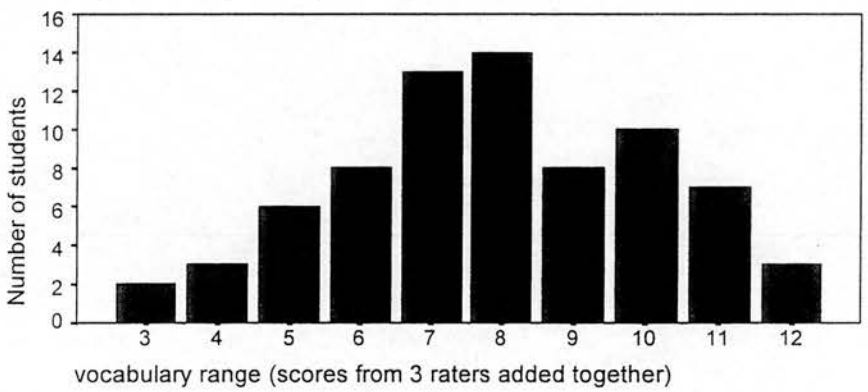
Distributions for scores on rater-judged constructs: school 4

VOCABULARY RANGE

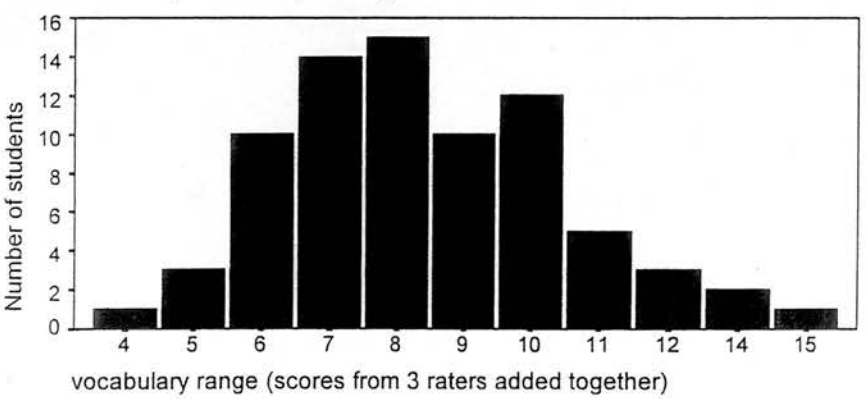
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

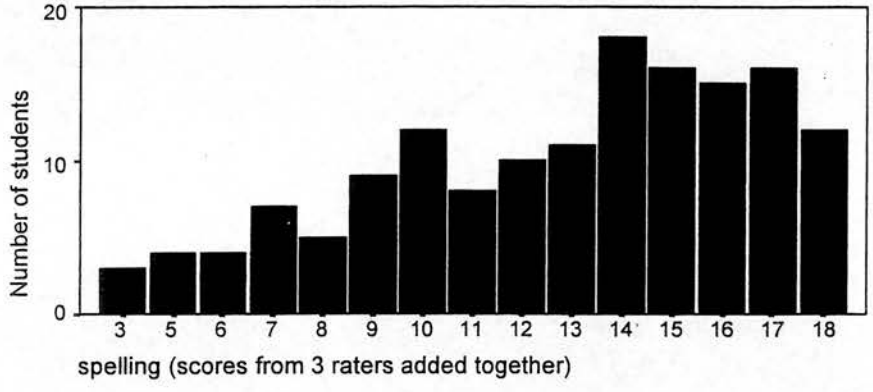


APPENDIX 10 (continued)

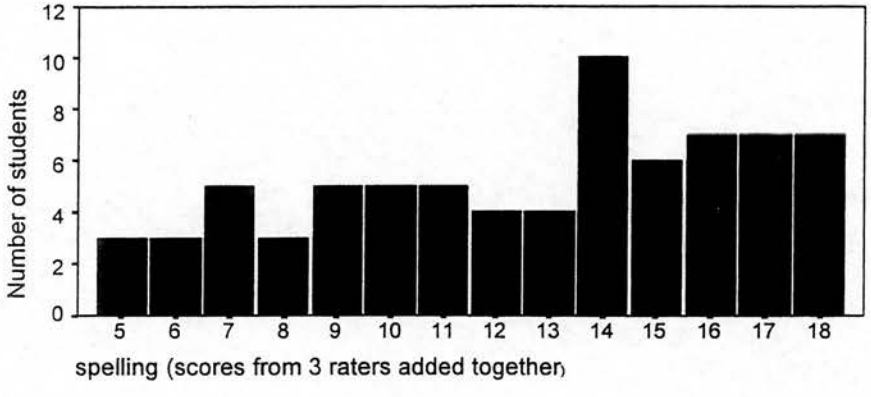
Distributions for scores on rater-judged constructs: school 4

SPELLING SCORES

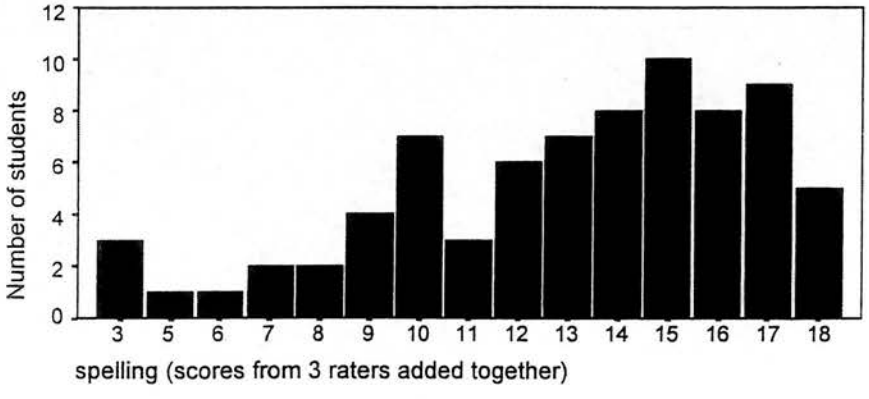
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

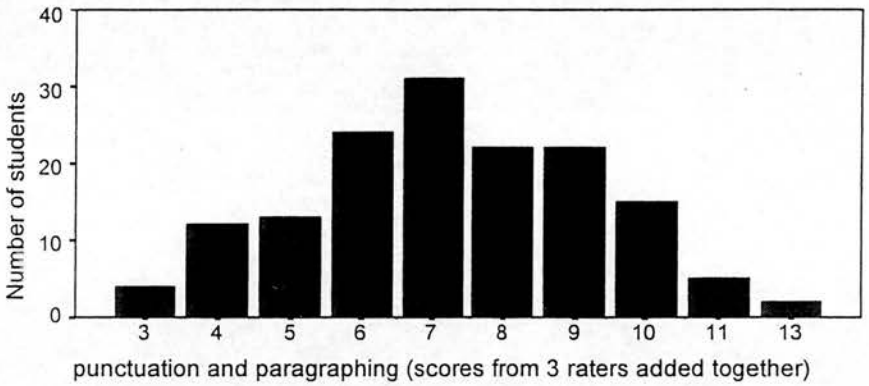


APPENDIX 10 (continued)

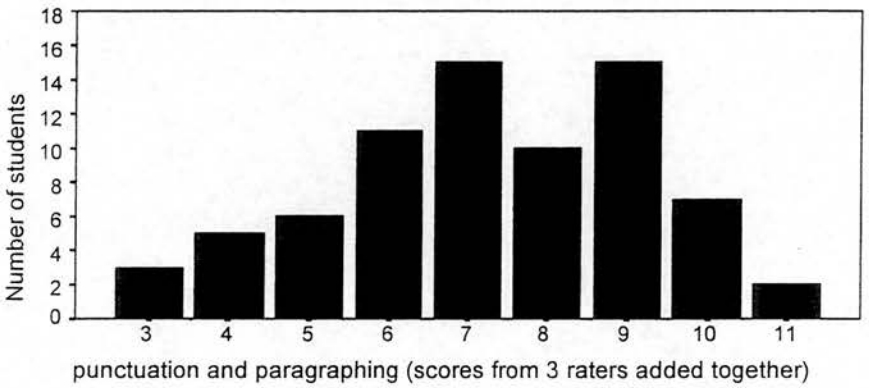
Distributions for scores on rater-judged constructs: school 4

PUNCTUATION AND PARAGRAPHING

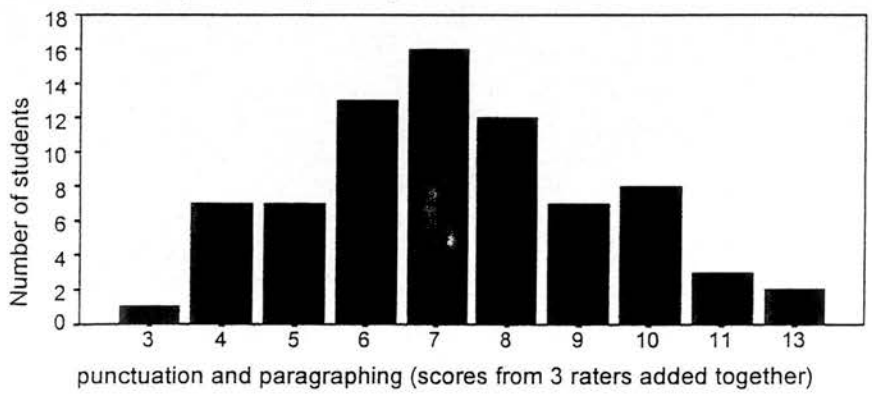
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

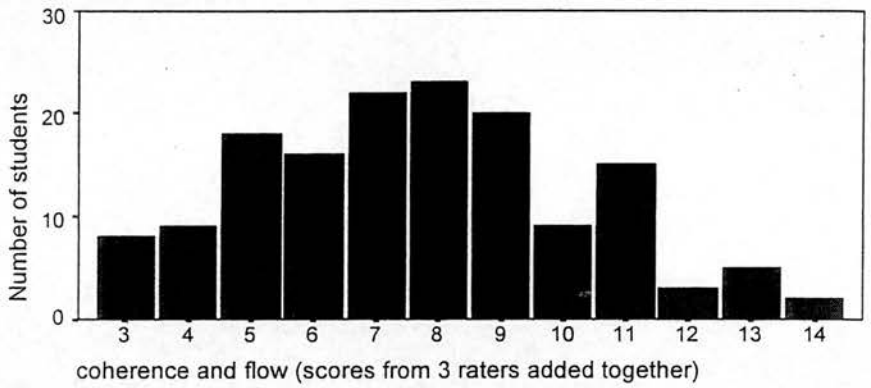


APPENDIX 10 (continued)

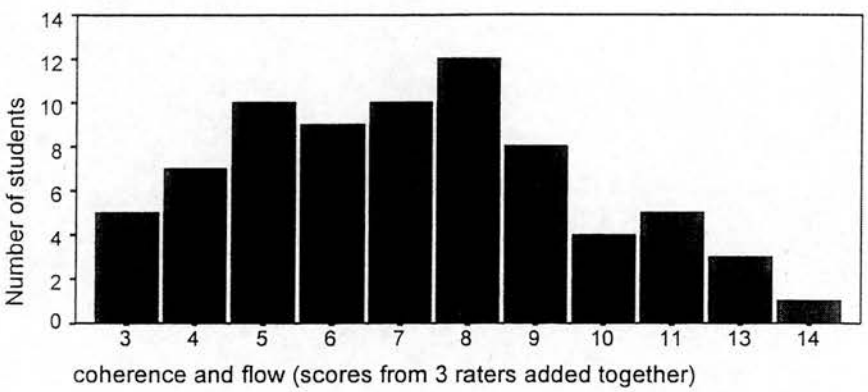
Distributions for scores on rater-judged constructs: school 4

COHERENCE AND FLOW

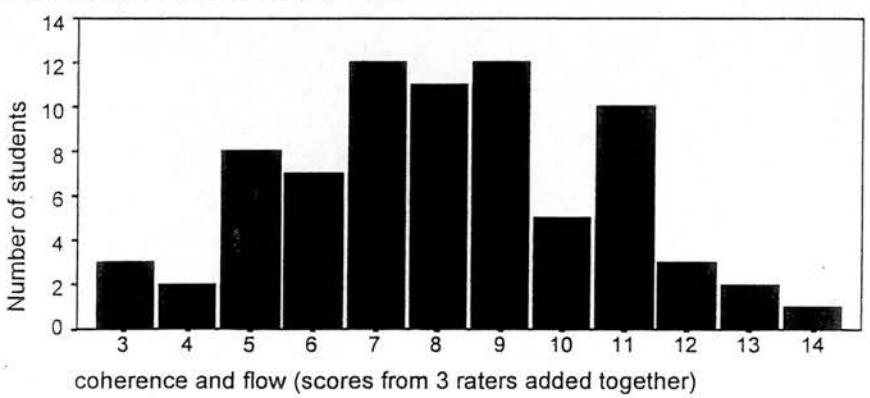
Complete data set (N = 150)



Control compositions (N = 74)



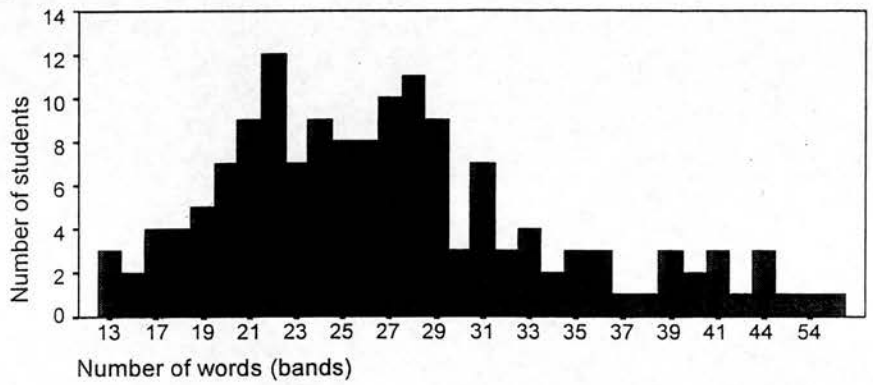
Experimental compositions (N = 76)



