Studies in Language Testing 32

Components of L2 Reading

Linguistic and processing factors in the reading test performances of Japanese EFL learners

Toshihiko Shiotsu

Series Editors
Michael Milanovic
and Cyril J Weir

UNIVERSITY of CAMBRIDGE
ESOL Examinations

CAMBRIDGE
UNIVERSITY PRESS

Components of L2 Reading

Linguistic and processing factors in the reading test performances of Japanese EFL Learners

Also in this series:

Fairness and Validation in Language Assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
Antony John Kunnan

Issues in Computer-adaptive Testing of Reading Proficiency
Micheline Chalhoub-Deville

Experimenting with Uncertainty: Essays in honour of Alan Davies
Edited by A. Brown, C. Elder, N. Iwashita, E. Grove, K. Hill, T. Lumley, K. O'Loughlin, T. McNamara

An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes
Edited by Cyril J. Weir, Yang Huizhong, Jin Yan

The Equivalence of Direct and Semi-direct Speaking Tests
Kieran O'Loughlin

A Qualitative Approach to the Validation of Oral Language Tests
Anne Lazaraton

Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002
Edited by Cyril J. Weir and Michael Milanovic

A Modular Approach to Testing English Language Skills: The development of the Certificates in English Language Skills (CELS) examination
Roger Hawkey

Issues in Testing Business English: The revision of the Cambridge Business English Certificates
Barry O'Sullivan

European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001
Edited by Cyril J. Weir and Michael Milanovic

IELTS Collected Papers: Research in speaking and writing assessment
Edited by Lynda Taylor and Peter Falvey

Testing the Spoken English of Young Norwegians: A study of testing validity and the role of 'smallwords' in contributing to pupils' fluency
Angela Hasselgreen

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution
Alan Davies

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J. Weir

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading

Linguistic and processing factors in the reading test performances of Japanese EFL learners

Toshihiko Shiotsu

Series Editors
Michael Milanovic
and Cyril J Weir

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

www.cambridge.org
Information on this title: www.cambridge.org/9780521157278

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Shiotsu, Toshihiko.
Components of L2 reading : linguistic and processing factors in the reading test performances of Japanese EFL learners / Toshihiko Shiotsu.
p. cm
Includes bibliographical references and index.
1. English language—Study and teaching (Secondary)—Japanese speakers. 2. Second language acquisition—Study and teaching. 3. Reading comprehension—Ability testing. 4. English language —Japan. I. Title.

PE1068.J3S48 2010
428.0071'2956--dc22
2010007519

ISBN 978–0–521–15727–8 Paperback

For my parents, Masato and Kiyoko Shiotsu

Contents

Acknowledgements

The preparation of this volume and the doctoral research on which it is based were possible owing to support from a number of individuals to whom I would like to express my gratitude.

I am very grateful to Professor Cyril Weir for his dependable guidance on my PhD project as my supervisor at the University of Reading and for his patience and kind support throughout the preparation of this volume as a series editor. At Reading, I was also fortunate to have Professor Barry O'Sullivan as my final PhD supervisor, whose friendly and effective support enabled me to complete the thesis on schedule. Without Cyril and Barry's help and kind recommendations, this publication would not have been here. In addition, I benefited much from stimulating and critical advice at the planning stage of the project from my initial PhD supervisor Don Porter and from the constructive comments and criticisms on my thesis by Sandy Urquhart and Mike Garman at the viva voce examination.

This study would not have been possible without the participation of the hundreds of EFL students as test takers at universities and colleges in Japan, and I am grateful to each of them for their time and effort.

Many others contributed to what came to be this volume, whether through assisting with data collection/item validation, providing bibliographic assistance, sending their research instruments for my use, or helping with my PhD application. I wish to thank: Alan Tonkyn, Amos Paran, Amos van Gelderen, Bill Lee, Catherine Walter, Diane and Norbert Schmitt, Hisako Uematsu, Jeong-Won Lee, Jodes Romero, Junko Yamashita, Kenji and Fukiko Numoto, Margot Haynes, Mariko Osaka, Mayumi Shiotsu, Motoyoshi Nagata, Paddy Harben, Rob Schoonen, Rory Britto, Sachi Yamauchi, Shigenobu Takatsuka, Shizuka Tetsuhito, Shoichi Tanaka, Tadamitsu Kamimoto, Takeshi Furukawa, Toshihiro Shimizu, Yubun Suzuki, Yuichi Yamasaki, Yusuke Ando, and nine anonymous ELT experts.

I also extend my thanks to Nagasaki Wesleyan Junior College and Kurume University for their generous support on my PhD work and to everyone there who helped me obtain the study leaves that allowed me to concentrate on the project. I wish to acknowledge the Ishibashi Foundation too for partially funding the study.

For the production of this volume, Lynda Taylor has helped me to clarify and improve the details of my work and to co-ordinate the necessary elements for this edition as the managing editor, and I have benefited much from her assistance.

Finally, my sincere *arigato* to my wife Mariko for all her help, patience, and cheers throughout the project.

The publishers are grateful to the copyright holders for permission to use the copyright materials reproduced in this book. Kansai University Press for Tables 92 and 93 from *New Horizons in Computerized Testing of Reading*, 2004; UCLES and Cambridge University Press for Table 41 from *An Investigation into the Comparability of Two Tests of English as a Foreign Language*, by Lyle Bachman, Fred Davidson, Katherine Ryan and Inn-Chull Choi, 1995.

Series Editors' note

Compared to the extensive research base on L1 reading available in cognitive psychology, L2 reading research in the Applied Linguistics field has lacked a solid empirical research base and as a result the predictability of L2 reading behaviour is seriously constrained. A more principled effort is needed to generate empirical evidence on the nature of L2 reading which draws on the rigorous investigation of the factors considered by cognitive psychologists to be significant for L1 reading comprehension.

This volume helps to establish such a research tradition. It makes good use of the methodology established in cognitively oriented L1 research and investigates whether those factors considered to influence reading comprehension skills among L1 readers do so among L2 readers. In the componential approach to modelling L1 reading ability a number of contributory factors have been empirically validated. However, research on their relative contribution either singly or in combination to explaining performance on second language reading tests is limited.

It emerges from Shiotsu's thorough overview of the L1 component skills literature that such variables as vocabulary knowledge, word recognition efficiency, phonological awareness, and working memory span account for a relatively larger amount of the variance in reading comprehension than do the others. Shiotsu investigates the contribution of these variables in L2 reading through theoretical and empirical scrutiny. In addition he looks at the importance of the L2-specific variable of grammar knowledge, whose relationship with L2 reading seems to merit more attention than it has been given so far in the reading research literature.

In Chapter 2 Shiotsu examines how reading has been researched and understood and details some of the knowledge and processing subskills which have been demonstrated to be important for reading abilities. The research questions below emerge from this review in relation to the Japanese context the author is working in:

> RQ1: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills – L2 vocabulary breadth, L2 word recognition efficiency, L2 working memory span, L2 syntactic knowledge, and language-independent metacognitive knowledge about the text and reading – account for the individual differences in L2 careful passage reading comprehension ability of Japanese EFL learners?

RQ2: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in L2 careful passage reading comprehension ability of the Japanese EFL learners as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

RQ3: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners?

RQ4: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

Chapter 3 considers the research methodology with an emphasis on the practicality and feasibility of the research variables and their instrumentation.

Chapters 4 and 5 describe a total of six preliminary studies which are used to assess the potential values of the initially selected variables for the subsequent main study and to evaluate the instruments and their content to make necessary refinements.

Chapter 6 reports on the nature of the main study providing detailed background and basic descriptive data. Chapters 7 and 8 report the main findings based on two different methods of analysis. A traditional method of analysis, conventional regression, is used to answer the research questions in Chapter 7, but in Chapter 8 Structural Equation Modelling is used to help consider the data from a slightly different perspective.

In Chapter 9 the research questions are revisited and the findings are discussed with reference to the existing research base reported in the literature review; the implications of the findings are critically evaluated, and recommendations for future research are made.

In contrast with previous findings in the literature in this field, this study offers support for the relative superiority of syntactic knowledge over vocabulary knowledge in predicting text reading comprehension test performance.

The volume offers the reader a valuable perspective on L2 reading ability, test score predictability, the potential effects of unequal instrument reliabilities in this research area, and test content validation. It provides rigorous empirical support for a number of the themes emerging from an earlier volume in the series – *Examining Reading* by Khalifa and Weir (2009). In particular it provides evidence for the view that linguistic comprehension is best divided into careful text processing ability and the speed of comprehension. It makes a valuable contribution to the methodology of using test data to enhance our knowledge of reading proficiency. As such it testifies to the centrality of language testing for developing our understanding of language proficiency and offers serious lessons to second language acquisition (SLA) researchers in the rigorous use of complex statistical procedures.

Cyril J Weir and Michael Milanovic
Cambridge – November 2009

Abbreviations

AGFI	adjusted goodness-of-fit index
AMOS	Analysis of Moment Structures
ANOVA	analysis of variance
AWL	Academic Word List
CALL	computer-assisted language learning
CAT	computer-adaptive test/ing
CB	computer-based
CBT	computer-based test
CFA	confirmatory factor analysis
CFI	comparative fit index
CLA	Communicative Language Ability
COPS	Clustered Objective Probability Scoring
df	degree of freedom
DSL	Dutch as a second language
EAP	English for academic purposes
EFA	exploratory factor analysis
EFL	English as a foreign language
ELT	English language teaching
ESL	English as a second language
ETS	Educational Testing Service
FCE	First Certificate in English
GFI	goodness-of-fit index
IELTS	International English Language Testing System
L1	native language
L2	second language
LCD	liquid crystal display
Max	maximum
MC	multiple choice
MCQ	multiple-choice question
Min	minimum
MS	mean square
msec	millisecond
NFI	Bentler-Bonnet normed fit index
NS	native speaker
PB	paper-based
PBT	paper-based test

PCA	principal component analysis
RMSEA	root mean square error of approximation
RST	Reading Span Test
SA	short answer
SAQ	short-answer question
SD	standard deviation
SE	standard error
SEM	structural equation modelling
SFL	Spanish as a foreign language
SLA	second language acquisition
SPSS	Statistical Package for Social Sciences
SS	sum of squares
TEEP	Test of English for Educational Purposes
TOEFL	Test of English as a Foreign Language
VLT	Vocabulary Levels Test

1 Introduction

The practice of developing and testing second language (L2) reading must follow as clear an understanding of this skill as possible, and such an understanding must be a result of rigorous research, both theoretical and empirical. Comparisons have been made between applied linguistic research on L2 reading and cognitive psychological research on L1 reading, both of which directly or indirectly contribute to our understanding of the nature and development of reading. Compared to cognitive psychology's extensive and cumulative research base resulting from successions of empirical data collection and theory and model generation, not only for description but for explanation and prediction of reading behaviour, L2 reading research effort within the applied linguistics community has been said to suffer from lack of a solid empirical research base (Bernhardt 1991a) and a resultant relative weakness in generalizability and predictability of L2 reading behaviour. It is therefore imperative that more principled effort be made to accumulate empirical evidence on the nature of L2 reading. While L2 reading research has at times drawn insights from the L1 reading literature, the issues it has addressed and emphasised have tended to be distinct from the ones researched through typical cognitive psychological approaches.