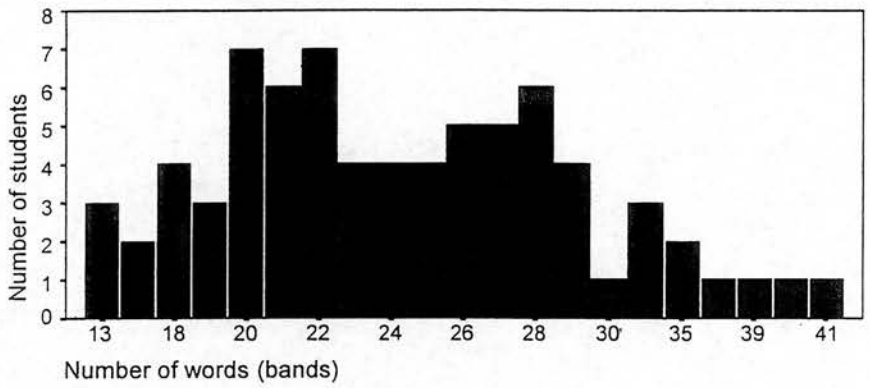
APPENDIX 11

Distributions for number of words per composition*: school 4

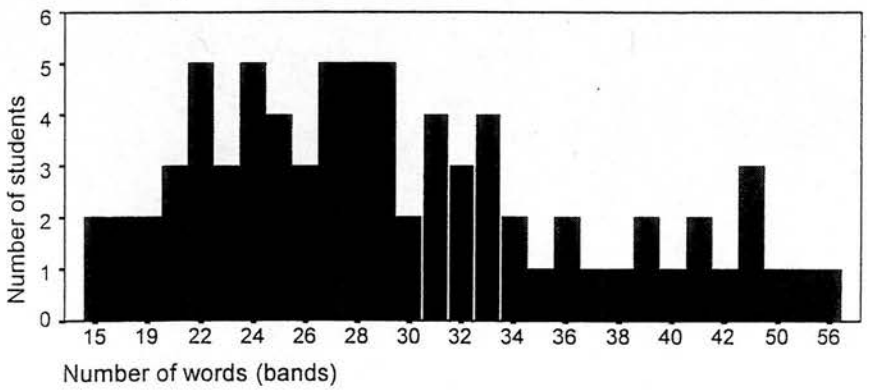
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

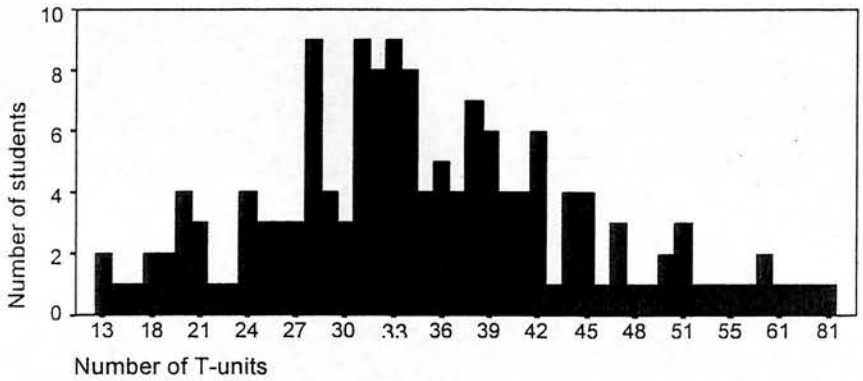


*Numbers of words are in bands of 10; *e.g.* 13 includes compositions with total numbers of words ranging from 130 to 139

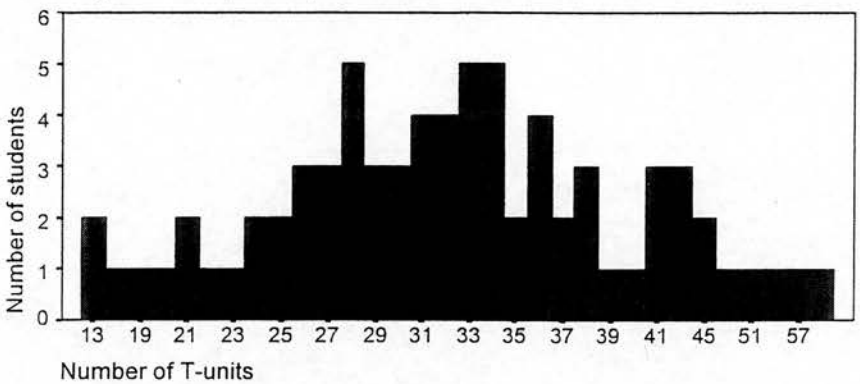
APPENDIX 12

Distributions for number of T-units per composition: school 4

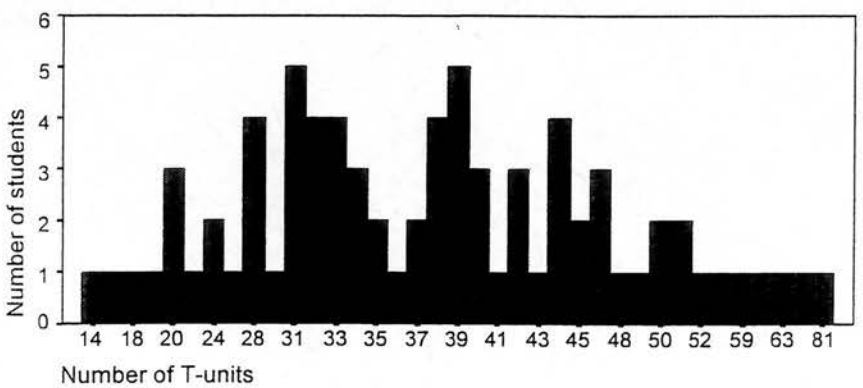
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

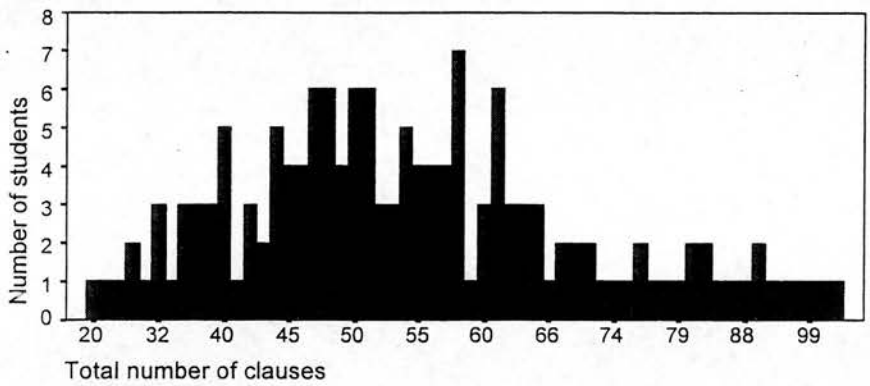


APPENDIX 13

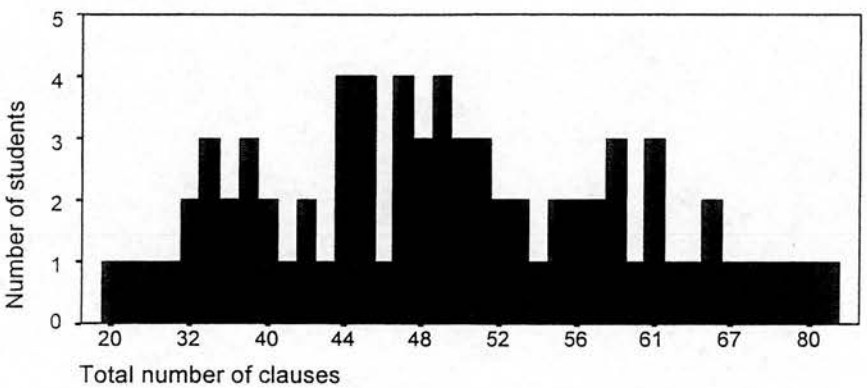
Distributions for numbers of clauses per composition: school 4

TOTAL NUMBER OF CLAUSES

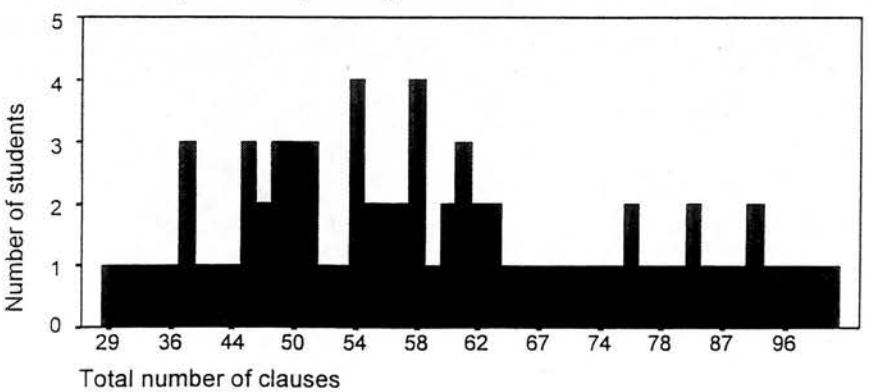
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

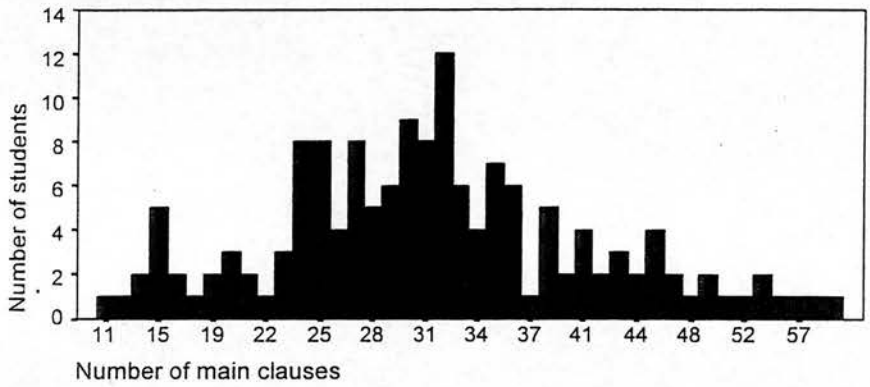


APPENDIX 13 (continued)

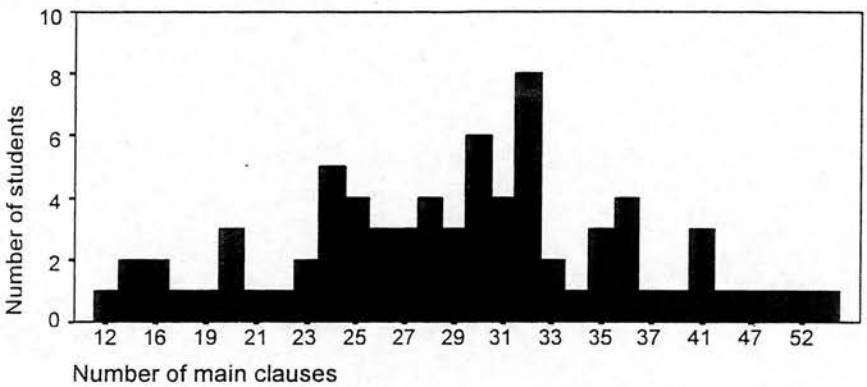
Distributions for numbers of clauses per composition: school 4

NUMBER OF MAIN CLAUSES

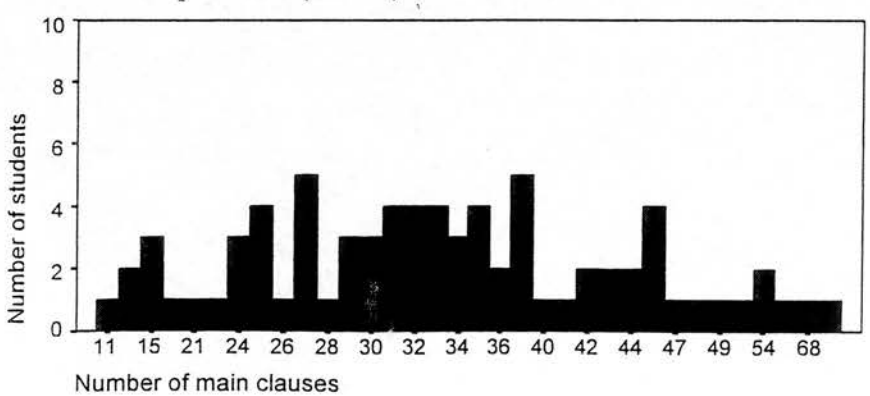
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

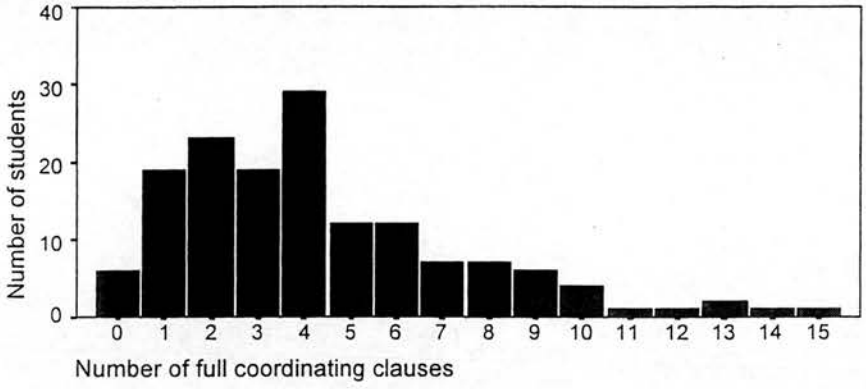


APPENDIX 13 (continued)

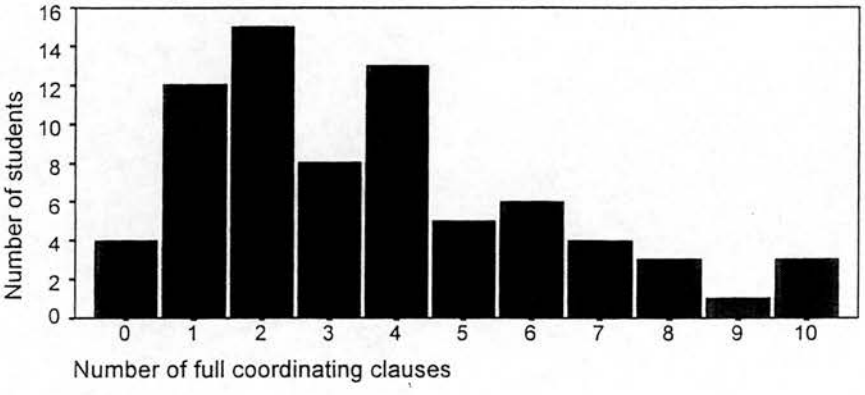
Distributions for numbers of clauses per composition: school 4

NUMBER OF FULL COORDINATING CLAUSES

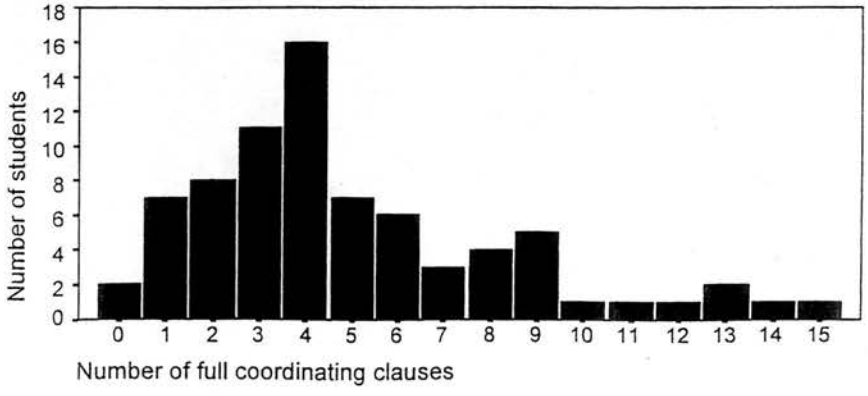
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