Two questions which characterise L2 reading research have been whether L2 reading difficulty arises from incomplete L2 knowledge or insufficient L1 literacy (Alderson 1984) and whether there is a linguistic threshold level which the L2 reader must achieve in order for his or her L1 literacy skills to be positively transferred to the task of L2 reading comprehension (Clarke 1978). These questions obviously relate to the facts that most individuals learning to read in L2 do not have the kind of sophistication in linguistic competence shared by the majority of the native speakers of the target language and that these L2 readers already have varying levels of literacy skills in their own L1.

While these features specific to L2 reading must be sufficiently highlighted and reflected in L2 reading research designs (Alderson 1984, Koda 1994), effort also seems to be justified which examines the factors considered to be significant for L1 reading comprehension. The advantages of such an effort would be the availability of methodology established in L1 research and its extensive body of research evidence, which enables us to compare L2 reading data with their L1 counterparts and to determine whether those factors considered to influence reading comprehension skills among L1 readers are also importantly related to those among L2 readers. This type of research has the potential of bringing closer the two fields of cognitive psychology and applied linguistics, since findings resulting from a common ground in terms of research design, variables, and instruments could be more easily compared or built into a common research base.

A review of L1 and L2 reading literature for methodological compatibility identifies one line of L1 reading research which has aimed to identify components of reading skills and possible sources of individual differences in reading abilities. It emerges from the overview of the L1 component skills literature that such variables as vocabulary knowledge, word recognition efficiency, phonological awareness, and working memory span account for a relatively larger amount of the variance in reading comprehension than do the others (Cunningham, Stanovich and Wilson 1990, Daneman 1991, Just and Carpenter 1992, Stanovich, Cunningham and Cramer 1984). It would seem most meaningful to subject such variables in L2 readers to theoretical and empirical scrutiny along with the L2-specific variable of grammar knowledge, whose relationship with L2 reading seems to deserve more attention than has been given so far (Urquhart and Weir 1998).

The present study initially surveys in Chapter 2 how reading has been researched and understood and details some of the knowledge areas and processing subskills which the researchers suggest as important for reading abilities. Discussions of these knowledge and skill areas will lead to the initial attempt to form a set of research questions for the present study.

Chapter 3 considers the research methodology that, if adequately employed, would answer the research questions initially posed. The emphasis in Chapter 3 will be on the practicality and feasibility of the research variables and their instrumentation. Taking account of the theoretical significance and the practicality constraints, revised research questions will be stated in this chapter.

Chapters 4 and 5 will describe a total of six preliminary studies to simultaneously evaluate the potential values of the initially selected variables for the subsequent main study and to evaluate the instruments and their content to make necessary refinements. The two chapters represent separate phases of the research programme each covering three studies.

Reports on the actual main study will begin in Chapter 6, which will present detailed background and basic descriptive data.

Chapters 7 and 8 will report on the main findings based on two different methods of analysis. The results of a series of multiple regression analyses to answer the research questions will be presented in Chapter 7, but an additional analysis which explores latent variables and helps consider the data from a slightly different perspective will be reported in Chapter 8.

What the data have indicated in Chapters 7 and 8 will be examined in Chapter 9, in which the research questions will be answered, the findings discussed in reference to the related research, the implications of the findings evaluated, and recommendations for future research proposed.

2 Literature review

Chapter overview

Research on reading in L2 requires an overview of how reading has been understood in the reading literature. This section will begin by briefly discussing how reading may be defined, followed by a developmental overview of the interactive process models of reading to capture some significant elements of reading which emerged from systematic L1 reading research. The focus will then shift to the difficulties of the process model approach in general and to the alternative approach to reading research, the component skills approach. This will lead to the discussion of the components and issues addressed in the L2 reading literature and the identification of some research variables that have been suggested as worthy of further investigation. The section will conclude by specifying the research questions emerging as a result of the literature review and theoretical and practical considerations.

Definitions of reading

In a well-quoted synthesis of previous reading research, Grabe (1991) avoided simple definitions of reading on the grounds that they 'typically misrepresent complex cognitive processes such as reading' (1991:378), underscoring the difficulty in defining reading. However, it is notable that, by way of rationalising his reluctance to provide a simple definition, Grabe actually states his view of reading as a 'complex cognitive process'. While this seems to be the position accepted by both the L1 and L2 reading researchers, Alderson and Urquhart's (1984) caution against equating reading with general cognitive activity is also commonsensical. What differentiates reading from other cognitive activities such as reasoning or mathematical calculation is the involvement of written language, or the text, faced by the reader. We may thus attempt to formulate our initial definition of reading as 'a complex cognitive process the individual is involved in while engaged with a written text'. This is still a very broad definition, but it seems rather premature to narrow our focus to which to commit ourselves when dealing with something as multi-faceted as reading. The divergence of views on reading even within a single academic discipline of cognitive psychology requires at least a brief survey of how reading has been understood. The following section reviews some previous work on reading which seems to be relevant for our attempt to formulate a balanced yet more focused and research-specific definition of reading.

Theories and models of reading

Processes of reading

Current efforts among reading researchers to theorise and model reading date back to at least the late 1960s, when Goodman (1967) proposed the notion of reading as a 'psycholinguistic guessing game'. He argued against the then prevalent view within the teaching profession which saw reading as a precise process involving exact, detailed, sequential perception and identification of letters, words, etc. and proposed an alternative view in which the reader is conceptualised as constantly making predictions or hypotheses on the linguistic or propositional contents of the text, relying heavily on the prior linguistic and non-linguistic knowledge available and cyclically confirming, disconfirming, or reforming the predictions made as the textual information is sampled rather than thoroughly processed. This significant role assigned to the reader's knowledge and guessing (or 'higher-level' processes) as the guiding force in reading, in contrast with the relative lack of emphasis on the role of the text (at the 'bottom'), has come to make his and a similar view (Smith 1971) the top-down model of reading. The model has earned popularity in the reading literature and pedagogy and exerted influence on the L2 reading material and curriculum (see Paran 1996), although its general credibility, and particularly its assertion that the more skilled readers guess more, have been significantly weakened as a result of a vast volume of empirical counter-evidence (cf. Stanovich 1991) based on data on readers' eye movements and vision (e.g., Just and Carpenter 1980, Rayner and Pollatsek 1989) and on the effects of context on word recognition (e.g., West and Stanovich 1978). Grabe and Stoller also state that 'few reading researchers actually support strong top-down views' (2002:32). Reading is certainly an active process; however, it does not appear to be entirely a psycholinguistic guessing game.

Conceptualisations at the other end of the top-down bottom-up continuum were also being developed at about the same time in the history of reading theory evolution. Gough (1972), among others, worked out a strictly sequential, bottom-up model, very much the type Goodman was trying to refute. Not only was Gough's model overshadowed by the popularity of the top-down approach to reading, its strictly serial stage-by-stage concept also suffered inconsistencies with experimental findings (Reicher 1969, Rumelhart 1977). However, Gough's model is also considered to have contributed more to the field of reading research than did its top-down competitor. In fact, Rayner and Pollatsek (1989) evaluate Goodman's and Smith's top-down model as 'so vague as to be untestable' but Gough's as 'very clear in what he thought was happening during reading' (1989:467) and gave the latter credit for making explicit testable predictions and stimulating a great deal of reading research. What the discussion on Gough's model made clear is that reading is not simply an act of serial bottom-up decoding, although one cannot initiate the act of reading without some sort of decoding of print.

One of the important results of the subsequent research effort was the emergence of the interactive models of reading, which took into account the claims of the two opposite approaches mentioned so far (although much less of the top-down models). Rumelhart introduced an interactive parallel processing model (1977) based on the idea and previous research which suggested that our perception of input at one level, be it letter, word, or syntax, is facilitated by the context in which we encounter it (e.g., the surrounding letters or the syntactic or semantic constraints). This interaction among various levels of knowledge has become a standard feature of subsequent models of the reading process that are still widely cited (Just and Carpenter 1980, Rayner and Pollatsek 1989, Stanovich 1980), though these models also maintained emphasis on the extraction of information through perceptual processes. Reading has thus been viewed as an interactive process in which various types of knowledge and textual information contribute to successful identification of letters, words, syntactic functions of words, and larger units of meaning.

The concept of interaction was also an integral part of the popularised schema-theoretic view of reading (Anderson and Pearson 1984, Carrell and Eisterhold 1983), which argued that the reader's schemata, or structured world/background knowledge, play a significant role when trying to make sense of the information presented in the text. In this framework, the phenomenon of interest seems to be general language comprehension, which was considered a function of the interaction between old knowledge stored in the comprehender's memory and new information presented via a linguistic message. Grabe (1991) distinguished between this type of reader-text interaction approach and the interaction-of-component-knowledge/skills approach previously described. He expressed scepticism on the former as it is difficult to demonstrate experimentally how prior knowledge is called up and used. Clapham (1996) emphasised the difficulty in assessing such knowledge as well. The role of prior knowledge must surely remain in our wider conceptualisation of reading, and the ultimate goal of reading is usually more than simply decoding the written symbols; however, the current interactive process approaches to reading stress the efficiency of bottom-up processes and integration of information within and between the memory structures (Just and Carpenter 1980, Rayner and Pollatsek 1989) and assign no central role to schemata.

The discussions so far have presented a view of reading as an interactive cognitive process involving various levels and types of reader knowledge for efficient processing of visually presented text. The models that represent such an understanding of reading have been identified as process models or stage models, since they typically identify distinctive stages of information processing which are interrelated with other stages to attain a certain goal, which is usually comprehension of the text. While such process models derive from careful synthesis of empirical evidence on L1 reading behaviour, no one is exhaustive. Each model mirrors the researchers' orientation towards their interests and research focus with its possible consequence being an emergence of a range of essentially interactive process models differing only in emphasis (Urquhart and Weir 1998). Rayner and Pollatsek (1989) warn when introducing their own model that it reflects their 'theoretical biases' and how they 'interpret the . . . evidence on the reading process' (1989:471). Such a cautionary note adds to our doubt with regard to the general usefulness of process models and discourages researchers from working out another variant of an interactive process model, at least until significant research findings have necessitated a major revision to the available models. Another difficulty seems to be with the lack of consideration of individual differences (see, however, Stanovich 1980) and differential purposes of reading within the same individuals. Modelling the cognitive processes of even a certain type of reading by a certain type of reader requires a synthesis of an enormous volume of empirical evidence. These difficulties, coupled with the extremely diverse background of L2 readers, may explain the scarcity of process models of general L2 reading (see Segalowitz 1986 and Paran 1994 for exceptions). Although these difficulties with process models are not easily surmountable, L2 reading research does need a framework in which to structure more empirical research for the purpose of characterising L2 reading. Thus, an alternative method of describing, explaining, and predicting reading comprehension ability will be discussed below, after a consideration of the different purposes of reading as they are related to different types of reading.

Purposes and types of reading

The previous section has alluded to the failure of the process models of reading to take account of the various types of reading associated with different purposes. Urquhart and Weir (1998) differentiate among five types of reading: Scanning, Search Reading, Skimming, Careful Reading, and Browsing. Although one might attempt to organise them in terms of the complexity of the cognitive operations required, one can more clearly distinguish them in terms of the reader's purposes. Browsing is contrasted with the rest by its lack of clear purpose. In that sense, it may be considered 'quasi-reading'. The first three of Scanning, Search Reading, and Skimming are expeditious reading, which means that priority is given to the speed of fulfilling a specific purpose at the expense of thorough text processing. The purpose of Scanning is to simply locate a set of textual strings which satisfy a certain search condition. Search Reading is for locating the answers to a pre-determined question without having to arrive at anything like a superordinate main idea of an extended discourse such as paragraphs or passages. Skimming is for gaining such a main idea, and contrary to Search Reading, without any prior clues as to the topic of the text. Careful Reading is different from any expeditious reading since it aims to deal with both the details and the general ideas, which must be constructed through comprehension and synthesis of the details, while the processing speed is secondary to digesting the majority of the text.

The tentative definition of reading stated earlier (see the section 'Definitions of reading') should now incorporate this purposefulness in reading: reading is a cognitive process the individual is engaged in with a written text for one or more specific goals such as quickly identifying particular facts stated in it or constructing a thorough semantic representation of most of it.

Clearly, no type of reading is superior to the others in any absolute sense. Nevertheless, any research which claims to deal with reading will have to specify what type or types it means by reading (cf. Taillefer 1996, Weir, Yang and Jin 2000). The primary interest of the present study is in Careful Reading at the passage level, which seems to be required in many academic, professional, and personal functions.

Components of reading

A somewhat different line of reading research, which perhaps complements the process model approach rather than competes with it, is what is known as the component skills approach or componential approach. Research in this category tries to identify the components of reading or to model the ability of reading through such components. Another goal of this approach is to account for the individual and developmental differences in reading performance in terms of the differences in specific component processes or knowledge sources. Carr and Levy (1990a) state:

> Many investigators believe that the kind of full characterization that results from component skills analysis is the only way to get an accurate picture of reading ability, how it changes developmentally, and what creates individual differences among readers who are otherwise roughly the same in developmental level (1990a:xi).

According to Hoover and Tunmer (1993), 'components' refer to some 'theoretically distinct and empirically isolable constituents' of reading (1993:4).

L1 componential research has argued, for instance, for separate word recognition and linguistic comprehension components (Hoover and Tunmer 1993). Their 'Simple View' of reading finds support in the L1 developmental or dyslexic situations, in which the reader has the linguistic competence to achieve listening comprehension but lacks the skill for visual word identification, or in hyperlexic situations, in which the reader can successfully decode the printed words but has deficiency in making sense of the results of this decoding. Urquhart and Weir's (1998) reaction to Hoover and Tunmer's Simple View points to the difficulty in identifying mutually exclusive components (word recognition subsumes access to mental lexicon, which is a part of linguistic comprehension), but the Simple View serves as a useful point of reference for the L2 componential approach and a reminder of the necessity to minimise the number of component distinctions to only the most meaningful ones.

Earlier L1 componential work with adult readers is found in such published studies as Jackson and McClelland (1979), Palmer, MacLeod, Hunt and Davidson (1985), Baddeley, Logie, Nimmo-Smith and Brereton (1985), Dixon, LeFevre and Twilley (1988) and Cunningham et al (1990).

Jackson and McClelland's (1979) research examined the correlates of the L1 reading speed of a group of university undergraduates and indicated that the largest proportion of the variance in their reading speed was accounted for by listening comprehension performance and the second and third largest proportions by their performance on letter-name matching and homonym matching reaction time measures respectively. The results have led the researchers to conclude that reading speed is dependent on the two main factors of general language comprehension ability and 'speed of accessing overlearned memory codes for visually presented letters' (Jackson and McClelland 1979:151).

A group of university undergraduates participating in the study by Palmer et al (1985) responded to a set of tasks yielding a total of 28 measures, which included reading comprehension, reading speed, listening comprehension, and various speeded measures of letter-, word-, and sentence-processing. Their correlational results indicated that reading comprehension was predicted very well by listening comprehension and in fact much better than by reading speed, which in turn correlated less well with listening comprehension. Their measures of visual stimulus-matching speed and speeded sentence-verification measures correlated well among each other but less well and differentially with the reading and listening comprehension and reading speed measures. They concluded from these and other results from a series of factor analyses that there is a modality-independent verbal comprehension component that can be dissociated from the reading speed component.

Baddeley et al's (1985) componential analysis with a group of adult L1 readers showed their lexical decision speed measure and sentence-span working memory measure to be accounting for the largest and roughly equivalent proportions of the variance in reading comprehension performance and their vocabulary measure to be explaining a smaller but significant proportion as well. Their second experiment, which did not include the lexical decision speed measure, also showed their working memory and vocabulary measures to be uniquely accounting for a significant portion of the reading variance. From these results, Baddeley et al claimed that vocabulary, lexical access speed, and some form of working memory are separable and important components of fluent reading.

Not only reading comprehension and reading rate but also the ability to make plausible inferences based on world knowledge were measured in Dixon et al's (1988) study with a group of university undergraduates reading in L1. Variance in their reading comprehension measure was best accounted for by the vocabulary score while sentence-span working memory and the knowledge of multiple meanings of specific words also uniquely explained additional variance. Reading rate and the inferencing ability measure showed a somewhat different pattern though vocabulary was again the strongest predictor for each criterion. These data allowed Dixon et al to claim multidimensionality of reading proficiency and importance of word knowledge even with skilled, mature readers.

Cunningham et al (1990) were interested in whether visual word decoding is a separate skill which contributes to explaining the individual differences in reading comprehension among their university undergraduates. Vocabulary size, listening comprehension, word and pseudoword reaction times, and working memory span were among the variables correlated most strongly with reading comprehension and on which good and poor readers differed significantly. A series of their multiple regression analyses, with reading comprehension as criterion, consistently revealed a significant effect of word decoding, measured through pseudoword reaction time, and their confirmatory factor analysis supported a 3-factor model yielding what they named the global verbal comprehension, word recognition, and reading comprehension factors.

L1 componential analyses referred to so far suggest several variables as potentially important for reading abilities among adults. Most of the studies (Baddeley et al 1985, Cunningham et al 1990, Jackson and McClelland 1979) showed the significant effects of some form of efficiency in accessing the lexical and sublexical information stored in long-term memory, which may be loosely termed 'word recognition efficiency'. Fast and accurate word recognition, which was once relegated as secondary to contextual prediction skill, is now considered to be a major determinant of reading success (Stanovich 1991), and the individual differences in the efficiency of this skill should be examined as a potentially important factor in L2 reading as well.

Other variables which predicted reading comprehension across different studies were vocabulary and working memory (Baddeley et al 1985, Cunningham et al 1990, Dixon et al 1988), while listening comprehension predicted reading speed in Jackson and McClelland (1979) and reading comprehension in Palmer et al (1985). Vocabulary knowledge is often claimed as a good predictor of reading skill (Beck and McKeown 1991, Daneman 1991, Laufer 1992a), and improvement in reading comprehension is also attributed to increase in vocabulary knowledge (e.g. Beck, Perfetti and McKeown 1982). The significant contribution of vocabulary knowledge in explaining reading comprehension variance found in the L1 componential analyses above coincides with these claims and encourages an inclusion of vocabulary as part of L2 studies.

Some researchers (Just and Carpenter 1992) also regard working memory as a major determinant of language comprehension skill, and significant correlations have been found between measures of working memory and reading comprehension (Daneman and Carpenter 1980, 1983, Masson and Miller 1983, Turner and Engle 1989). The results of the componential studies summarised earlier corroborate the idea that working memory may be yet another important variable in L2 reading.

While the previous sections have dealt with L1 componential research with adult readers, concepts representative of developmental reading research concerned with children may provide additional perspectives appropriate for the development of L2 reading skills among adults. Levy and Carr (1990) summarise the findings of the componential analyses in their edited collection (Carr and Levy 1990b), most of which examined children reading in their L1, and highlight some of the regularities therein. They first identify deficits in phonological processing as significantly related to reading failure among any population of readers although they add that adequate phonological processing does not guarantee success in reading comprehension. They also acknowledge the contribution of general language skills distinct from word recognition skills but overlapping with working memory processes. Studies reported in Carr and Levy (1990b) thus seem to indicate phonological processing skill as another important variable, while importance of most of the knowledge or subskill areas suggested as related to adult L1 reading success in the previous paragraphs seem to have received concurrent support from this set of research evidence.

Phonological processing skill, or phonological awareness as it is more often called in the reading literature, has been known to predict early L1 reading skill acquisition (Stanovich et al 1984) and has also been a focus of some L1 reading acquisition research which compared it with working memory in terms of predictability of early reading achievement. Leather and Henry (1994), Hansen and Bowey (1994) and Cormier and Dea (1997) were all concerned with phonological awareness and working memory and the relationship of each of the two variables with children's reading development. They all found each of the two measures to correlate significantly with their measures of children's reading achievement and to make a unique contribution in predicting reading achievement.

To summarise the L1 componential studies reviewed thus far, the individual differences in reading ability may be accounted for by individual differences in vocabulary knowledge, word recognition efficiency, phonological awareness, and working memory, which overlapped with general verbal comprehension skill as measured through listening comprehension tasks. An implication of these L1 reading studies for applied linguistics is that it may be meaningful to shed light on the component processes of L2 readers to explore what types of knowledge or skills explain the individual differences in L2 reading ability. It is perhaps a productive practice to consider whether some of the variables which emerged as explaining the L1 reading variance can be theoretically and practically accommodated in a componential analysis of L2 reading. The following section focuses on the efforts from the L2 researchers to characterise the skill of L2 reading and to identify what may be treated as candidate 'components' of L2 reading.