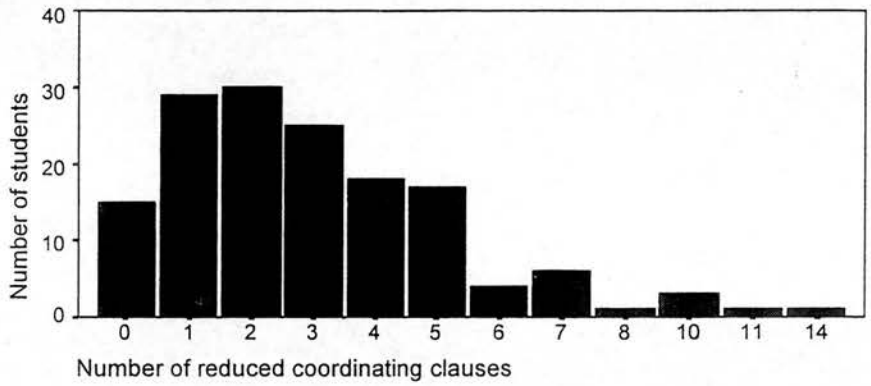


APPENDIX 13 (continued)

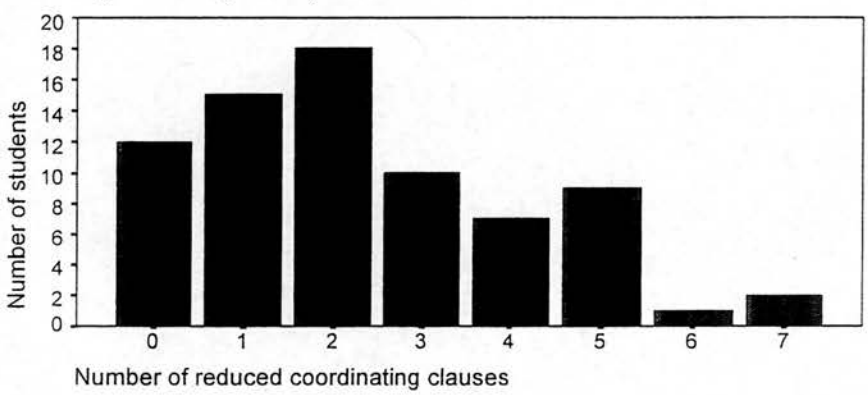
Distributions for numbers of clauses per composition: school 4

NUMBER OF REDUCED COORDINATING CLAUSES

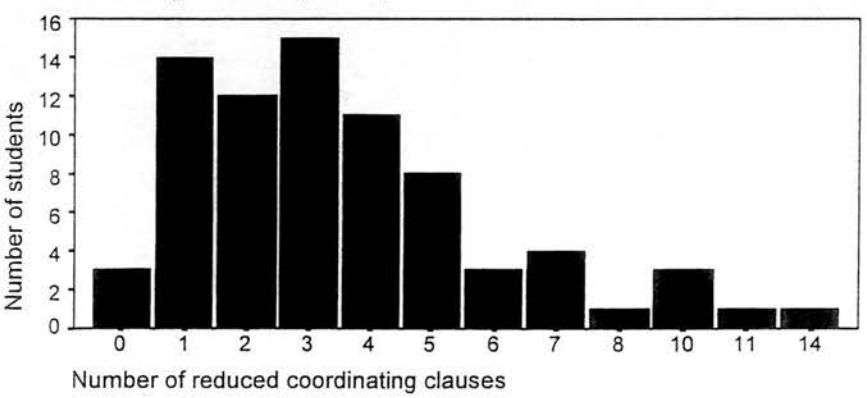
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

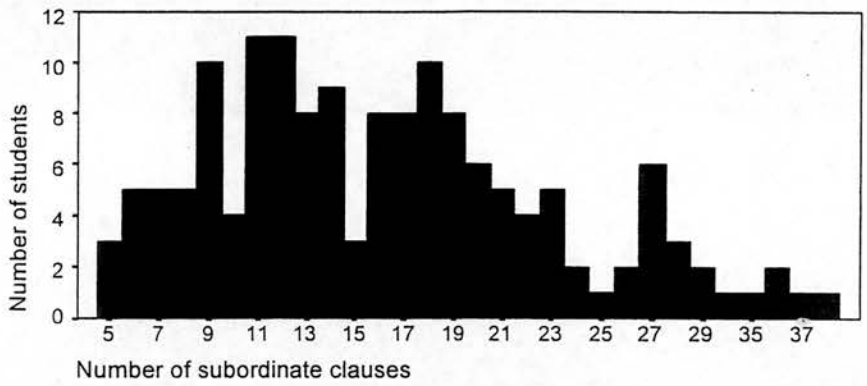


APPENDIX 13 (continued)

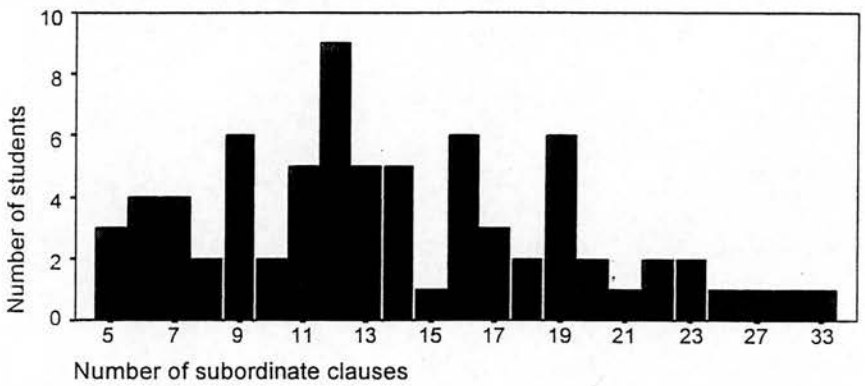
Distributions for numbers of clauses per composition: school 4

NUMBER OF SUBORDINATE CLAUSES

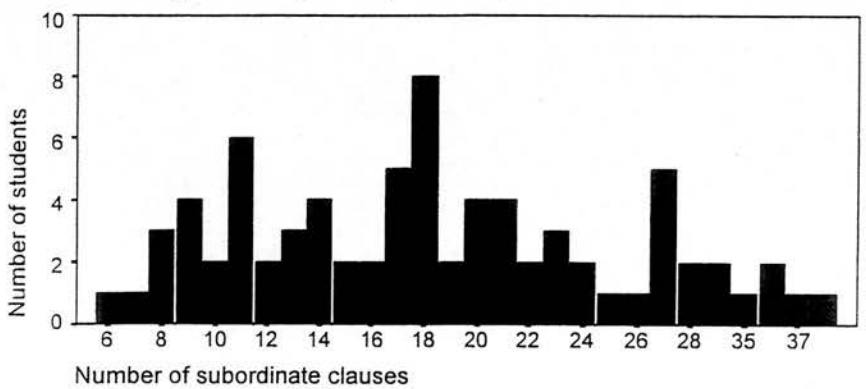
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

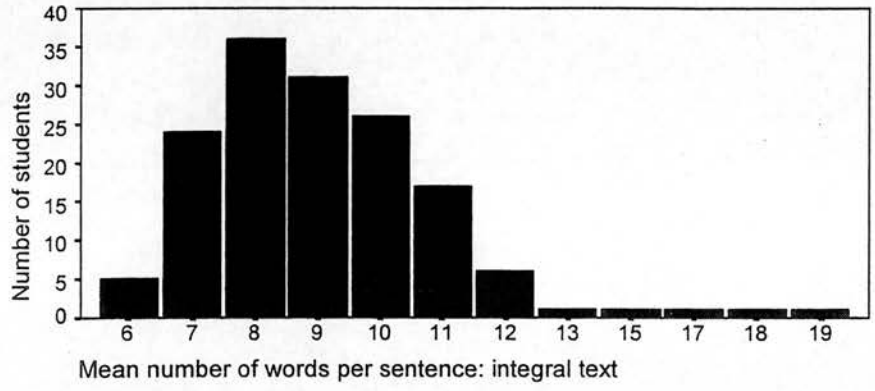


APPENDIX 14

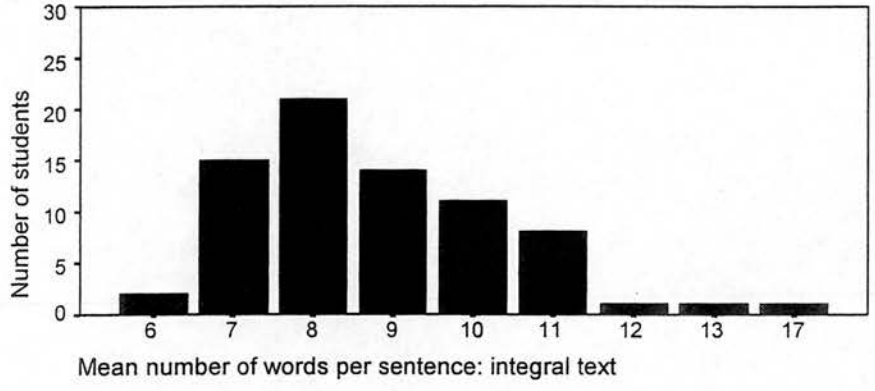
Distributions for mean numbers of words per sentence*: school 4

INTEGRAL TEXT

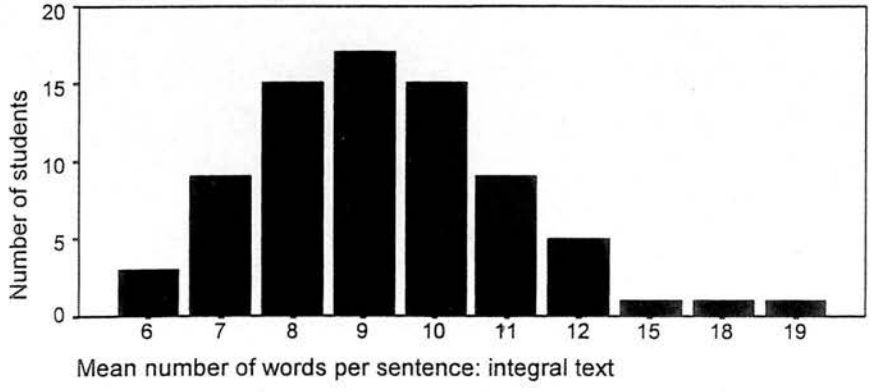
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)



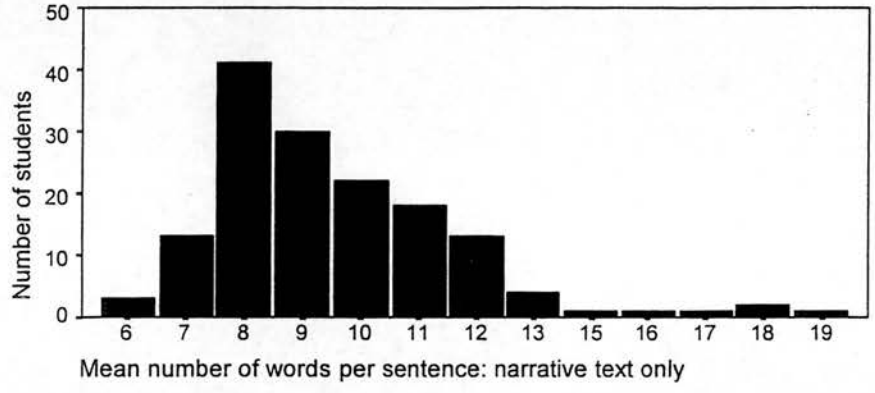
* Mean number of words per sentence is given to the nearest whole word

APPENDIX 14 (continued)

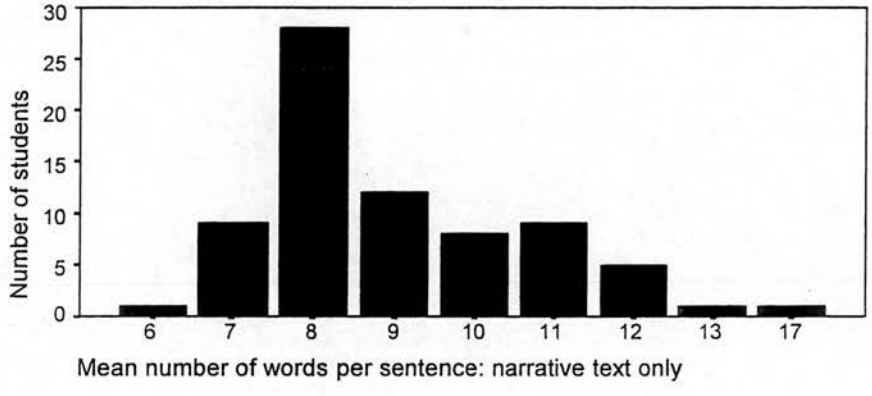
Distributions for mean numbers of words per sentence*: school 4

NARRATIVE TEXT ONLY

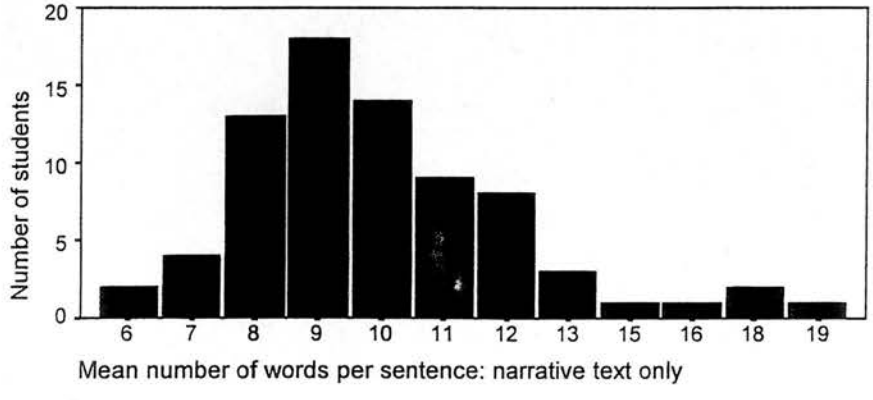
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

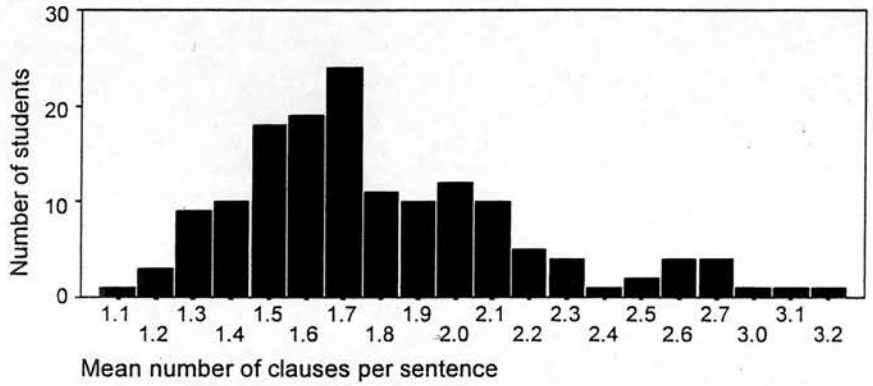


* Mean number of words per sentence is given to the nearest whole word

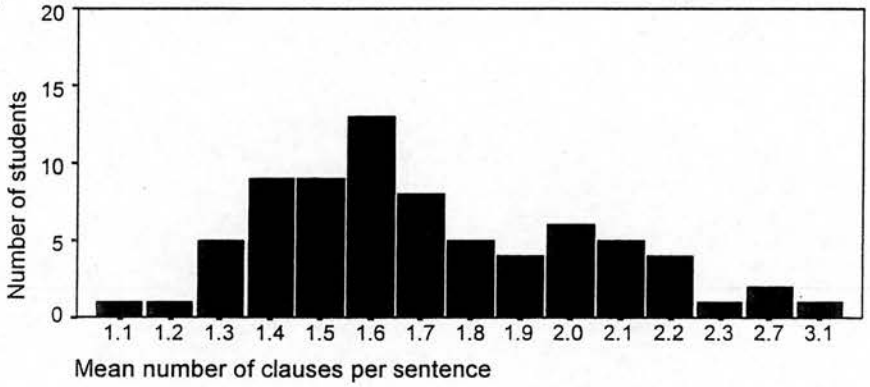
APPENDIX 15

Distributions for mean number of clauses per sentence: school 4

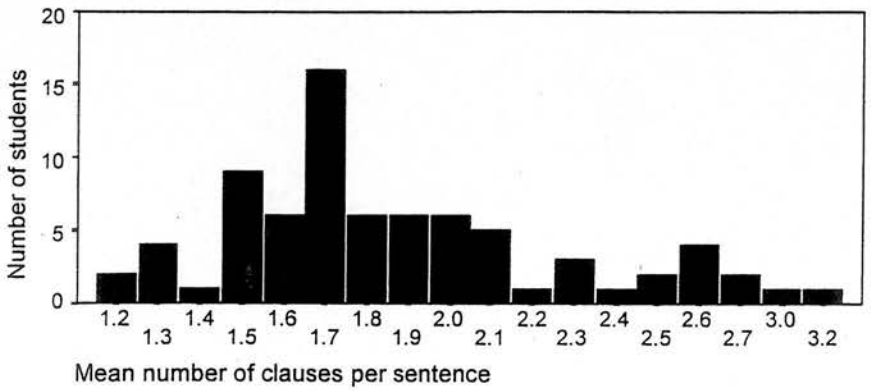
Complete data set (N = 150)



Control compositions (N = 74)



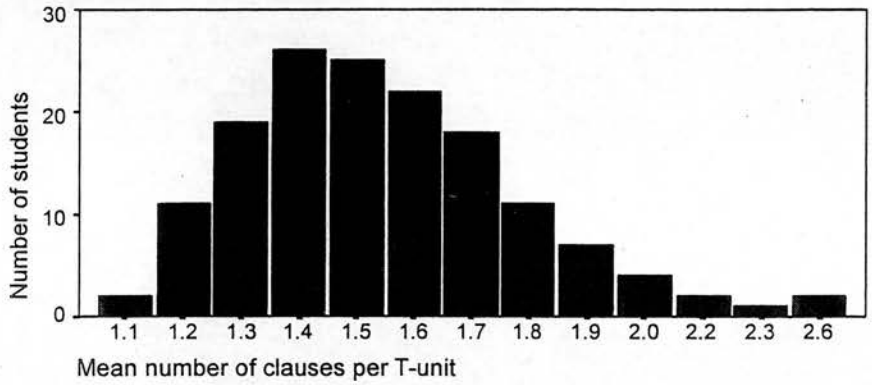
Experimental compositions (N = 76)



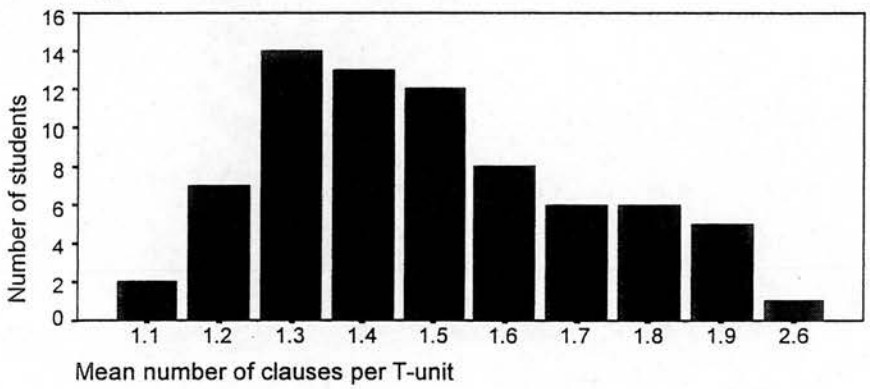
APPENDIX 15 (continued)

Distributions for mean number of clauses per T-unit: school 4

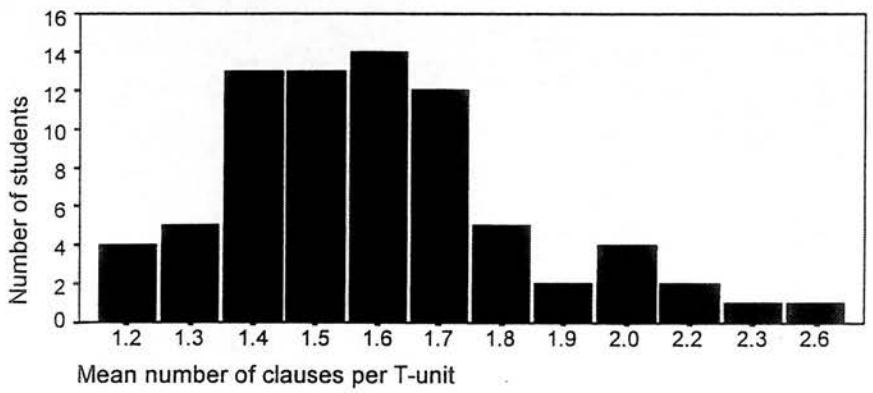
Complete data set (N = 150)



Control compositions (N = 74)



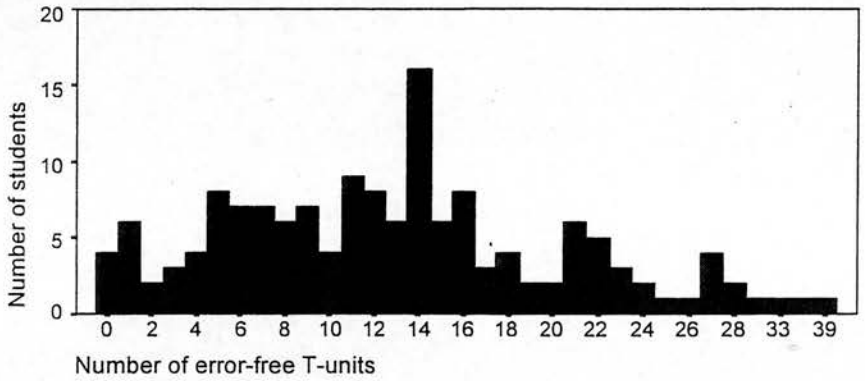
Experimental compositions (N = 76)



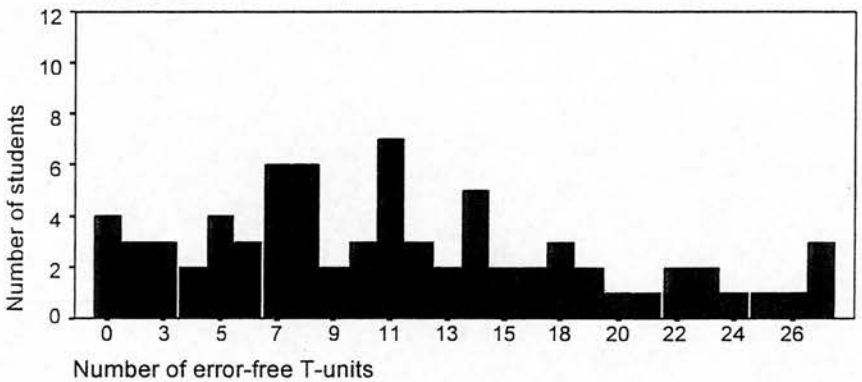
APPENDIX 16

Distributions for number of error-free T-units per composition: school 4

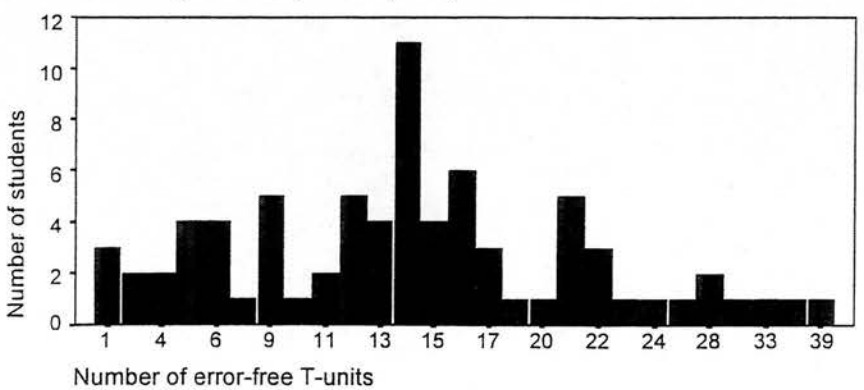
Complete data set (N = 150)



Control compositions (N = 74)



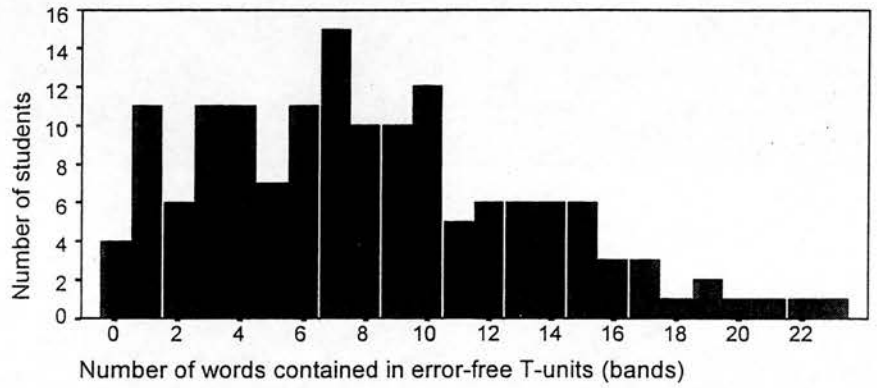
Experimental compositions (N = 76)



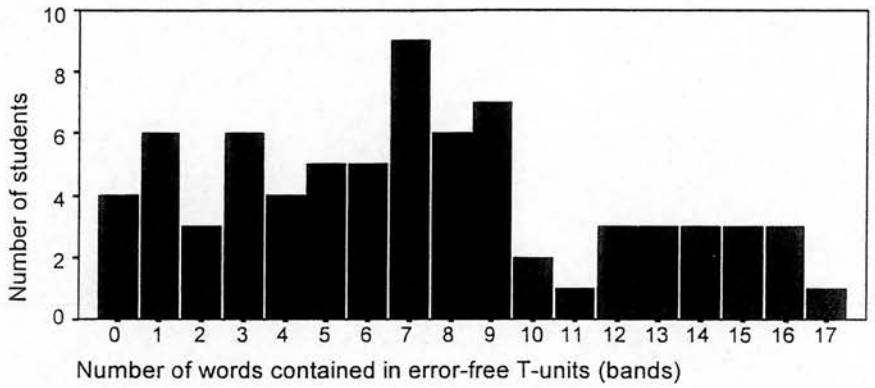
APPENDIX 16 (continued)

Distributions for number of words contained in error-free T-units per composition*: school 4

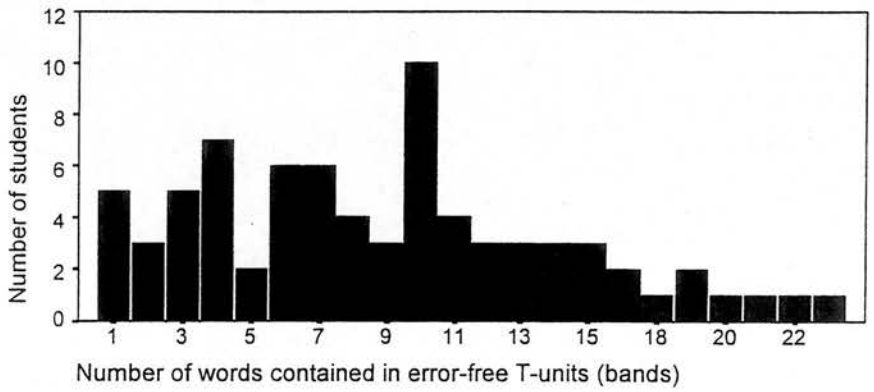
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

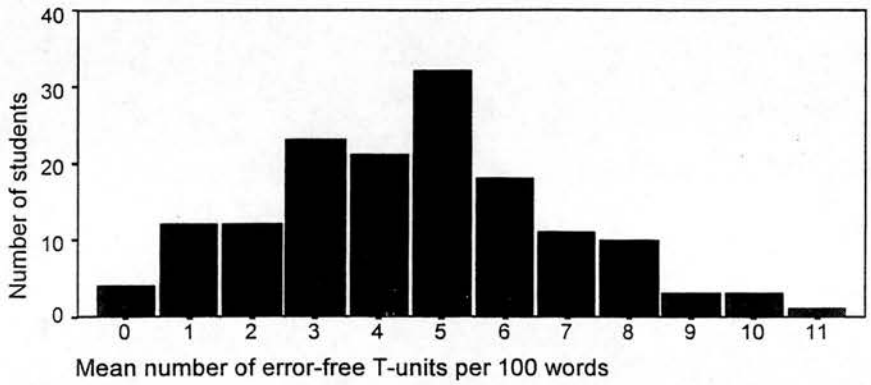


*Numbers of words contained in error-free T-units are in bands of 10; *e.g.* 5 includes compositions with number of words in error-free T-units ranging from 50 to 59