L2 reading research and components of reading

L2 reading research to date has been overwhelmingly componential in nature, be it theoretical or empirical. Early theoretical attempts to identify some key subskills or knowledge areas of L2 reading were later complemented by data-driven research. This section describes the previous L2 research which explored the possible source(s) of individual differences or important component(s) in L2 reading.

Among some early theoretical attempts at modelling EFL (English as a foreign language) reading is Coady's tripartite model (1979), which described the reader's background knowledge, conceptual abilities, and process strategies as interacting with each other to enable comprehension. The developmental dimension of Coady's model depicted the changes in the relative involvement of six process strategies as the EFL learner's skill at reading improves. Coady hypothesised that for beginning readers the relative amount of involvement of the process strategies would be in the order of grapheme-phoneme correspondence, grapheme-morphophoneme correspondence, syllable-morpheme information, syntactic information, lexical meaning, and contextual meaning, while this order would gradually shift over time and reverse as the readers become more advanced. Coady's model and Field's (1985) modification to Coady's model, based on her observation of Chinese ESL (English as a second language) readers, were clearly prepared under the influence of top-down conceptualisations, but their enumeration of six process strategies and the hypothetical differences in the relative amount of their involvement across the proficiency levels seem to have more in common, at least in appearance, with componential models. Coady's and Field's hypotheses were later explored by Chern (1993) in a study with a group of Chinese ESL readers; however, the focus of the study was limited to the use of strategies in dealing with unknown words in the text.

Dichotomising early theoretical positions on the reader factors determining L2 reading success, Alderson (1984) identified Coady (1979) and Jolly (1978) as supporting the view that an L2 reading problem is a reading problem and Yorio (1971) as claiming it is a language problem. To say that it is a reading problem is to attribute L2 reading difficulty to the lack of adequate reading skills in the reader's L1, while to specify it as a language problem is to attribute it to the reader's imperfect L2 knowledge and L1 interference. To explore this question of L2 reading as 'a reading problem or a language problem', Alderson surveyed the previous literature and tentatively reported that it appears to be both and that the evidence is in favour of the view that it is more of a language problem than a reading problem for L2 readers with low level of L2 proficiency.

Of great relevance to this statement are the 'short-circuit' hypothesis (Clarke 1980, also referred to as 'linguistic ceiling' in Clarke 1978) and the notion of a 'threshold level of linguistic competence' (Cummins 1979), both of which Alderson (1984) refers to. These constitute some early attempts to theorise the relationships among L1 reading, L2 reading, and L2 proficiency based on sets of published learner data, and their points can be summarised as follows: Mental operations for reading acquired through the learner's L1 reading experience can be applied to the L2 reading situation (therefore, performance in L1 reading is expected to correlate with that in L2), but this positive transfer of reading skill can occur only after the learner has reached a certain 'threshold level of L2 linguistic competence', without which the L2 linguistic difficulty is so great that it 'short circuits' the system of even those individuals who are good readers in their L1 (thus, no significant correlation is expected between L1 reading and L2 reading among the learners with linguistic competence below the threshold level). Formulation of this explicit hypothesis as well as Alderson's (1984) observation on the unsettling nature of the empirical evidence for this hypothesis have brought the test of linguistic threshold to the foreground of L2 reading research and led to a burgeoning of research activities focusing on the issue (Perkins, Brutten and Pohlmann 1989, Hacquebord 1989 and Bossers 1989, both reported in Bossers 1991, Brisbois 1995, Carrell 1991, Hulstijn 1991, Lee and Schallert 1997, Schoonen, Hulstijn and Bossers 1998, Taillefer 1996, Yamashita 1999). This line of research is also componential in the sense that it tests whether L1 reading and L2 linguistic competence are significant sources of individual differences in L2 reading (i.e., important components of L2 reading) and whether the relative contributions of these factors to L2 reading comprehension performance differ across a certain point (a threshold) in the L2 learners' development of linguistic competence. Brief summaries of such studies will follow.

Perkins et al (1989) correlated the L1 reading and L2 reading comprehension test scores of L1-Japanese EFL readers in three proficiency bands defined by their TOEFL (Test of English as a Foreign Language, ETS) scores. They observed an increase in correlation coefficient across the three proficiency bands. Perkins et al judged the substantial correlation for their highest proficiency group to be 'conclusive evidence that a transfer of reading skills has begun' (1989:8) at this proficiency level. Such results may mean that skills necessary for L1 reading comprehension become an important component of L2 reading comprehension after a certain point in the learners' L2 proficiency development, although little information can be extrapolated on what factors explain the L2 reading variance of the learners at lower proficiency levels.

Hacquebord (1989, as cited in Bossers 1991) examined, in the Dutch as a second language (DSL) context, the L1 and L2 reading comprehension performances and L2 knowledge of L1-Turkish secondary school students who had emigrated to the Netherlands in their early childhood. L2 knowledge as defined by vocabulary test scores correlated more strongly with L2 reading than did L1 reading, and the correlation coefficients diminished when the same group was examined two and a half years later. Hacquebord ascribed the reduction in L1 and L2 reading correlation to L1 loss and described the reduced Dutch reading and knowledge correlation as approaching the pattern obtained for native speaker (NS) readers of Dutch. Whatever the cause of the decline in the correlation coefficients, the L2 knowledge measure was always a stronger predictor of L2 reading comprehension.

Another study with DSL learners (Bossers 1989, reported in Bossers 1991) also revealed a significant contribution of L2 knowledge in explaining the L2 reading comprehension variance. Bossers' L1-Turkish subjects participating in DSL courses at the tertiary level provided data on their L1 and L2 reading skills and L2 knowledge as defined by a composite index of grammar and vocabulary test scores. Both L1 reading and L2 knowledge contributed to explaining the variance in L2 reading, but L2 knowledge emerged as a stronger predictor of L2 reading than L1 reading. When separate analyses were conducted for the two subgroups of lower and higher ability, only L2 knowledge was a significant predictor of the lower group's L2 reading whereas only L1 reading significantly predicted the higher group's L2 reading, offering support for the linguistic threshold hypothesis.

Carrell (1991) analysed the L1 and L2 reading performance of two groups of college students in the USA, L1-Spanish ESL learners and L1-English Spanish as a foreign language (SFL) learners. Carrell's data showed that, for her ESL group, L1 reading was a stronger predictor than L2 proficiency level of the L2 reading comprehension, but for her SFL group, L2 proficiency was a stronger predictor than L1 reading of L2 reading comprehension. Carrell's results may suggest differential relationships among L1 and L2 reading and

L2 proficiency level as a function of the subjects' L1 and L2 or the L2 learning environment (whether L2 is a second or foreign language). Carrell also speculates that the absolute proficiency level of her SFL learners could have been lower than that of her ESL learners and that L2 proficiency level may have been a more significant component for the L2 reading of this lower-proficiency group. Such accounts remain unsubstantiated and are for future research to test empirically.

Proposing the concept of lexical threshold, Laufer (1992b) argued that '(t)he prevalent position nowadays is that for L1 reading strategies to be effective in L2 reading, the reader must pass a threshold of foreign language ability . . . This threshold is, to a large extent, lexical' (1992b:2). She used L1 and L2 reading comprehension and L2 vocabulary size data obtained from L1-Hebrew and L1-Arabic EFL readers at an Israeli university and reported that (1) lexical level correlated more strongly with L2 reading than did L1 reading, (2) the minimal vocabulary level at which there were more 'readers' than 'non-readers' was the 3,000-word level, and (3) the largest inter-group difference in reading comprehension performance was observed between the learners belonging to the 2,000- and 3,000-word levels. Laufer interpreted these results as evidence that 'the turning point of vocabulary size for reading comprehension is 3,000 word families' (1992b:4). Contrary to what the threshold hypothesis would predict, L1 and L2 reading comprehension measures correlated most strongly among the learners at or below 2,000-word level. Therefore, the results confirmed the significance of L2 lexical knowledge but not of L1 reading.

Brisbois (1995) obtained data on L1 and L2 reading as well as L2 vocabulary and grammar knowledge from beginning-level and upper-level adult French as a foreign language students. She reported that: (1) her beginner group's L2 reading correlated significantly with their L1 reading performances, L2 vocabulary and L2 grammar knowledge; (2) for her upper-level group, L2 reading only correlated significantly with one measure of L1 reading; (3) correlations between L2 vocabulary and L2 grammar were significant for both the beginner group and the upper-level group; (4) L1 reading, L2 vocabulary and L2 grammar collectively accounted for her beginners' L2 French reading variance, and they were also significant predictors of one measure of her upper-level learners' French reading. Thus, both the L1 literacy component and the L2 knowledge component appear to have been related to L2 reading for Brisbois's learners of French.

Research by Bernhardt and Kamil (1995) involved L1-English learners of Spanish at three instructional levels and led them to a general conclusion that L1 reading variables account for between 10 and 16% of the L2 reading variance whereas L2 proficiency accounts for 30 to 38%. Bernhardt and Kamil's data corroborated those studies indicating the superiority of L2 language factor in predicting L2 reading performance and the somewhat less though still significant L1 reading factor implicated in L2 reading. Bernhardt and Kamil also raise the question: what accounts for the other 35–50% variance in L2 reading comprehension scores? The empirical research is beginning to characterise L2 reading as better predicted by the learners' L2 knowledge/proficiency level in general and especially until a certain level of L2 is acquired. It also supports the significance of general reading skills, manifested by the learners' L1 reading performance, particularly after a certain L2 level has been achieved. Along with further efforts to rigorously research this issue through methodological refinement, concurrent work for uncovering the remaining variance by considering new dimensions or components of L2 reading should be meaningful.

Lee and Schallert's (1997) study attempted to verify the linguistic threshold hypothesis with a large sample of EFL learners in the Korean secondary school context. Their subjects provided L1 and L2 reading comprehension and L2 knowledge data. General correlational data suggested a stronger L2 reading-knowledge relationship than L1–L2 reading relationship. Their multiple regression data showed that both L2 knowledge and L1 reading measures made significant contributions in explaining the L2 reading variance but a greater proportion was explained by L2 knowledge than L1 reading scores. Lee and Schallert also divided the entire sample into different proficiency bands based on their L2 knowledge scores and found a steady increase in L1–L2 reading correlation beyond a certain division of the bands, which they interpreted as evidence of a threshold level of L2 proficiency for their Korean students. The findings from this large-scale study add support for the significance, in L2 reading comprehension performance, of the L2 knowledge component across the entire L2 knowledge continuum and the involvement of language-independent reading component beyond a certain linguistic knowledge threshold.