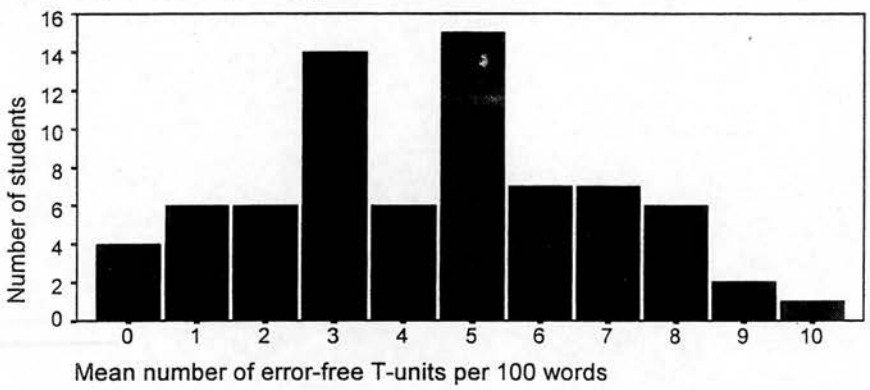
APPENDIX 16 (continued)

Distributions for mean number of error-free T-units per 100 words*: school 4

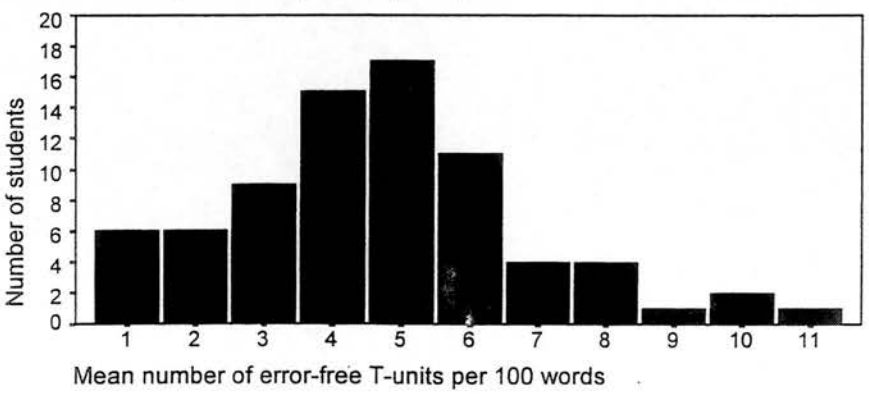
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

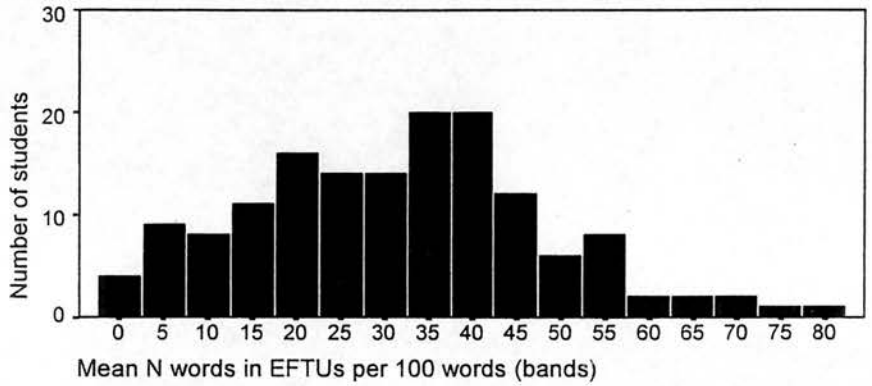


* Mean numbers of error-free T-units per 100 words are given to the nearest whole T-unit

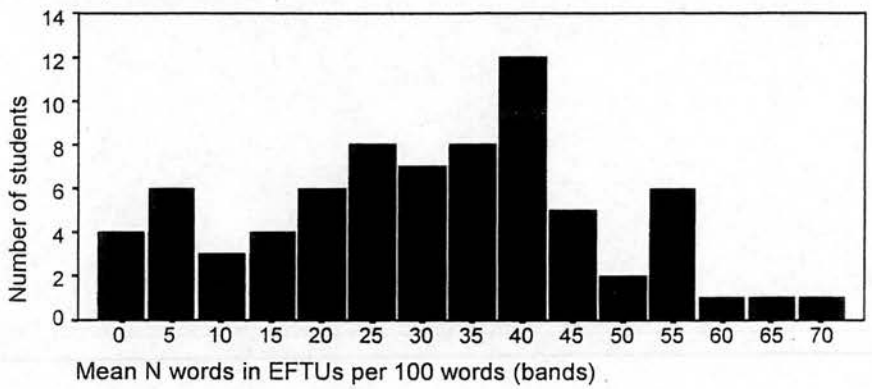
APPENDIX 16 (continued)

Distributions for mean number of words contained in error-freeT-units per 100 words*: school 4

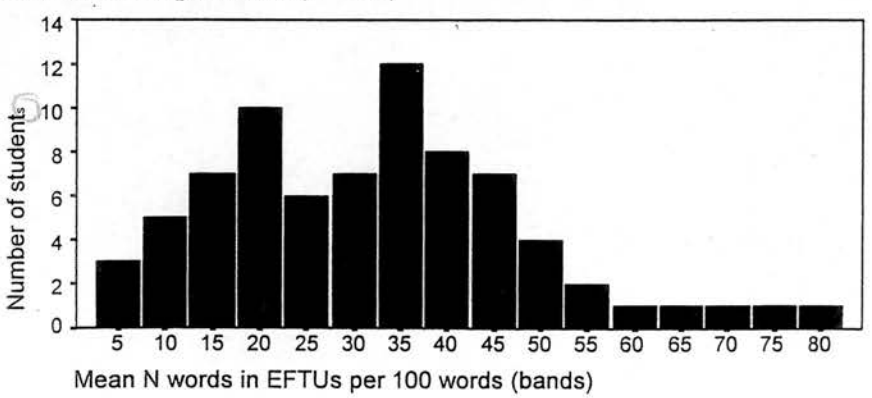
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)



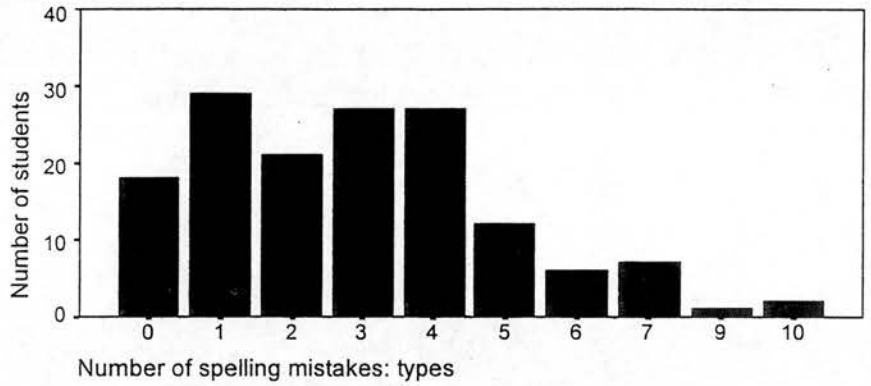
*Mean numbers of words contained in error-free T-units per 100 words are in bands of 5; *e.g.* 5 includes compositions with mean numbers of words in error-free T-units per 100 words ranging from 1 to 5

APPENDIX 17

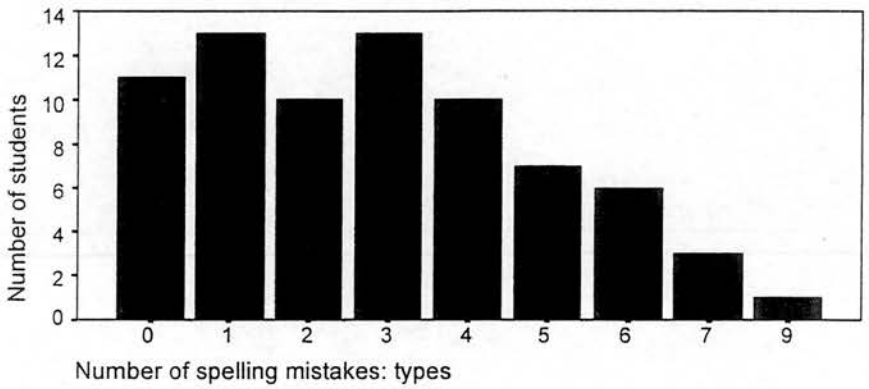
Distributions for numbers of spelling mistakes per composition: school 4

SPELLING MISTAKES: NUMBER OF TYPES

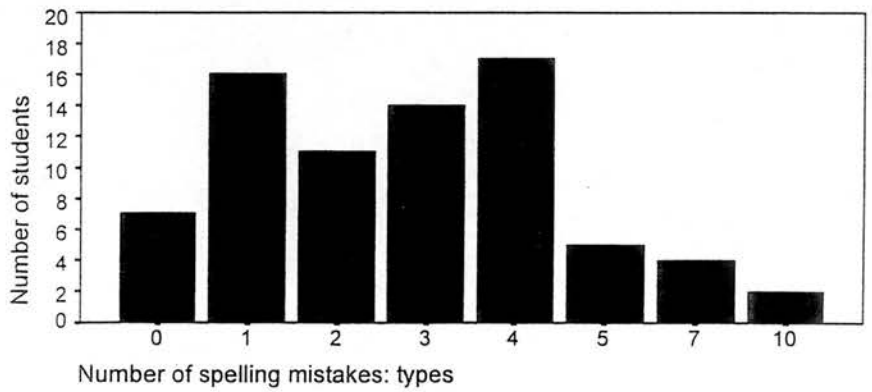
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)

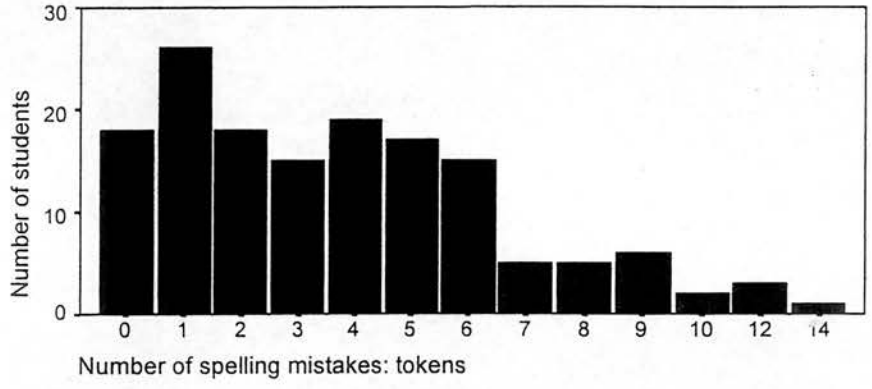


APPENDIX 17 (continued)

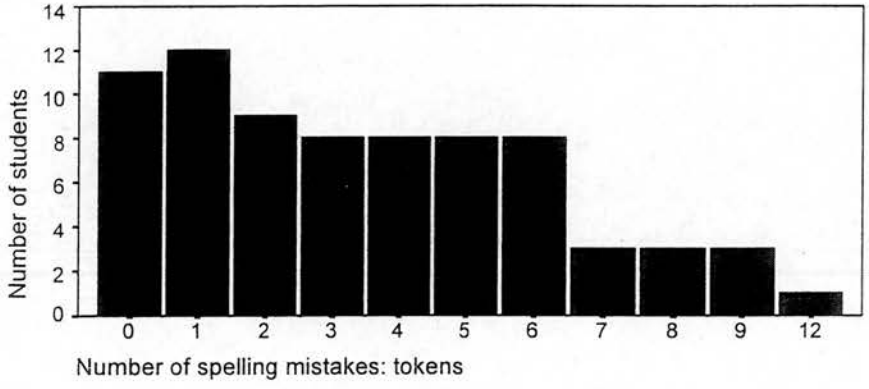
Distributions for numbers of spelling mistakes per composition: school 4

SPELLING MISTAKES: NUMBER OF TOKENS

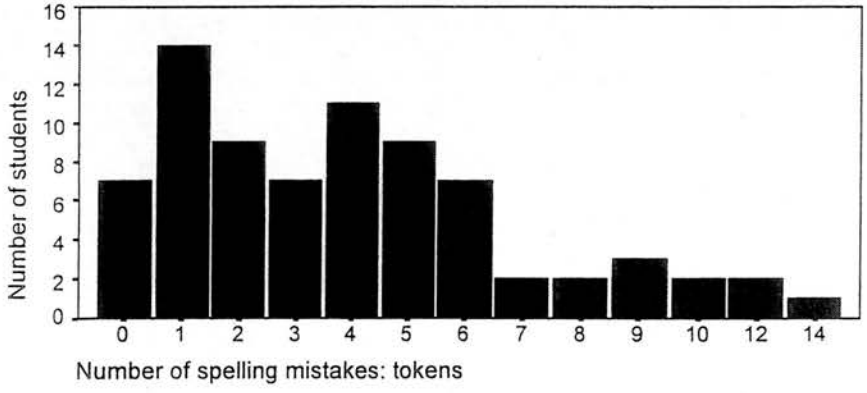
Complete data set (N = 150)



Control compositions (N = 74)



Experimental compositions (N = 76)



APPENDIX 18

Rules for correcting punctuation:

- Sentences which begin with "And", "But" or "So" are acceptable as sentences if these have a capital letter, are preceded by a full-stop, exclamation mark or question mark and contain a main clause.

  Example:
  We stayed in the house at night. *And we went home next day.*

- Sentence fragments (i.e. which do not constitute a main clause) are to be added to the relevant sentence.

  Examples:
  After we had dinner. We went to bed. = 1 sentence
  When we got tired. We take a rest. = 1 sentence
  After it. I saw a lot of gold in front of me. = 1 sentence

- A text unit which is clearly intended to be a sentence but which has no main verb because of grammatical inaccuracy is considered to be a sentence.

  Example:
  We back to the house the next day.

- When there is a comma between clauses but no connector, and the second clause contains a subject or subject pronoun, this constitutes two sentences.

  Examples:
  I and my friend Ken got lost, we find an empty house to spend a night, people said an empty house was full of ghosts, but we don't belived about these story. = 3 sentences
  We can't find the way home, we [missing verb] *too tired to walk.* = 2 sentences

- When two consecutive clauses are not separated by a comma or joined by a connector but both clauses contain a subject or subject pronoun, this constitutes two sentences.

  Examples:
  Yesterday, I stayed at home it was a boring day. = 2 sentences
  We go into the house, and saw white smoke move above the floor we are very afraid. = 2 sentences

- When two or more consecutive clauses with the same subject are separated by a comma, and there is no connector but the subject is ellipted, this constitutes one sentence so long as these clauses form a coherent sequence.