Work in the Netherlands by Schoonen et al (1998) explored the roles of metacognitive knowledge, defined as 'knowledge about one's cognition and about regulation of that cognition', and language-specific vocabulary knowledge in L1 and L2 reading comprehension performance of L1-Dutch EFL learners at three different grade levels of primary education. In L1 reading, metacognitive knowledge and L1 vocabulary scores accounted for two thirds of the reading comprehension variance, and the proportion of the reading variance explained by the former increased as a function of the learners' grade level. In L2 reading, L2 vocabulary score was the better predictor for both of the top two grades, but metacognitive knowledge added to the amount of variance explained. The linguistic knowledge factor is evident from the strong predictive power of the vocabulary score; however, the increase in the involvement of language-independent metacognitive knowledge for the higher grade subjects has also been observed, indirectly supporting a developmental shift in the relative importance of language-specific and language-independent components in L2 reading. Of interest is Schoonen et al's claim that their L1 and L2 vocabulary tests were not entirely language-specific measures (although mainly so) since L1 and L2 vocabulary measures appeared to be strongly correlated and explained the reading variance across language. Their subsequent caution against trying to discover an exclusively language-specific or more general reading component seems warranted as even vocabulary knowledge contains both language-specific knowledge and more general conceptual ability.

Taking the multi-dimensionality of reading into account, Taillefer (1996) applied the threshold concept to two reading tasks involving different mental operations: 'Scanning' of passages for specific key words/numbers and 'Receptive Reading' of passages for meaning, the latter usually being the only type studied in research cited so far. Taillefer's two groups of L1-French EFL readers at the tertiary level were matched in L1 reading scores but differed significantly in L2 knowledge. When the entire sample was analysed together, L1 Scanning significantly contributed to the prediction of L2 Scanning while L2 knowledge did not have any significant effect for this type of L2 reading. For L2 Receptive Reading, both L2 knowledge and L1 Receptive Reading were significant predictors. Of the two significant predictors of L2 Receptive Reading, L2 knowledge had the stronger effect. A separate analysis of the high L2 knowledge group revealed that only the L1 Scanning score was a significant predictor of L2 Scanning score, while for L2 Receptive Reading, L2 knowledge was the only significant predictor for this group. Within the low L2 knowledge group, no independent variable significantly accounted for either L2 Scanning or Receptive Reading. From these and other results, Taillefer has shown that the language threshold varies in reading tasks of different cognitive complexity. Her study introduced the concept of differential relationships between the components and products of L2 reading depending on the types of operations involved in the reading task. Taillefer is also right in pointing out that 'further experiments with different measures of L2 proficiency are necessary' (1996:474). The studies directed towards the threshold issue have usually included only one L2 knowledge/proficiency variable (whether a vocabulary score, grammar score, or an averaged vocabulary and grammar score), implying that it is a unidimensional construct. As Taillefer has shown, examining the different dimensions of a construct previously treated as unitary may lead to findings that could otherwise not have been unearthed. L2 reading research may benefit more from a study considering multiple dimensions or components of L2 knowledge/proficiency for L2 reading skills.

Yamashita (1999) analysed data from her Japanese EFL readers and reported that L2 linguistic knowledge was a better predictor of L2 reading than was L1 reading of her entire sample, and between the two sub-domains of linguistic knowledge, L2 vocabulary exceeded L2 grammar knowledge in the amount of explained L2 reading variance. When her sample was divided into three subgroups according to their L2 reading ability level, their L2 reading performance was best accounted for by their L2 vocabulary for the top two levels but by the L2 grammar knowledge for the bottom group. L1 reading ability was implicated in the L2 reading ability for the top subgroup only and even in that group its explanatory strength was weaker than L2 vocabulary knowledge.

Most of the studies on linguistic threshold cited above (Bernhardt and Kamil 1995, Bossers 1989, Brisbois 1995, Carrell 1991, Lee and Schallert 1997, Schoonen et al 1998, Taillefer 1996) examined which independent variable (L1 reading comprehension skill or L2 linguistic competence) explained the larger proportion of variance in the dependent variable (L2 reading comprehension skill) in a multiple-regression analysis, a procedure routinely adopted in the componential analyses of L1 reading. This methodological commonality seems to provide a useful ground for further componential analysis of L2 reading comprehension skill, which may or may not converge with existing L1 reader data.

There is, however, an important difference between the typical componential analyses of L1 reading and the L2 threshold studies listed above. L1 componential studies have examined a variety of factors such as lexical decision or access speed, phonological awareness, working memory, vocabulary size, etc. simultaneously. The L2 reading studies above have, in contrast, involved no more than two factors at a time: L2 proficiency or linguistic competence and L1 reading comprehension. Given the consistently larger effect found for the L2 knowledge/proficiency variable than for L1 reading, and given the amount of L2 reading variance unaccounted for by the usual pairing of one L2 knowledge/proficiency variable and one L1 reading comprehension variable, a further consideration of the independent variable and particularly finer analysis of L2 knowledge or proficiency seems warranted to help increase the amount of L2 reading variance explained. It should be noted that some of the threshold studies summarised above (Bernhardt and Kamil 1995, Carrell 1991, Perkins et al 1989) operationalised L2 proficiency inadequately. Perkins et al (1989) used a TOEFL score for this construct, but a significant portion of TOEFL taps reading comprehension. Carrell (1991) and Bernhardt and Kamil (1995) judged the subjects' L2 proficiency by their membership in instructional level, but reading comprehension skill and its development might well have been an important aspect of advancing through the instructional levels. In explaining reading comprehension variance, one wishes to avoid a situation where the level of independent variable is explicitly or implicitly influenced by reading comprehension skill. Perhaps the construct of language proficiency is not compatible with any factor study of reading unless it is more precisely defined, such as oral/aural proficiencies or vocabulary or grammatical knowledge. Besides such better-defined L2 variables, some of the knowledge areas or subskills shown to be importantly related to reading outcomes in L1 componential studies may be meaningfully evaluated for their predictability of L2 reading outcomes.

Outside of the linguistic threshold research, the idea that more than a few subskills or knowledge areas may underlie L2 reading performance is better accepted. As has been mentioned in the early part of this section, Coady (1979) led the multi-factor conceptualisation by hypothesising the reader's use of six process strategies in L2 reading comprehension. Developed under the influence of Coady's model and the L1 interactive models is Bernhardt's (1986) Constructivist model, which depicted the three text-based factors of word recognition, phonemic/graphic decoding, and syntactic feature recognition and three extra-text-based factors of intra-textual perception, prior knowledge, and metacognition, each interacting with all the others to enable construction of text meaning. (Bernhardt seems to have used the term word recognition for the use of vocabulary knowledge in this 1986 publication.) While this model claimed to be data-generated, it was not clear about the relative involvement of the factors in the reading outcome, rendering the model little more than a factor enumeration.

Bernhardt's later model, a multi-factor model of L2 literacy (1991a), was significantly clearer than her first about the relative importance of the factors at work at different stages of L2 reading development. Like Coady's model (1979), which hypothesised gradual changes in the learners' use of six process strategies as their proficiency improves, Bernhardt's model (1991a) described the changes in the relative amount of involvement of five subskills or knowledge areas (excluding metacognition from the six listed for her Constructivist model above) along the dimension representing an increasing ability of L2 reading. Five unique curved lines represented the relative rate of errors in reading which can be attributed to the five factors at any given point on the reading development dimension. This model was based on a large set of empirical L2 reading data (n=300) and methodical transcription of reading recall protocol. Unfortunately, the analysis on which the model was based is largely qualitative, focusing on the researchers' interpretation of the sources of the learners' misreadings of various parts of the texts. Quantitative and statistical reasoning of the type usually employed in psychology or applied linguistics before constructing any model of human behaviour or ability were not offered for triangulation, which left the model standing without full underpinnings. That is probably why Bernhardt commented, after reproducing this model in her similar discussion of the L2 reading factors, 'Validation (of this model) lies in the future' (1991b:40). No reference to an L1 componential approach to reading appears with Bernhardt's model (1991a, 1991b), but it would be a natural methodological source for the empirical validation of her model since it postulates multiple factors of reading differing in importance. Some of the factors shown to cause misreadings in Bernhardt's qualitative analysis can be subjected to multiple regression analyses to discover which of them explain the reading comprehension variance among L2 readers. Of the factors Bernhardt includes, phonemic/graphic decoding skill has been a standard variable in L1 componential analyses, and word knowledge and syntactic knowledge are the two most frequently examined aspects of L2 linguistic competence for threshold studies. L2 componential research including these variables to measure the effects of each on the reading comprehension outcomes must help validate or invalidate the theory advanced by Bernhardt, and the multiple regression analysis seems appropriate for an investigation of this nature.

One study which considered a similar set of variables to Bernhardt's (1991b) and analysed the data with the multiple regression design is Haynes (1989, also reported more concisely in Haynes and Carr 1990). Her study differed sharply from the typical L2 threshold studies since it simultaneously considered a number of variables and not simply one L1 reading variable along with an L2 proficiency variable. It is a rare piece of work which combines the characteristics of the cognitive psychological approach to reading skill-components and the applied linguistics focus on the L2 readers (see also Nassaji and Geva 1999, reported below). It thus seems to be an excellent reference for further componential analyses of L2 reading and will be described in detail below.

Haynes collected data from 60 L1-Chinese EFL readers at a Taiwanese university on a total of nine independent and five dependent variables. Since her primary interests were the effects of orthographic knowledge and the efficiency of visual processing on L2 reading outcomes, the independent variables had five speeded tests of number matching, letter string matching, pseudoword matching, real word matching, and synonym/antonym decision, but they also included listening comprehension, vocabulary range, grammar knowledge, and L1 reading comprehension. For dependent variables, one reading comprehension, two reading speed, and two contextual word learning measures were adopted.

Three points from Haynes's correlational data merit our attention. First, there was a clear dissociation between the correlates of comprehension and the correlates of speed: i.e., the performances on the speeded tasks of visual matching and synonym/antonym decision correlated with reading speed measures but not with the reading comprehension measure. Second, listening comprehension, vocabulary range, and grammar measures correlated with all five of the dependent variables. Third, all independent variables correlated with the two contextual word learning measures. These results underscore the distinctiveness of the dimensions of L2 reading performance as well as the pervasive character of the L2 linguistic knowledge in L2 reading outcomes.

Although it has been specified earlier that the current study will focus on Careful Reading at the passage level rather than other kinds of reading (e.g., Skimming, Scanning, etc.), reading speed also seems to form an interesting subject of research. Haynes's first set of multiple regression results with the two reading speed measures as dependent variables showed the significance of number matching efficiency, synonym/antonym decision latency, and grammar knowledge. Haynes inferred from such results that a general visual processing efficiency, lexical-semantic processing efficiency, and a grammatical or syntactic factor contribute to the speed of L2 reading and questions the arguments by Ulijn (1981, 1984) that L2 reading requires little syntactic processing but much lexical-conceptual processing.

A separate multiple regression analysis revealed that, of the L2 reading comprehension variance, L1 reading accounted for the first 5%, the orthography effect (the difference between pseudoword and letter string matching efficiencies) an additional 4%, listening comprehension an additional 20%, and vocabulary range an additional 3%. This highlighted the significance of the listening comprehension or general L2 comprehension component in L2 reading comprehension while the effects of L1 reading, orthographic knowledge, and L2 vocabulary range appeared marginal.

The third set of her multiple regression analyses tested how the predictor variables explained the variances in the two contextual word learning scores. Of the variance in the scores, number matching efficiency and L1 reading together accounted for about the first 19% (with number matching explaining consistently larger 11 to 14%), the orthography effect an additional 6 to 7%, and vocabulary range an additional 15 to 26%. Haynes's multiple regression data suggested that such factors as general visual processing efficiency, lexical access efficiency, knowledge of L2 grammar, general L2 comprehension ability manifested via listening comprehension, and vocabulary range contributed more significantly to certain types of L2 reading outcome, while orthographic processing efficiency and L1 reading ability can account for smaller proportions of both L2 reading comprehension and contextual word learning variances.

Another study which attempted to examine some component processes in L2 reading is Nassaji and Geva (1999). With a group of 60 advanced ESL readers from L1-Farsi background, they examined the effects on reading performances of several component skills, including phonological, orthographic, lexical-semantic, sentence-semantic, and syntactic processing. Based on composite scores of accuracy and speed, their regression analyses indicated that lexical-semantic, sentence-semantic, syntactic, and orthographic processing tasks consistently made significant contributions to the prediction of reading comprehension and rate. Caution is necessary when interpreting Nassaji and Geva's results since (1) their composite 'efficiency' scores confound accuracy and speed, (2) their predictor measures vary in reliability, and (3) their sentence-semantics variable is difficult to clearly distinguish from reading comprehension of sentences (albeit isolated).

The type of L2 componential approach adopted by Haynes (1989) and Nassaji and Geva (1999) may provide important information on what kinds of knowledge and subskills and combinations of them can explain the individual differences in the L2 reading outcomes but this is exactly the type of approach lacking in the field of L2 reading research. Also, these valuable attempts benefit from methodological improvement. For instance, Green (1991) recommends 50 plus eight times the number of predictor variables as the minimum sample size for multiple regression. Haynes's and Nassaji and Geva's study would be meaningfully expanded by further consideration of the variables with a more appropriate sample size.

The most difficult decision to make seems to be on the selection of the variables, since the theoretical need for a larger number of variables conflicts with the practical limitations on the sample size as well as the need for a reasonably small amount of testing through which the individuals can be asked to provide ability data. This balance between theoretical interest and research feasibility is an issue requiring careful consideration for the current study as well and one that will be addressed directly in the 'Methodology' section. At this point, however, it is the theoretical interest that is the criterion for further discussion.

Knowledge areas and psycholinguistic processing subskills

It seems appropriate now to return to the knowledge areas or psycholinguistic processing subskills identified as significant in the section on L1 componential analyses – vocabulary knowledge, word recognition efficiency, phonological awareness, and working memory – so that each can be evaluated for its compatibility with L2 componential analysis in light of the L2 reading theories and data referred to thus far. An additional variable, knowledge of target language syntax, must also be discussed since it has often been considered a significant part of L2 linguistic knowledge in threshold research or integrated as a contributory strategy or factor in L2 reading theories.

L2 vocabulary breadth and L2 reading

In both L1 and L2 reading research (e.g. Baddeley et al 1985, Bernhardt 1991a, 1991b, Dixon et al 1988, Laufer 1992b) there is an emphasis on some sort of word knowledge or aspects of vocabulary, which coincided with the growing interest among the L2 researchers in the general value of teaching and researching vocabulary knowledge (Coady and Huckin 1997, Huckin, Haynes and Coady 1993, Meara 1983, 1987, 1992, Nation 1990, 2001, Schmitt and McCarthy 1997). The idea that the knowledge of words is basic to the comprehension of messages composed of them is intuitively appealing and has received support from L1 reading researchers. According to Beck and McKeown, 'In factor analytic studies of reading comprehension, vocabulary knowledge has consistently emerged as a major component . . . correlating very highly with comprehension ability' (1991:805), and Daneman goes so far as to say it is 'one of the best single predictors of reading comprehension performance' (1991:524). To proceed to investigate the role of such knowledge in L2 reading, we must be clear about what kind of knowledge it is. Vocabulary is defined as 'a set of lexemes (the smallest unit in the meaning system of a language distinguishable from other similar units), including single words, compound words and idioms' (Richards, Platt and Platt 1992:400), and it is the consensus among vocabulary researchers that lexical knowledge is multi-dimensional, involving such features of individual words as their semantic range, occurrence or co-occurrence probability, syntactic behaviour, association with other words, derivations, pronunciations, etc. (Meara 1996, Nation 2001, Richards 1976). However, the dimension of vocabulary which appears most often with its name in reading research is vocabulary range or size. This is also referred to as the 'breadth' dimension of lexical competence and is equated with the quantitative individual attribute often expressed in the number of words the person is estimated to know the meaning of. Vocabulary breadth must be evaluated for its relationship with L2 reading, along with some other dimensions of lexical and other linguistic competence suggested as dissociable from it, and not be treated as synonymous with linguistic competence, as was done in some studies into the relationship between L2 knowledge and reading (e.g. Hacquebord 1989, Schoonen, Hulstijn and Bossers 1998).

L2 word recognition efficiency and L2 reading

Another dimension of lexical knowledge that was among the variables supported in L1 reading research was word recognition efficiency. While vocabulary size usually concerns one's knowledge or lack of knowledge of the words in the language, word recognition efficiency concerns the general ease with which one can process the 'known' words or constituents of words. Efficient word recognition is associated with the automaticity concept (e.g. Shiffrin and Schneider 1977), which has been applied to the theories of L2 development in general (e.g. McLaughlin, Rossman and McLeod 1983) and L2 word recognition in particular (Segalowitz and Segalowitz 1993, Segalowitz, Watson and Segalowitz 1995). The theory assumes that human mental resources are limited at any given moment, and success in such complex cognitive skills as reading comprehension requires automatisation of the so-called 'lower level' processes such as letter or word identification. This ensures that sufficient mental resources are left available for other 'higher level' processes such as temporary storage of phonological or propositional representations of the text or integration of them with previously presented ones for a global comprehension of the discourse. Reflecting the growing concern among some researchers on the need to develop automaticity in the decoding skills of L2 readers (e.g. Eskey 1988, Grabe 1991, Segalowitz, Poulsen and Komoda 1991), L2 visual word recognition research has produced enough work to justify a publication of literature review (Koda 1996). This is a significant step forward since it means at least one subskill of reading with extensive L1 work (Stanovich 1991) now also has a substantial L2 research base, and avenues for further research have been illuminated. At the same time, as expressed in Koda's account, 'Compared with cognitive and metacognitive processes, L2 decoding has received far less attention, and as a result, much remains unexplored' (1996:450). Thus, despite the growth of the L2 word recognition research, it is still in serious need of empirical research expansion. Further systematic studies of the relationship between L2 word recognition skills and L2 reading comprehension skills are needed to help us to determine if word recognition efficiency is as significant in L2 reading as has been shown in the L1 reading literature (Stanovich 1991).

Word recognition requires both lexical knowledge and sublexical knowledge, which further consists of the knowledge of physical distinctions of the letters, identities of letters differing in case, orthographic regularity or probability of letter sequence, and grapheme-phoneme correspondence. The speeds of access to lexical and sublexical knowledge areas have been considered in many multivariate componential analyses of L1 reading (Baddeley et al 1985, Cunningham et al 1990, Dixon et al 1988, Jackson and McClelland 1979, Palmer et al 1985). In sharp contrast with this state is the lack of their L2 counterparts.

Among the few componential studies which considered the L2 readers' individual differences in word recognition efficiency are Brown and Haynes (1985), Haynes (1989), and Nassaji and Geva (1999). The first two studies adopted, among many measures, latencies in processing existing English words and nonwords, which were further divided into pseudowords and irregular letter strings. Pseudowords referred to strings of letters which followed some orthographic regularities found in existing English words (e.g. prossing) and irregular letter strings to those which violated such regularities (e.g. prngesis).

As nonwords are presumably novel in letter sequence, they are considered to require word-internal analysis, ruling out the possibility of any individuals relying on the whole word approach.

The inclusion of the nonwords also enabled the researchers to measure the effects of visual familiarity, or lexicality (by comparing the word recognition latencies and pseudoword recognition latencies), and that of orthographic familiarity (by comparing the pseudoword recognition latencies and irregular string recognition latencies). Better readers seemed to benefit more from orthographic familiarity, presumably due to their developed sublexical decoding speed, which must be crucial in encoding new word forms and meanings efficiently and effortlessly for both text comprehension and lexical and reading development.

Haynes (1989) additionally measured her Taiwanese EFL readers' synonym/antonym decision speed and found it to be correlated significantly with their passage reading speed.

Nassaji and Geva's (1999) orthographic processing measure focused on knowledge of permissible orthographic sequences in English. Each test item was a pair of nonwords, such as *gmub-gnub,* where only one member (*gnub* in this example) contained a sequence that can occur in English and is to be selected as the correct answer.

Despite the continued interest in developing efficiency in lower level processes for successful L2 reading (e.g. Segalowitz, Segalowitz and Wood 1998), such early attempts as Brown and Haynes (1985) and Haynes (1989) to explicate the relative significance of rapid lexical/sublexical recognition skills have rarely been followed up in other componential research of reading abilities.

L2 phonological awareness and L2 reading

Phonological awareness, the reader's sensitivity to the constituent sounds in words, is considered to be related to the skill of word recognition and reading achievement in general. More specifically, phonological awareness is said to correspond to our ability to segment a stream of syllables or phonemes, identify a particular syllable, phoneme, or rhyme in such a stream, and manipulate the sequence by adding, deleting, or substituting syllables, phonemes, or rhymes (Goswami and Bryant 1990). Ability to perform these tasks that seem rather easy to adult speakers of an alphabetic language are not easily available to the early readers and speakers of non-alphabetic language. The involvement of phonological awareness in children's reading development has been well noted among L1 reading researchers (e.g. Cormier and Dea 1997, Goswami and Bryant 1990, Hansen and Bowey 1994, Leather and Henry 1994, Levy and Carr 1990, Stanovich et al 1984). Whether this applies in the case of individuals learning to read in L2 with different orthographic patterns or scripts from their L1 can constitute an interesting question and one that is starting to draw attention of the researchers dealing with L2 readers.