  Example:
  We went in the house, went upstairs, saw a man. = 1 sentence

  However, if more than three consecutive clauses are joined in this way, these should be divided into two sentences. More than six should be divided into three sentences.

- An interjection followed by an exclamation mark, question mark or full-stop is an independent sentence. An interjection followed by a comma is part of the sentence which follows.

  Examples:
  Oh! It a ghost! = 2 sentences
  Oh, it was Mary. = 1 sentence

APPENDIX 18 (continued)

Sample uncorrected and corrected composition (punctuation): school 4

Composition 29 (uncorrected)

One day, I went to a small island with a friend. There were only a few people lived on there and there were a few house.

That day morning, I met my friend at the pier which had ferry went to the island. After two hours, we reached the island and went on the island for a walk. When I was walking on the island, I only saw a few house, a few people, and many trees. At last, we saw a large forest, so we went in it. When we went in it, we saw many special animals, special plants. After a long time, the sky turned dark, and we wanted to left, but when we were leaving, we lost the way, we still walked, at last, we walked through the forest, but we were on the opposite side, and we saw a very old house, and we were tired so we decided to spend the night at the house. I had read a book about this house, it said the house was full of ghosts. But we were very tired, so we went in, when we went in, the light turned on, and the door closed. we were very afraid. after a few minutes, the long table on the sitting room full of food, and a voice told me ate the food. We were very hungry, so we ate the food. At that time, the voice told me went up stair and slept in the room, so we went up stair and went in the room it told me.

Next day, we got up and down stair, we saw the table full of food, and the voice told me ate them, and we ate them, after we had eaten the food, we said thank you and left. Although the house was full of ghost, but I though they were good ghosts.

(16 sentences)

Corrected version

One day, I went to a small island with a friend. There were only a few people lived on there and there were a few house.

That day morning, I met my friend at the pier which had ferry went to the island. After two hours, we reached the island and went on the island for a walk. When I was walking on the island, I only saw a few house, a few people, and many trees. At last, we saw a large forest, so we went in it. When we went in it, we saw many special animals, special plants. After a long time, the sky turned dark, and we wanted to left, but when we were leaving, we lost the way. We still walked. At last, we walked through the forest, but we were on the opposite side, and we saw a very old house, and we were tired so we decided to

spend the night at the house. I had read a book about this house. It said the house was full of ghosts. But we were very tired, so we went in. When we went in, the light turned on, and the door closed. We were very afraid. After a few minutes, the long table on the sitting room full of food, and a voice told me ate the food. We were very hungry, so we ate the food. At that time, the voice told me went up stair and slept in the room, so we went up stair and went in the room it told me.

Next day, we got up and down stair. We saw the table full of food, and the voice told me ate them, and we ate them. After we had eaten the food, we said thank you and left. Although the house was full of ghost, but I though they were good ghosts.

(22 sentences)

APPENDIX 19

Coding guidelines for T-units

Definition:

A T-unit is one main clause and all its dependent clauses with the following rules:

Coordination:

A coordinating clause containing a subject is an independent clause and constitutes a T-unit in its own right.

Examples:
We ran into the forest *// and there was a big tree.*
After a few minutes, the rains stop *// and we wanted to went out of the forest // but it was fog.*
Then we wanted to went out *// but we could not because we could not opened the door.*

A coordinating clause with no stated subject (i.e. because of ellipsis) is dependent on its associated main clause and is part of the same T-unit.

Examples:
We were very afraid and fall down the stairs. = 1 T-unit
Then I ran to my house very fast and put some water, food and clothes in the bag. = 1 T-unit

Subordination:

A subordinate clause is a dependent clause.

Example:
He wanted me to kill a person because he hated the person very much. = 1 T-unit

So occurring in the middle of a sentence and acting as a connector introducing the result of an immediately preceding clause is a subordinator: the ensuing clause is not independent.

Example:
I thought he was ill, so I helped him to come to his home = 1 T-unit

So occurring at the beginning of a correctly punctuated sentence is a conjunct and marks the beginning of a new T-unit.

Example:
I thought it was very easy. *// So I said "Yes!"* = 2 T-units

Verbless sentences:

A verbless text unit which appears to be a grammatically incorrect clause (because of omission of a main verb) is coded in the same way as if the verb had not been omitted.

Examples:
We very tired. = 1 T-unit
The light suddenly on. = 1 T-unit

Grammatically correct verbless expressions, including single-word interjections, which are punctuated as sentences, and are clearly not part of another sentence, are classed as independent T-units.*

Examples:
He said that he wanted my blood! *// All of my blood!*
"3 wishes? // You can give me wishes?"
Oh! // I saw a very ugly man who just like a ghost.

An interjection, such as *oh*, followed by a comma, which appears to be part of a sentence, is not an independent T-unit.

Examples:
Oh, she was the pretty woman in the world! = 1 T-unit
Oh, that was a very unlucky night for me. = 1 T-unit

Reported dialogue:

A reporting verb + subject immediately previous to directly reported dialogue is the main clause for the first text unit within the reported dialogue: thus the first reported text unit is dependent on the reporting clause; subsequent text units may be independent.

Examples:
We loudly say, "Oh! // Ghosts!"
And he said, "You are a good boy. // You help me. *//* I give three wishes to you."

A reporting verb + subject immediately following directly reported dialogue is the main clause for the last text unit within the reported dialogue: thus the last reported text unit is dependent on the reporting clause; previous text units may be independent.

Examples:
"Come! *// Come!" John said happily.*
"Are that real thing. *// Are you have a dream?" my friend said.*

* **Note**: this kind of verbless T-unit was excluded from calculations involving mean length of T-unit

Sample coded composition: school 4

Composition 213

On last holiday, I and my friends went to the camping. // But when we went to the forest, we got lost. // The sky was black, // but we could not find it out the forest. // In that night, we walked and walked. // Soon we found a house. // It was very big house. // When we opened the door, it was very dark and dust. // I shouted, "Have anyone in that", // but nobody answered. // Suddenly Mary said "I hear people say a night at the empty house was full of ghosts". // I said, "I do not believe it, // we are find the bedroom to sleep now." // Jane said, "I'm very afraid..." // I said, "Don't worry about it." // So we up the stairs to the second floor. // When we opened the first door, suddenly the door open it. // We were very afraid. // But nobody in it. // Jane said, "We are left this house, // I am very afraid." // Mary said, "But we are not find other house in the forest. // In the forest is very danger." // So we were opened the second door. // I said, "Oh! // It a bedroom." // So we were came in the bedroom and slept in that. // But sometime we were hear a strange sound. //

In the next day, the sun was very shiny. // We were woke up and left the strange house. // Soon we found the road to get out the forest and went home. // This holiday was very unforgettable. //

APPENDIX 20

Clauseless production units excluded from calculation of mean length of T-unit: school 4; control compositions

Oh! *15 times*
OK! *5 times*
Help! *twice*
Oh, yes! *twice*
Yes! *twice*

3 wishes?
A skull!
About 15.
All of my blood!
Bye bye!
Five years!
Ghosts!
Ha!
Ha ha!
Hey, boy!
No problem.
Of course!
Oh dear!
Oh my god!
Oh, no!
OK?
Okay!
Sure!
Very frightened!
Wah!
Wo!

Sandy! *twice*
Mary!
Sally!

APPENDIX 20 (continued)

Clauseless production units excluded from calculation of mean length of T-unit: school 4; experimental compositions

Oh! *fourteen times*
OK! *six times*
Yes! *six times*
No. *four times*
Of course! *three times*
Oh dear! *three times*
Australia? *twice*
But why? *twice*
Bye! *twice*
Hello! *twice*
No! *twice*
No problem. *twice*
Oh, no! *twice*
Okay! *twice*

3 wishes?
A monster!'
Ah!
Ah.....Ah."
All right.
Ghosts!
Girlfriend?
Good luck to you!
Just a body.
Magician!
Magic-stick!
Maybe!
Me?
No wars.
Oh yes!
Please!
Sorry!
Sure.
Thanks for your wishes.
Um!
Wa!
What!
What?
Wo!
Yes, of course.
You silly girl.

Tom! *six times*
Miranda.
Winnie!

APPENDIX 21

Examples of error-free and with-error T-units: school 4

Definition of *error*:

- Error refers to *grammatical error* or *incorrect use of vocabulary*.
- Spelling mistakes and poor or wrong punctuation are irrelevant.
- Errors of usage are to be discounted unless, as a result of the error, the sentence of which the affected T-unit forms part does not make sense in relation to the rest of the text.
- Vocabulary which is not quite appropriate but which nonetheless results in a sentence which makes sense in relation to the rest of the text is to be considered acceptable.
- Text which is awkward but does not violate a specific grammatical rule is to be regarded as *error-free*.

Examples of T-units coded as *error-free:*

Next day, when I got up. I was on my bed,

She was very hungry and dirty,

She said that she was very happy to meet me and gave three wishes to me.

When we walked through the forest, we got lost in it.

That condition was I should do some good things every day.

We went into the timple.

I thought he may be knew magic.

Then, we opend it.

We went upstair and looked around the house.

Suddenly there was a storm. // After it, we saw a house in front of us.

He used a black canvas to cover his face except his eyes.

And Josephine followed me.

Then we slept again.

Our clothes were wet, so we went into it.

It just had a table in the middle of the house. *(Previous sentence contained "house" to which "it" may refer.)*

I was very frightened at that time.

Examples of T-units coded as *with-error:*

One day, I saw a magician was play in the street.
He worn beautiful,
I thought he was come to another city.
he would like choose someone to play.
He was play wonderful, so many people were enjoy him.
I was see him what do he play wonderful.
I felt happy because magic was a exciting things.
That house had a lot of special substances were made by timber.
The house's owner was a old man and also was a magician.
The magician had a little strange, who offered to made three wishes come ture.
We were happy that had three wishes come ture.
The magician was agreed to us.
Finally, we were a rich people, had a good girl friends.
But they didn't come their home on next week.
I never see Sam and Tom every day.
Someone said me about the forest that was very terror,
Suddenly, the house's door broke down.
One night, I and my friend, Peggy got lost in our picnic.
We knocked the house more time.
and we open the door for myself.
We clean the house and visited this house.

APPENDIX 22

Identified clause types: school 4

Primary clause types

Main clause
Full coordinate clause
Reduced coordinate clause
Defining relative clause
Non-defining relative clause
Subordinate clause

Subordinate clauses were further coded as:

complement clauses

finite complement clause
examples:
I didn't know *what he will do*; I said *that I will give it to him*; I saw *there was a woman*; Please tell me *why am I here*

to- infinitive complement clause
examples (post-verb):
I don't know *what to do*; We wanted *to find a place for a picnic*
(post-adjective):
It was not easy *to do that*; I was ready *to go*

bare infinitive complement clause
examples:
Let me *see the house*; He made me *do that*

-ing complement clause
example:
I saw *a woman flying above our head*

adverbial clauses of:

time
examples:
When we got up, it was raining; *After we had dinner*, we went to sleep.

manner
example:
We did that *as quick as we can.*

condition
example:
If you help me, I will give you three wishes.

concession
example:
Although it was dark, we can see the ghost.

cause/reason
example:
I ate the food *because I* [missing verb] *very hungry.*

purpose
example:
We went in the house *to spend a night.*

result
example:
At that night, we can't sleep very well, *so we made a fire.*

other clauses:

this category was used for clauses which did not fit into any of the above categories or which were unclear.

Coding

Clauses were coded using the same set of guidelines as for correcting punctuation (see Appendix 18). Thus *Yesterday, I stayed at home it was a boring day* was coded as two main clauses: *I and my friend Ken got lost, we find an empty house to spend a night* was coded as two main clauses and one subordinate clause: *We went in the house, went upstairs, saw a man* was coded as one main clause and two reduced coordinating clauses.

Text units which functioned as clauses but which had no verb due to grammatical inaccuracy were coded as clauses.

Example:
John said me to come home *but I not with my friend to the bus.*

Coordinating complement clauses which were the complement of an ellipted main clause were coded as subordinate clauses and not as coordinating clauses.

Examples:
I saw him running to the house *and opening the door.*
I hope I have a nice house *and live there with my family.*

APPENDIX 23

Frequencies of identified subordinate clause types: school 4

Table 9.1 ***Frequencies of identified subordinate clause types: school 4; four classes***

Subordinate clause type	**Lower level control class:** N = 34	**Lower level experimental class:** N = 33	**Higher level control class:** N = 40	**Higher level experimental class:** N = 43
finite complement	186	228	200	305
to- infinitive complement	59	88	86	94
bare infinitive complement	3	11	13	13
-ing complement	2	7	14	8
adverbial of time	44	68	51	99
adverbial of manner	2	1	3	4
adverbial of condition	10	25	13	16
adverbial of concession	1	3	9	13
adverbial of cause/reason	37	29	58	51
adverbial of purpose	39	38	50	73
adverbial of result	24	25	42	62
defining relative	22	51	47	81
non-defining relative	0	5	1	7
other	1	0	15	9
total N subordinate clauses	430	579	602	835

APPENDIX 24

Verb coding system and sample coded compositions: school 4

1. Primary codes

R = Regular verb: simple past
I = Irregular verb: simple past
M = past Modal
B = copular Be
P = Passive
T = *to-* infinitive: Tb = bald infinitive
O = Other verb form correctly used: OX = Other verb form inappropriately used
U = Uncoded (either the verb does not fall clearly into any of the other categories or it is not clear what form/tense is intended by the student)

Supplementary codes
(These categories are not discussed in the quantitative analysis. Reporting verbs are included in regular and irregular verb counts. Missing copular is included in Be error counts.)