Durgunoğlu, Nagy and Hancin-Bhatt (1993) have shown that phonological awareness developed in alphabetic L1 (Spanish) can account for individual differences in word recognition skill in alphabetic L2 (English) among beginning readers. Koda (1998) and Wade-Woolley (1999) reported on studies involving phonological processing of adult L2 readers belonging to two different L1 groups each. Koda (1998) explored the effects of differing

L1 alphabetic experience on L2 phonological awareness by comparing the phonological awareness of L1-Chinese (non-alphabetic) and L1-Korean (partially alphabetic) ESL learners at an American university and analysed its relationship with their English decoding and reading comprehension skills. The two groups, which had been matched based on a TOEFL listening subsection score, did not differ in phonological awareness or decoding performance but only the L1-Korean group's phonological awareness and decoding significantly accounted for their reading comprehension variance. Wade-Woolley (1999) compared L1-Japanese (non-alphabetic) and L1-Russian (alphabetic) groups of ESL learners who showed similar group performance on TOEFL reading and her correlation results suggested that the Japanese group relied less on phonological processing than on orthographic processing. (Nassaji and Geva (1999) considered what they called phonological *processing* skill, but it will not be discussed in detail here as it is different from phonological *awareness* as conceptualised and operationalised in the literature.)

These represent some interesting but rare attempts at exploring the relationship between phonological awareness and L2 reading among the learners with disparate alphabetic background and encourage more research on the role of phonological awareness in L2 reading.

L2 verbal working memory span and L2 reading

Many researchers consider working memory as an important aspect of normal language functions (Cowan 1996, Gathercole and Baddeley 1993), successful text comprehension and production (Kintsch and van Dijk 1978), and L2 learning (Service 1992). Working memory capacity refers to the simultaneous information storage and processing functions of memory for complex cognitive activities (Baddeley and Hitch 1974) and is regarded by some as a major source of individual differences in verbal comprehension (e.g. Just and Carpenter 1992). Complex span tasks (Daneman and Carpenter 1980) that simultaneously require verbal processing and short-term storage have repeatedly predicted reading comprehension among adults (Baddeley et al 1985, Cunningham et al 1990, Daneman and Carpenter 1980, 1983, Dixon et al 1988, Engle, Nations and Cantor 1990, Masson and Miller 1983, Osaka and Osaka 1994, Turner and Engle, 1989) and children (Leather and Henry 1994, Swanson 1992, Swanson and Alexander 1997). Researchers concerned with adult L2 learners or bilinguals have also explored the concept of L2 verbal working memory (Da Fontoura and Siegel 1995, Geva and Ryan 1993, Harrington and Sawyer 1992, Möbius 1995, Nassaji and Geva 1999, Osaka and Osaka 1992, 1994, Osaka, Osaka and Groner 1993). However, only one published study (Harrington and Sawyer 1992) specifically addressed the relationship between complex-span working memory and L2 reading comprehension among adults (for children's data, see Geva and Ryan 1993 for correlation between their L2 working memory test and cloze scores). Harrington and Sawyer found a significant correlation between complex-span working memory of their L1-Japanese EFL learners and their Vocabulary and Reading Comprehension section score on TOEFL. Since the latter was a composite index of vocabulary size and reading comprehension, their study did not precisely address the role of working memory in reading comprehension. It would seem meaningful to measure the L2 readers' working memory and reading comprehension skills to evaluate their relationship as well as to investigate how much unique variance in reading comprehension working memory can account for.

L2 syntactic knowledge and L2 reading

While syntax is rarely explored as a factor in normal adult L1 reading, it is sometimes considered in research on children's reading development (Tunmer, Nesdale and Wright 1987) and very frequently in L2 reading research (e.g. Berman 1984, Bossers 1989, Brisbois 1995, Lee and Schallert 1997, Perkins et al 1989, Taillefer 1996, Yamashita 1999). Richards et al (1992) define syntax and grammar in a similar manner. According to them, one meaning of grammar is 'the way in which linguistic units such as words and phrases are combined to produce sentences in the language' (1992:161) and syntax is concerned with 'how words combine to form sentences and the rules which govern the formation of sentences' (1992:370). Grammatical or syntactic knowledge is thus primarily concerned with the well-formedness (or ill-formedness) of a sentence or parts of a sentence such as a clause or a phrase. Knowledge of such aspects as the possible transformations or movements of elements, and different realisations of word forms based on the co-occurring words that control them can be considered a part of syntactic knowledge. Although a wider conceptualisation of grammar seems to subsume much more than syntax, the current study will use the term grammar to refer to syntactic knowledge.

L2 reading studies such as Bossers (1989), Taillefer (1996), and Lee and Schallert (1997) included a composite index of vocabulary and grammar knowledge, while others (e.g. Carson, Carrell, Silberstein, Kroll and Kuehn 1990, Perkins, Brutten and Pohlmann 1989) either directly or indirectly consulted scores on such standard L2 proficiency tests as the TOEFL or the Michigan Test, in which the grammatical knowledge section occupied a significant proportion. While this underscores some researchers' awareness of grammar knowledge as an indispensable aspect of the overall L2 competence or proficiency in studies of L2 reading, it does not reflect any interest in exploring the relationship between grammar knowledge itself and reading in L2.

Results from a small number of L2 reading research which compared the effects of syntax/grammar knowledge with others' are mixed. Brisbois (1995) and Yamashita (1999) reported that grammar was generally a weaker predictor than vocabulary knowledge. Nassaji and Geva's (1999) measure of syntactic processing skills predicted reading ability but not as well as their vocabulary measure did, while Haynes (1989) identified grammatical knowledge as significantly and uniquely accounting for L2 reading speed.

Large-scale data in Alderson's (1993) research showed grammar test scores to be very predictive of reading test scores; in fact they correlated with scores on different reading tasks more strongly and consistently than such reading tasks correlated among themselves. Of interest is Alderson's own admission that he was unable to avoid a degree of 'contamination' of the grammar variable since his grammar measure involved visual processing of sentence semantics, i.e., a kind of reading. Alderson recommends, 'It might be fruitful to attempt to devise tasks that require no processing of meaning and to contrast performance on them with performance on more meaning-oriented tasks as well as on a series of defined reading tasks.' (1993:218)

The degree to which L2 syntactic knowledge predicts L2 reading comprehension and how it interacts with other potential knowledge factors such as vocabulary range should be examined with more rigour than most L2 reading research has proceeded with so far.

L2 listening comprehension and L2 reading

At least one L1 reading study suggested that our L1 reading comprehension could be accounted for mostly by our L1 listening comprehension (Palmer et al 1985). However, since the listening instrument in their study was an aurally presented version of a passage reading comprehension test, as the authors themselves point out, the observed correlation must be interpreted with caution. Some studies involving L2 readers have also found a link between L2 listening comprehension and L2 reading comprehension. Brown and Haynes (1985) found a significant correlation between the two skills among their subjects with Spanish, Arabic, and Japanese L1 backgrounds. Additionally, Haynes (1989) found listening comprehension to be the best predictor of her Taiwanese university students' reading comprehension ability. These results constitute an initial database from which to tentatively infer that, in L2, comprehension performance in listening can predict that in reading.

L1 reading and L2 reading

While summaries of threshold studies (see pp. 12–22) have indicated that L2 reading comprehension ability was generally better explained by L2 linguistic knowledge than by L1 reading skill and efforts should focus on the different dimensions of L2 knowledge, the role of L1 reading skill does not seem to be negligible (e.g. Carrell 1991, Haynes 1989, Lee and Schallert 1997). Though secondary to the L2 component areas addressed above, an aspect of the L1 reading ability will be included in further discussions of the design of the present study.

L2 reading speed

Throughout the discussions so far, the primary interest of the current study has been comprehension ability of the L2 readers rather than their speed of reading. Although the review of previous research has naturally paid most attention to reading comprehension, the results concerning reading speed were incorporated as well whenever relevant. To return to some of such speed related results briefly, in the L1 context, Jackson and McClelland (1979) found listening comprehension had the largest overlap with reading speed, and Palmer et al (1985) suggested that the reading speed factor is dissociable from verbal comprehension in general. One of the most relevant L2 reading studies by Haynes (1989) produced results which converged with Palmer et al's (1985) conclusion. Nassaji and Geva (1999), on the other hand, showed their L1-Farsi ESL readers' comprehension and speed measures to be highly correlated, and their two criterion measures of comprehension and speed had similar patterns of relationship with their respective predictor variables. Despite the attention that the L2 componential studies have paid to L2 reading rate, our research base is still extremely limited and it would appear to be meaningful to attempt to understand more about this significant variable, besides our main interest of passage comprehension ability.

Tentative research questions

The previous sections have discussed seven areas of knowledge or skill – L2 vocabulary size, L2 word recognition efficiency, L2 phonological awareness, L2 working memory, L2 syntactic knowledge, L2 listening comprehension, and L1 reading – and proposed the need to examine these areas as potential factors explaining the individual differences in adult L2 reading comprehension ability. This can be framed as a research question in the following manner:

> Q1: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills – L2 vocabulary size, L2 word recognition efficiency, L2 phonological awareness, L2 working memory span, L2 syntactic knowledge, L2 listening comprehension, and L1 reading skill – explain the individual differences in L2 careful reading comprehension ability?

Subsumed under this research question is the relationship between each of the knowledge or skill areas and L2 reading comprehension ability. Examining such relationships individually is the first step toward a fuller explanation of individual differences in L2 reading ability. This is, however, more or less where many of the L2 reading studies rest the case (e.g. Alderson 1993, Harrington and Sawyer 1992, Laufer 1992a). Simply establishing a relationship between a subskill and reading outcomes does not truly allow us to compare the relative significance of the subskills in accounting for the overall reading comprehension ability. When more than one subskill equally show a significant relationship with reading ability, those subskills may do so since they are themselves related, in which case the multiple subskills may actually represent a common underlying trait, or since they are differentially related to reading, in which case the multiple subskills are rather unrelated among each other. The question must therefore include the extent to which the knowledge and skill areas relate with each other.

This exploration of the interrelationships among the subskills will address a number of subsidiary research questions. For instance, are EFL learners with a large sight vocabulary fast at visual identification of basic words? Do individual differences in phonological awareness coincide with those in word recognition efficiency, and/or with vocabulary size? Are sight vocabulary size and syntactic knowledge separate components or do they in fact overlap to any significant extent? What relationship does working memory span have with sight vocabulary size and/or syntactic knowledge? The degrees of these interrelationships seem to form distinct areas of research in themselves. At the same time, determining the degrees of such interrelationships serves another important purpose. When these subskills are evaluated for the degree to which they account for the individual differences in reading ability, one should be interested in how much each subskill can uniquely contribute to the explanation of the reading ability differences. Even if each subskill seemed to correspond to the overall reading comprehension ability well, the more closely interrelated the subskills are among each other, the smaller the unique contributions that each of these subskills will make in accounting for the individual differences in reading. Analysis of the interrelationships among the potential subskills is thus recommended as a means of addressing several subsidiary questions and as a necessary prerequisite of the further analysis of the extent to which each subskill uniquely accounts for the reading ability differences across individuals.