V = reporting verb
Z = missing copular
$ = verb contained in directly reported dialogue

2. Tense codes (for verbs coded as *copular Be*, *Passive* or *Other*)

f = simple present
g = present continuous
h = *was going to* + infinitive
j = simple past
k = past continuous
w = past perfect
y = present perfect

3. Assertion codes

D = Declarative
N = Negative
Q = Interrogative

4. Accuracy codes

C = correct form
E = incorrect form
a = agreement error but otherwise correct form

5. Types of error (for verbs coded as *Regular simple past, Irregular simple past, past Modal, or copular Be*)

Regular and irregular simple past declarative

E1: present tense used instead of past tense
*examples: goes (*for *went); laugh (*for *laughed)*

E4: irregular verb made regular
examples: catched; sleeped; feeled

E6: present tense BE + past tense V
*examples: is made (*for *made); am fainted*

E7: past tense BE + root V
*examples: was listen (*for *listened); was disappear; was stand*

E8: past tense BE + past tense V
*examples: was occurred; were said (*for *said); was fell*

E9: present tense HAVE + past tense V
*example: have saw (*for *saw)*

E10: present perfect for simple past
*example: have seen (*for *saw)*

E11: V+ing
*example: running (*for *ran)*

E12: irregular past participle
*example: known (*for *knew)*

E13: past continuous instead of simple past
*example: was taking (*for *took)*

EU: uncoded error

Regular and irregular simple past negative

E1: present tense used instead of past tense
*examples: don't know (*for *didn't know); don't answers*

E2: present tense negative auxiliary + past tense V
examples: don't wanted; don't believed; not remembered

E3: past tense negative auxiliary + past tense V
examples: didn't spent; didn't believed

E5: present tense negative BE + root V
*example: is not listen (*for *didn't listen)*

E8: past tense negative BE + past tense V
*example: was not believed (*for *didn't believe)*

E9: present tense negative HAVE + past tense V
*example: haven't saw (*for *didn't see)*

EU: uncoded error

Regular and irregular simple past interrogative

E1: present tense used instead of past tense
*example: do you see (*for *did you see)*

EU: uncoded error

Past modals

E1: present tense used instead of past tense
*examples: can't find (*for *couldn't find); must to find (*for *had to find)*

E2: present modal + simple past
examples: can studied; can't found; will designed; will lost

E3: past modal + simple past
examples: could went; could not walked; should woke

EU: uncoded error

BE

E1: present tense used instead of past tense

EU: uncoded error

Sample coded compositions

Composition 200

One day, My friends and I went <IDC> to have <TC> a picnic. We were <BjDC> happy, because we had wanted <OwDC> to have <TC> a picnic for a long time.

On the way, the weather turned <RDC> bad suddenly. We became <IDC> unhappy, but we kept <IDC> walking <UC> until we'd found <OwDC> a good place. When we got <IDC> into a hut, It rained <RDC>. One of my friend said <VIDC> that there were <BjDC> ghosts in this hut. We were <BjDC> very frightened. Suddenly, We heard <IDC> a strange voice from a man. We felt <IDC> very very frightened and all my friends ran <IDC> away except Fredy and I, because we wanted <RDC> to know <TC> what had happened <OwDC>. Then I saw <IDC> a white thing flying <UC> above my friend,

Fredy's head. I screamed <RDC>. My friend then look <RDE1> above, but he couldn't see <MNC> it, because It had already pass <OwDE> through the wall. I told <VIDC> him what had happened <OwDC> but he didn't believe <RNC> me. I wanted <RDC> to leave <TC> the house, but Fredy didn't let <INC> me go <TbC>, so I stayed <RDC> there. We sat <IDC> on a chair. After that, we heard <IDC> the strange voice again. It was <BjDC> horribled. The white thing appeared <RDC> again. It was <BjDC> a man who was <BjDC> transparent. We were <BjDC> very very very frightened, but I heard <IDC> a voice of my friend. Oh! Those was make <PjDE> by my friends.

Composition 88

It was <BjDC> a sunny day, so I phoned <RDC> to my friend to ask <TC> her to had <TE> a picnic with me. She said <VIDC>, "Would we go <$> to somewhere that we never gone <$>. We could spend <$> one day in there and I would bring <$> a map. We willn't lost <$>." That's <BjDE1> a good idea so we got <IDC> on a bus and then by ferry. We also walked <RDC> for a long time and had <IDC> a rest in a comfortable place. We had <IDC> our lunch then sleeped <IDE4> about 1 hour. The birds sang <IDC> on the trees. The wind blew <IDC> gentlely. That was <BjDC> we never feel <U> at home.
Quickly, It's <BjDE1> the time to go <TC> home. But when we walked <RDC> to the nearest bus stop, the last bus had gone <OwDC>. We was <BjDCa> so worried, we can't go <MNE1> home tonight. We walked <RDC> back to the hill tirely. But luckly, we saw <IDC> a oldest house on the top of the hill, it looked <RDC> dirty and dangerous. But nobody live <RDE1> there so we came <IDC> in the house. There was <BjDC> dark, only a few candles with small light. It was <BjDC> the ghost house in the story book, but it was <BjDC> the only place for us to stay <TC>. We was <BjDCa> hungry but we didn't like <RNC> the food in the house. In the second floor, we stayed <RDC> in the one of the bedrooms. We can't sleep <MNE1> because the voice was <BjDC> terrible, the wind was <BjDC> cold and the windows can't close <MNE1> and made <IDC> a loud noice. We hope <RDE1> morning came* <IDC> at soon, but it was <BjDC> the midnight, a long time to the morning. We was <BjDCa> so fraidtened and we don't remember <OfDC> how we did go <IDEU> home.

*__Note:__ simple past forms are coded as correct or incorrect as regards form only. Although *came* is not used appropriately, it is a correct past form.

APPENDIX 25

Words which were recategorized as high frequency vocabulary (first 500 words) for Web VocabProfile analysis: school 4

School vocabulary:

classmate, exam, exams, football, geography, volleyball

"Global" English :

barbecue, cinema, sandwich, supermarket, television, T.V.

Transport vocabulary and words commonly seen on signs:

airport, exit, ferry, toilet

Words contained in the task rubric:

condition, conditions, ghost, ghosts, magician, special, true, wish, wishes

Hong Kong place names:

Aberdeen, Belilios, Chau Cheung, Chengdu, Hong Kong, Lantau, Mongkok, Peak, Saikung, Territories, Tsim-Sha-Tsui

Fictional companions:

Alex, Alice, Amy, Andy, Anita, Anna, Annie, Ariel, Billy, Bobby, Candy, Chan, Christina, Chu, Daisy, David, Eric, Esther, Eva, Fannie, Fiona, Franny, Fredy, Harry, Ivy, Jacky, Jane, Janet, Jenny, Jimmy, John, Johnson, Jonny, Jordy, Josephine, Judy, Karen, Ken, Kim, Lam, Leo, Leona, Lesley, Li, Linda, Louise, Maggie, Maria, Martin, Mary, McDonald, Meidy, Michelle, Micky, Nacky, Paul, Paula, Peggy, Peter, Raymond, Sally, Sam, Sandy, Sarah, Stephanie, Sue, Tim, Tom, Tommy, Vens, Vicky, Vivian, Wai, Wendy, William, Wilson, Winnie, Wong, Yoki, Yuen

APPENDIX 26

Rules for the Internal Word Frequency List

1. Rules for what constitutes a *type*

Nouns:

A singular noun, its regular plural and singular and plural possessive forms constitute a single type.
Example: *girl*, *girls*, *girl's* and *girls'* = one type

A singular noun ending in a *consonant* + *y* and its plural *-ies* form are a single type.
Example: *baby* and *babies* = one type

An irregular plural constitutes a different type from the corresponding singular form (since it is possible to know the singular form without knowing the plural form and vice versa).
Example: *person* and *people* = two types

Adjectives:

An adjective, its *-er* comparative and *-est* superlative constitute a single type regardless of any minor spelling changes.
Examples: *old*, *older* and *oldest* = one type; *happy*, *happier* and *happiest* = one type; *big*, *bigger* and *biggest* = one type

Adverbs:

An adverb and its associated adjective are two types.
Example: *quick* and *quickly* = two types

Ordinary verbs:

A root form, simple present forms and *-ing* form constitute a single type regardless of any minor spelling changes.
Examples: *buy*, *buys* and *buying* = one type; *come*, *comes* and *coming* = one type

The simple past form for both regular and irregular verbs constitutes a different type from the root form.
Examples: *like* and *liked* = two types; *buy* and *bought* = two types

The past participle of a *regular* verb is counted as being the same type as the corresponding simple past. The past participle of an *irregular* verb is a new type.
Examples: *liked* (simple past) and *liked* (past participle) = one type; *ate* and *eaten* = two types; *bought* (simple past) and *bought* (past participle) = two types

Auxiliaries:

Forms which cannot be computed from a base form are independent types.
Examples: be, *am*, *are* and *is* = four types; *have* and *has* = two types

Contractions:

Contractions are independent types.
Examples: *had* and *hadn't* = two types; *could* and *couldn't* = two types; *I'll* is one type and is distinct from the two types *I* and *will; it's* = one type

Other words:

All other words (e.g. prepositions, conjunctions, articles, pronouns) are single types.

Closely related homonyms:

Closely related homonyms (sometimes called homomorphs) are separate types.
Example: rain (noun) and *rain* (verb) = two types

Excluded words:

Proper names of people
Names of Hong Kong locations
non-words: *examples*; *wo*, *wah*, *um*, *ha*, *dok* (*I heard a "dok" sound*)
unidentifiable words

2. Rules for frequency rating

Types are assigned to a frequency band according to their number of occurrences in the data (compositions), with the following additional rules for verbs:

if a simple past form occurs more frequently than the corresponding root or present form, the past form *type* subsumes the root form *type* (which includes *-s* and *-ing* forms) and all forms are grouped together to calculate the frequency rating; present and root forms are then assigned to the same frequency band as the past form *type*
examples: *rang* subsumes *ring* and *ringing*; *was* subsumes *is* and *am*

if a root or present form occurs more frequently than the corresponding past form, the past form *type* may appear in a higher (lower frequency) band than the root form *type*
examples: can = band 3, *could* = band 4; *know* = band 4, *knew* = band 5

if the root form *type* and the past form *type* have the same number of occurrences, neither subsumes the other and both may appear in the same band as being two independent *types*
example: I've and *I'd* both appear at band 9

APPENDIX 27

Internal Word Frequency List and sample coded composition: school 4

Level 0 types

a
and
I
the
to
was/is/am
we

APPENDIX 27 (continued)

Level 1 types

but
had/has/have/having
he
house/houses/house's
in
it
me
my
of
saw/see/seeing
said/say/says/saying
that
very
went/go/goes/going
were/are
you

APPENDIX 27 (continued)

Level 2 types

afraid
at
because
came/come/coming
didn't/doesn't/don't
door/doors
found/find/finding
friend/friends/friend's
ghost/ghosts/ghost's
not
on
one
out
so
some
suddenly
then
there
this
walked/walk/walking
wanted/want
when
wish/wishes (*noun)*

APPENDIX 27 (continued)

Level 3 types

after
all
can
couldn't/can't
day/days/day's
felt/feel/feeling
for
forest/forests/forest's
got/get/getting
heard/hear
him
magician/magician's
man
many
night/night's
no
old/older/oldest
people/people's
she
told/tell
they
thought/think/thinking
three
time/times
up
with

APPENDIX 27 (continued)

Level 4 types

about
again
an
asked/ask
back *(adverb)*
be
believed/believe
big/bigger
could
dark
did/do/does/doing
down
empty
frightened/frighten
girl/girls
give/giving
happy/happiest
her
home/home's
if
I'm
into
know
last
left/leave/leaving
long/longer
looked/look/looking
lost/lose
lot/lots
money
oh
opened/open/opening
our
quickly
ran/run/running
sleep/sleeping
thing/things
true
us
what
will

APPENDIX 27 (continued)

Level 5 types

also
answered/answer/answers
any
around
arrived/arrive
as
ate/eat/eating
away
beautiful
became/become/becomes/becoming
bed/beds
book/books/book's
by
called/call/calling
condition/conditions
decided/decide
disappeared/disappear
dream/dreams *(noun)*
few
first
floor
food/foods
from
front
full
gave
good
happened/happen/happening
help/helping
here
hill/hills
his
it's
just
knew
life
light/lights *(noun)*
like *(similar to)*
loudly
made/make/making
magic
met/meet
minute/minutes
more
morning
mother/mother's
must
near/nearest
never
next
noise/noises
now
o'clock
only
other *(adjective)*
place/places
play/playing/plays
rain/rains/raining
rain *(noun)*
road/roads
room/rooms
sat/sit/sitting
shouted/shout/shouting
sky
slept
someone/someone's
something
sound/sounds *(noun)*
special
still
story/stories
strange
them
tired
too
took/take/taking
tree/trees
turned/turn/turning
two
voice/voices
way
which
who
why

window/windows
woke/wake
woman/woman's
world/worlds
year/years
yes
your

APPENDIX 27 (continued)

Level 6 types

ago
along
always
anything
appeared/appear
bad
bag/bags
bedroom/bedrooms
before
began/begin/beginning
behind
black
bought/buy/buying
camp *(noun)*
change *(verb)*
closed/close
clothes
cold/colder
cried/cry/crying
dirty
evening
every
eye/eyes
face/faces
family/family's
fast/faster
father
finished/finish
fire/fire's
fly/flies/flying
followed/follow/following
hair/hairs
hand/hands
hasn't
haven't
hope *(verb)*
horrible
hour/hours
how
hungry
inside
island
knocked/knock
late/later
leg/legs
let
like *(verb)*
lived/live/lives/living
loud/louder
moment
mountain/mountains
much
name
needed/need
nobody
nothing
OK
once
or
picnic *(noun)*
please
promised/promise
put
really
remember
rich/richest
school/school's
screamed/scream/screaming
second *(adjective)*
should
small/smallest
smiled/smile/smiles/smiling
soon
spend
stair/stairs
stand/standing
started/start
stayed/stay
stop *(verb)*
street
sun
Sunday
table
talked/talk/talking

third
through
together
until
used/use
wear/wears/wearing
where
white
wind/winds
worried/worry
would

APPENDIX 27 (continued)

Level 7 types

afternoon
agreed/agree
alone
although
angry
animal/animals
another
anyone
baby/babies
beach
bird/birds
blew/blow/blowing
blood
body
boy/boys
brave
broke/break/breaking
building
camping
candle/candles
car/cars
caught/catch/catching
chair/chairs
changed
clever/cleverest
country/countries
course (*of course*)
dead
dear
died/die
dinner
discovered
dreaming
everyone/everyone's
exam/exams
except
fell/fall/falling
finally
five
football
forever
forgot/forget
furniture
God
gone
hadn't
happily
head
health
heavy/heavier
helped
hide/hiding
hit/hits *(verb)*
holiday/holidays
hoped
hotel
immediately
jumped/jump/jumping
lady/ladies
large/larger
lighted/light/lighting
listen/listening
little
love
lunch
map
maybe
midnight
move/moving
new
offered/offer
outside
parent/parents
path/paths
person
photo/photos
picked/pick
played
police
poor
read/reading
remembered
replied/reply
rest *(noun)*

sad
same
Saturday
scared/scare
shadow
shock
slowly
stick
stopped
swim/swimming
tent/tents
terrible
thanked/thank
thanks
these
today
top
tried/try
T.V.
under
upstairs
village/village's
visit/visiting
wall/walls
wasn't/isn't
water
weather
well *(adverb)*
wet
whole
wish *(verb)*
wore
work *(noun)*
yesterday

APPENDIX 27 (continued)

Level 8 types

above
accepted/accept
already
answer/answers *(noun)*
apple
asleep
aunt/aunt's
Australia
ball
bat/bats *(animal)*
begged
beside
best
birthday
both
box/boxes
bread
breakfast
bright
broken
brother
brought/bring
bus
cake/cakes
cannot
card/cards
carefully
castle
cat/cats
ceiling
child
children/children's
cleaned/clean
clear *(adjective)*
clearly
climbed/climb
clock/clock's
cloud/clouds
colour
corner
countryside
covered/cover/covering
difficult
doctor
dog/dogs
dollar/dollars
drank/drink/drinks/drinking
dress/dresses *(noun)*
dust *(noun)*
early
easy
end *(noun)*
enjoyed/enjoy
everybody
everything
everywhere
excited/exciting
exit *(noun)*
fable/fables
fainted
false
fan/fans
far
ferry
film/film's
firstly
flat *(noun)*
flower/flowers
foot
Friday
fright
funny
future
game/games
garden
gold
grandmother
grass/grasses
grassland/grasslands
great
greedy
green *(adjective)*
ground
half

handsome
happiness
hard
hat
hated/hate
healthy
heart
hello
help *(noun)*
high/higher
hold/holding
hole/holes
Hong Kong/Hong Kong's
hopes *(noun)*
horror
hospital
hundred
hurried/hurry
hurt
hut
idea
I'll
important
interesting
itself
keep
killed/kill
kind *(adjective)*
kind/kinds *(noun)*
kitchen
lastly
laughed/laugh/laughing
liked
listened
locked/lock
lonely
lovely
luckily
lucky
machine
mark/marks *(noun)*
marry
matter
may *(modal)*
meal
men
middle
might *(modal)*
million/millions
mirror/mirrors
monster/monsters
month
most
mouth
moved
myself
nice
nine
off
park *(noun)*
party
passed/pass
past
peace
phoned/phone
picture/pictures
plane
prayed/pray/praying
problem
promise *(noun)*
pull
pushed/push/pushing
question/questions *(noun)*
quick
quietly
rained
rainy
rang/ring/ringing
red
result/results
right *(correct)*
river
rock/rocks (noun)
round
rushed/rush
saved/save
second/seconds
shiny
ship
shirt/shirts
shoe/shoes
shop/shops/shop's
short/shorter
side

silly
sing/singing
six
skull
sleep *(noun)*
somewhere
sorry
spent
spider/spiders/spider's
stone/stones
stood
storm
stranger *(noun)*
strong
stupid
summer
sunny
sure
surprised/surprising
surrounded/surround
sweetly
tall
tank
taxi
television
temple
ten
test/tests
than
that's
their
there's
thief
thirsty
though
threw
toilet
tomorrow
torch/torches
towards
travel/travelling
trouble
ugly
unhappy
unlucky
visited
waited/wait/waiting
war/wars
watch *(verb)*
week/weeks
weren't
while
wife
wished
wrong
you're
yourself

APPENDIX 27 (continued)

Level 9 types

accident
acetic
achievement
acid
across
activity
actually
added
address
adventure
advice
affect
afterwards
air
airport
alarm *(noun)*
alarmed
alive
allowed/allow
almost
alright
altogether
America
ancestor
anybody
anymore
anywhere
appointment
appreciate
area
argue
arm
armchair
athletic
attention
back *(part of body)*
background
badly
badminton
bank
barbecue
barking
basket
basketball
bath
Batman/Batman's
battery
beam
beard
beat/beating
beckoning
been
beer
beggar/beggars
Beijing
bell
better
between
bicycle/bicycles
bit *(noun)*
blind
blindness
blue
board
boat
bogus
bone
booking
bookshop
boost
boring
born
borrow
bottle
bottom
bowl
boyfriend
branches
brand
bravely
bridge
bubbles
bulb
burn

business
bus-stop
busy/busiest
bye
cafe
calm
Canada
cancel
canvas
cared/care
career
careful
careless
carelessly
carpet
carried
cartoon
case
causes
cave
celebrate
cellar
century
champion
chance
chased
chat
check
chemically
chest
chicken
China
Chinese
chocolate
choice
chose/choose
Christian
Christmas
church
cinema
circle
circus
city
class
classical
classmate
cloth
cloudy
coat
coffee
Coke
colourless
comfortable
comforted
community
compass
compensate
competition
complete
computer
concentrated
cone
consent
conserve
continued
controlled/control
cooked/cook/cooking
cool
corrected
corridor
cost
court
crashed/crash
crazy
cream
crossed/cross/crossing
crown
crows
cruel
crystal
cup
cured/cure
curious
cut
cute
dad
daily
damage
danced/dancing
danger
dangerous
darkness
daughter
daydreaming

deaf
death
December
deep
deeply
deer
delighted
demanded/demand
density
designed
desk
destination
destroyed
diary
difficulty
digging
dim
dining
direction/directions
disappointed
discussing
disease/diseases
dishes
distribute
doll
donated
done
doubt
downstairs
drain *(noun)*
draw
drawer
dressed
drifted
dripping
dropped
drove/driving
drunk
dry
due
during
dusty
each
ear/ears
earth
easily
eaten
eerie
egg
eight
either
electric
electrical
electricity
eleven
else
embrace
encouraged
encouragement
ending *(noun)*
England
English
enough
entered
entrance
equipment
error
escape
Europe
even
events
ever
evil
examination
example
excuse
exercises
experience
experiment
explain
explored
extremely
fact
farm *(noun)*
farmer
farming
fat
fault
feared
fearful
fed
fever
field
fifteen

fifteenth
fight
finger
fish *(noun)*
fishing
flash
flaw
fled
flew
floated
flock
flowing
fog
foggy
footpath
footsteps
forced
forgave/forgive
forgotten
form *(noun)*
formed
fortunately
fortune
fountain
four
free
friendship
frightenedly
frightful
frightfully
fruit
fur
further
gate
gentlemen
gently
geography
girlfriend
given
glass
golden
goodbye
government
gradually
grandfather/grandfather's
grateful
greedily
grew/grow
grey
group
growth
gun/guns
Halloween
hammer
handbag
handicapped
handle
harbour
hardly
hearsay
heavily
height
helicopter
hers
herself
hey
hi
honest
horrid
horse
hot
however
human
hung
hunter
hurry *(noun)*
husband
ice
I'd
ill
illusion
imagined/imagine
importance
included
indeed
independent
Indians
injure
insect/insects
instant
instead
intelligence
intention
introduce

invited
iron
its
I've
jacket
Japan
Japanese
Jesus
jewellery
job
joke
journey
juice
July
jungle
keeper
kept
key
kicked
kindly
kite
knife
knob
known
labyrinth
laid
lake
lamps
land
lay/lie/lying
layer
lazy
lead
legends
less
let's
lichen
lightbulb
lighter *(noun)*
lightly
lightning
line
lion
liquid
lit
lively
lock *(noun)*
look *(noun)*
lorry
loss
lover
low/lowest
luck
major
male
marbles
mark *(verb)*
market
married
match/matches *(to light fire)*
May
meanwhile
mentally
mess
method
metre/metres
mice
military
mind
mine
Miss *(title)*
missed/miss
mist
modern
Monday
monkey
moon
moonshine
mount
mouse
moustache
movement
Mum
mustn't
narrow
naughty
nearby
nearly
neck
necklace
nervous
nests
net
news

nodded
non-organic
noodles
north
nose
noticed/notice
nowhere
number
observed
occurred
ocean
often
okay
onto
opinion
opposite
orange
order
organ
organic
others
otherwise
ours
ourselves
outdoors
outlook
over
owl
owner
packed
paid/pay
pain
painted
pair
pane
paper
patient *(adjective)*
Peak
pen/pens
perfect
perhaps
period
phone *(noun)*
photographs
physically
picnicking
piece
pier
plan *(noun)*
planned
plants
pleased
plenty
point *(noun)*
pointed
policeman
policemen
politely
pollution
pool
pop
poster
power
practised
prepared/prepare
present *(noun)*
president
pretty
primary
prizes
processed
produced
product
project
protect
provide
pulled
punished/punish
puppy
quarter
queen
question *(verb)*
quiet
quite
rabbit
rack
radio
rats
ray/rays/ray's
reached
ready
real
realize
reappeared

reason
relatives
remaining
rent
repaired
report
reptiles
required
rest *(verb)*
restaurant
revision
ring/rings *(noun)*
rode/ride
roles
rose/rise/rises/rising
rubbish
rule
runaway
safety
sand
sandwich
sang
scar
scenery
scream *(noun)*
sea
secondly
secret
seek
seemed/seem/seems
seen
sent/send
September
serious
set
seven
several
sex
Shanghai
shape
she'd
sheep
shell/shells
shelves
she's
shine/shining
shook
shopping *(noun)*
shot
shoulder
shouldn't
show
showers
shrouded
shut
shy
sick
signal
silence
silent
silently
simple
since
Sir
sister
site
sixteen
skating
sleepy
slightly
slope
slow
smart
smell *(noun)*
smelled
smoke *(noun)*
snake
snow *(noun)*
snowing
sofa
softly
soil
sold
solid
solution
somebody
sometime
sometimes
son
song
specially
spoke/speaking
spoons
sports

sportsman
star/stars
state
statement
station
steep
step *(noun)*
stepping
stole/stealing
store/stores
straight
stranger *(comparative adj.)*
student/students
studied/study
style
subjects
substance/substances
successful
such
suggestion
suit *(clothes)*
summary
sunlight
super
Superman
supermarket
supper
surely
surface
surprise *(noun)*
surroundings
swam
swayed
sweat
sweater
Sweden
sweeping
sweet/sweets
switch *(noun)*
switched/switch
Switzerland
sword
symptoms
Taiwan
taught/teach
tea
teacher
team
tea-time
telephone
temper
territory/territories
terror
thankful
themselves
therefore
they're
thick
thin
thirdly
thirty
those
thousand
throat
thunder
tickets
tidy
tightly
timber
timid
tiredly
title
tombs
tonight
touched/touch
tourists
toward
town
toys
train/training
translucent
transparent
transport
trap/traps
treasure *(noun)*
trick *(verb)*
trip *(noun)*
twelve
twenty
umbrella
unable
uncomfortable
understood/understand
unforgettable

unfortunately
uninteresting
university
unluckily
unusual
unusually
upon
upset
usually
uttered
valuable
vanished
vase
view
violent
visitor
volleyball
wages
waiter
warm *(adjective)*
warmed
warned/warn
wash
waste
watch *(timepiece)*
watched
weak/weaker
weakness
wealth
web/webs
we'd
weekend
we'll
well-known
we're
what's
whom
whose
widow
wild
windy
wine
winning
winter
without
wolf
won
wondered/wonder
wonderful
won't
wood/woods
wooden
word
worked/work
workers
worn
wouldn't
wrote/write
yell
yet
you'll
young/younger
youth
zoo

APPENDIX 27 (continued)

Sample coded composition

Composition 25

4-last 6-Sunday 7-afternoon , 2-when 0-I 0-was 6-buying 6-clothes 3-with 1-my 2-friend , 2-suddenly 4-an 3-old 3-man 2-came 9-toward 1-me 0-and 1-said ,

9-excuse 1-me 9-miss . 0-I 0-am 0-a 3-magician . 0-I 3-can 4-give 1-you 3-three 2-wishes , 1-but 1-you 1-have 0-to 4-do 2-one 5-special 5-condition 3-for 1-me . 3-can 1-you ? 0-the 3-old 3-man 1-said 9-surely .

1-you 1-are 0-a 3-magician . 1-you 3-can 4-give 1-me 3-three 2-wishes . 0-I 2-don't 4-believe 1-it . 4-if 1-you 1-are 0-a 3-magician , 0-I 2-want 0-to 1-have 3-many 5-beautiful 6-clothes , 0-a 1-very 4-big 1-house 0-and 0-a 4-lot 1-of 4-money , 0-I 1-said 7-happily .

8-sure . 1-you 3-can 3-get 3-all 4-things 1-but 1-you 4-will 3-get 3-no 2-friends 4-if 1-you 2-want 7-these 4-things , 1-he 1-said .

6-OK , 0-I 1-said .

5-now 1-you 1-go 4-home . 0-the 1-house 4-will 6-much 4-bigger 8-than 6-before . 1-in 5-your 1-house 2-there 4-will 4-be 3-many 5-beautiful 6-clothes 0-and 0-a 4-lot 1-of 4-money 1-in 5-your 9-drawer , 0-the 3-magician 1-said .

0-I 1-went 4-home 4-quickly 0-and 0-I 1-saw 3-many 5-beautiful 4-things 1-but 3-all 1-of 1-my 2-friends 1-had 7-gone 2-out . 7-although 0-the 3-three 2-wishes 1-had 2-come 4-true 1-but 0-I 0-was 1-very 8-unhappy 2-because 0-I 6-haven't 5-any 2-friends .

APPENDIX 28

Comparison of lower- and higher-level non reading-scheme classes: school 4

Table 9.2 ***Comparison of lower- and higher-level non reading-scheme classes: school 4***

	Lower-level control class: N = 34		Higher-level control class: N = 40			
	Mean	Standard Deviation	Mean	Standard Deviation	T-value	Sig. Level
overall quality	5.94	2.43	8.63	2.13	5.053	.000
grammatical complexity	5.32	1.60	7.88	2.37	5.307	.000
grammatical accuracy	4.59	1.70	7.15	2.10	5.678	.000
vocabulary range	6.91	2.09	8.73	1.96	3.842	.000
*spelling**	12*		14*		1.90**	.057
punct. & paragraphing	6.35	1.84	8.03	1.79	3.953	.000
coherence and flow	5.71	1.69	8.48	2.63	5.272	.000
number of words	235.5	56.58	267.0	58.27	2.348	.022
words per sentence***	8.62	1.62	9.22	1.97	1.428	.158
words per T-unit****	7.72	1.03	7.93	1.14	0.850	.398
N subordinate clauses	12.65	5.46	15.05	6.11	1.768	.081
N subordinate clauses per 100 words	5.41	1.80	5.69	1.93	.643	.522
N error-free T-units	7.47	5.97	14.80	6.72	4.921	.000
N error-free T-units per 100 words	3.04	2.21	5.57	2.06	5.083	.000
N words contained in error-free T-units	46.38	39.61	101.10	40.39	5.859	.000
N words in error-free T-units per 100 words	18.88	14.55	38.28	13.26	5.997	.000
N types (VocabProfile)	103.47	19.29	119.88	20.75	3.499	.001
% text in first 500 words	84.20%		82.92%		not sig. tested	
% text in second 500 words	9.04%		9.59%		not sig. tested	
% text in second 1000 words	5.56%		6.22%		not sig. tested	
lexical originality quotient	2.8		3.7		not sig. tested	

Range of possible scores for rater-judged constructs = 3 to 18
**Medians are reported for spelling as the distribution was not normal*
*** This is the Z value for a Mann-Whitney U test*
****Narrative text only*
*****Excluding clauseless expressions*