The L2 reading literature (see pp. 12–22) has often indicated that subskills change their importance for overall reading skill as a function of the learners' reading ability (Bernhardt 1991b, Bossers 1989, Carrell 1991, Coady 1979). Following the linguistic threshold analyses, L2 component skills analysis can observe the relationships between the potential subskills of L2 reading with overall L2 reading comprehension within some subgroups of the L2 readers demarcated according to their L2 reading ability. Subskills may relate to reading ability differently across the subgroups. Some subskills may be more closely linked to reading performance of more proficient readers than to that of less proficient ones, or vice versa. Alternatively, they may relate to reading to the same extent regardless of which level of reading proficiency group was observed. Thus, the second research question is:

> Q2: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) explain the individual differences in L2 reading comprehension ability as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

The subsidiary questions that are addressed within Q1, the interrelationships among the subskills, can be examined within the subgroups of L2 readers as well.

An additional interest, which emerged as a byproduct of a review of reading comprehension research, is in reading speed. This dimension of reading outcome should deserve a separate research question. Therefore the following two questions are added:

> Q3: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills (listed in Q1) explain the individual differences in L2 reading speed?
>
> Q4: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) explain the individual differences in L2 reading speed as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

The research questions formed above would have to be tentative. This is because the feasibility of adopting every single one of the proposed variables in a fully multivariate study is yet to be established. Each of the knowledge and skill areas in the research question must be considered carefully in terms of their research feasibility. The 'Methodology' section that follows in Chapter 3 will describe the process through which each variable was evaluated more on that basis. The research questions are thus subject to revision as a result of this evaluation.

3 Methodology

Chapter overview

The present chapter begins with an exploration of the appropriate methodological direction and sample requirements for pursuing the research questions initially formed in the previous chapter. The prospective participant group is then proposed, and the learner's skill areas identified as tentative research variables in the previous chapter are reexamined from the perspective of research feasibility and are operationalised for constructing a realistic research programme. The operationalisations of the variables will lead to a revised set of research questions, and the chapter will conclude with a description of the general design and the procedure for the study as well as the information on data collection and data processing.

General direction of methodology

The tentative research questions formed in the previous section should provide a guiding force in finding methodological direction for the study. Pearson's product-moment correlation will initially help us compare the skill and knowledge variables in the general degree to which they overlap with the reading measures and identify some of the more significant variables. However, correlations cannot index the independent contributions these skill variables may make in accounting for reading abilities because they ignore the amount of overlap these skill variables may have with each other. The methodology which overcomes this weakness and which is relatively popular for this type of research is multiple regression. Partial regression coefficients obtained via multiple regression analysis can indicate the degree to which each skill or knowledge variable covaries with the criterion reading ability variable when its overlap with all the other skill variables has been statistically removed; thus they are suitable for evaluating the strength of each skill variable's independent effect in accounting for the reading test performance. The review of the literature above has shown that both L1 and L2 studies which aimed to identify components or important explanatory factors of reading made frequent use of multiple regression analysis (e.g. Bernhardt and Kamil 1995, Carrell 1991, Dixon et al 1988, Jackson and McClelland 1979, Lee and Schallert 1997, Taillefer 1996). As in these previous studies, an index of reading ability can be specified as the 'criterion' (or 'dependent') variable and indices of knowledge or skill areas, such as vocabulary breadth or the speed of visual word recognition, can form the 'explanatory' (also 'predictor' or 'independent') variables. Those skill or knowledge variables that resulted in statistically significant partial regression coefficients can be considered important. If more than one skill variable emerge as statistically significant, standardised partial regression coefficient (or beta coefficient) associated with each predictor variable will indicate the relative significance of the explanatory variables.

Another type of analysis that can be adopted to explore reading ability is factor analysis. While conventional regression methodology is concerned only with the relationships among the observed variables, factor analysis attempts to identify a smaller number of unobserved or latent variables which can account for the relationships among the observed variables. Weir, Yang and Jin (2000) and Negishi (1996) adopted factor analysis to study the latent structure of L2 reading test performances. Although regression analysis is the principal tool in our attempt to answer the research questions, use of factor analysis should also enable an approach to the data from an additional angle. If any latent variable(s) should appear to account for the correlations among the observed reader variables, the nature of the latent variable(s) can be speculated and a model of the sample group's ability will be formulated.

Finally, if any such model should be formulated, it can also be subjected to confirmatory factor analysis (CFA), which allows us to statistically test our models of relationships among the observed and latent variables in terms of a goodness of fit between each of the formulated models, or equations, and the set of empirical data actually obtained. Palmer et al (1985) and Cunningham et al (1990) are examples of L1 reading studies heavily drawing on this methodology. Also known as covariance structure analysis or structural equation modelling (SEM), this methodology has been utilised less widely than regression or exploratory factor analysis (EFA) in the community of L2 researchers. There seems to be a growing interest in it (cf. Kunnan 1998), and its use has been increasing among researchers concerned with language proficiency and testing (cf. Kunnan 1995, Purpura 1999, Sasaki 1993, Schoonen et al 1998).

Sample size

Since multiple regression analysis serves as the core of the analysis tools for this study, the minimum sample size required for this procedure should be the initial goal. One formula available in the literature for estimating the minimum sample size for the procedure is 50 + 8 times the number of independent variables (Green 1991). The discussion in the previous chapter has identified seven potential areas that may be observed as independent variables. If all of these seven knowledge and skill areas are to be included in the multiple regression analysis, then 106 would be the minimum sample size. Since the explanatory variables may be further categorised into smaller subcomponents, which would increase the number of variables, conservative estimates should be adopted at this point. A target sample size of 200 should be a safer estimate. As for exploratory or confirmatory factor analysis, Bryant and Yarnold (1995) warn that a large sample size is required. A rough guideline they offer for estimating the minimum sample size for factor analytic research is 5 to 10 times the number of observed variables, which would require a minimum sample of 35 to 70 persons. Since a sample size of 35 to 70 is by no means large compared to the factor analytic research cited above, it is speculated that the guideline is meant for research involving a much larger set of independent variables than the present one. Even after taking such factors into consideration, the more conservative figure of 200 proposed above seems sufficiently large for our purposes. Therefore, the present study attempts to reach that sample size.

Participants

The target sample is to be drawn from L1-Japanese groups at the tertiary level of education in Japan. It is firstly the expected availability of participants to the researcher which led to this decision, and secondly it is the need that the researcher personally feels of increasing our understanding of the abilities of the Japanese EFL learners.

Students from a broad spectrum of English proficiency – national and private 4-year university entrants, undergraduates and postgraduates, some English majors and others non-English majors (but all with secondary-level EFL education) – are invited to participate in the research project. The common L1 background eliminates the effect of L1 differences and allows us to focus on the internal factors of linguistic knowledge and psycholinguistic processing skills. Also, the relatively homogeneous secondary-level EFL curriculum in Japan reflecting governmental guidelines is expected to keep the effects of teaching methods and the divergence of EFL development pattern at least as small as any other L1 groups. Nevertheless, the study will employ a questionnaire to elicit information about any outstanding exposure to or experiences with the target language so that unusual patterns emerging in the data might be accounted for by any uniquenesses of particular individuals.

To summarise, a minimum of 200 L1-Japanese EFL learners at the tertiary level of education in Japan are to constitute the main participant group for the study.

Feasibility considerations

Although a total of seven explanatory knowledge and skill areas have been identified as potentially important, they need to be considered again from the viewpoint of research feasibility, particularly in light of the available opportunities for the Japanese university EFL students to participate in the study. From the researcher's experience as an EFL instructor at Japanese universities, it is estimated that the longest possible time that can be asked of EFL classes for any research-related testing is 90 minutes even if the course instructor or director finds educational relevance in the tests. This must include the time for distributing and collecting the papers and explaining the procedures; thus the actual testing time would be even shorter. Individual testing of any nature must be scheduled outside of their class time, and the researcher's experience again predicts difficulty in recruiting 200 students for out-of-class participation requiring more than 30 minutes of their time even if payment is offered. One of the conditions to be met, therefore, in determining the actual number and the nature of variables is that the data collection for all of those variables must fit in 80 minutes of group-administration plus 30 minutes of individual administration. Those variables on which data can be collected through a questionnaire are considered separately, since a questionnaire may be filled out at the learner's leisure and need not affect the total test time.

Another issue related to research feasibility is the amount of preparation required for the measurement instruments. Since multiple measurement instruments are to be prepared by a single researcher with the obvious need for piloting with some Japanese sample groups, it would be virtually impossible to complete the entire study in a reasonable length of time if each of the instruments were to be constructed completely anew. Therefore, every possibility of adopting existing instruments that may be suitable for reliable and valid measurement of Japanese university students' abilities should be sought, and such existing instruments may be modified through item analyses and/or content validation. If no existing instrument should become available for the measurement of a particular variable, that variable may need to be reconsidered or excluded from the research design despite its theoretical significance.

It is with these practical considerations in mind that the research variables are examined again and an initial attempt is made to operationalise them below.

Operationalising the criterion variables

Attempts are made in this section to operationalise the knowledge and skill variables that have been identified as worthy of further analyses. Further survey of previous L1 and L2 reading research should help identify the most appropriate form of instrument available or modifiable for the current research.

L2 reading comprehension ability

Ability to comprehend the ideas expressed in written discourse in a narrative or expository passage through a careful reading can be measured in several different ways. Comprehension is a result of an internal, mental process without any observable sign of success or failure, unless it is explicitly indicated by the reader. To elicit this reader indication, several methods of reading comprehension measurement are available (Alderson 2000, Cohen 1993, Heaton 1975, Hughes 1989, Weir 1993).

Two of the most commonly used formats include responding to some questions regarding the contents of the text or reproducing the reader's version of the text. Both types may be answered in the reader's L1 and/or L2. The first type may involve multiple-choice questions (MCQ) or short-answer questions (SAQ), which may or may not allow the reader access to the actual text when answering the questions. The second type may be subdivided into full recall or summary, in either free production or more structured gap-filling type, which seems to have some affinities with the structured SAQs. All types have some advantages and disadvantages. Unless they are constructed with great care, MCQs can distract the reader from the act of text reading, and they can be answered without recourse to the text or only by eliminating the distractors and without really comprehending the text. However, they are expected to deliver the highest level of objectivity and reliability, and for this reason were included in TOEFL's reading comprehension section (cf. Spolsky 1995). All other question types allow for subjectivity in responding, some to a larger extent than the others. Reproduction of the text, whether it is the summary or 'all that can be remembered', will have to involve the skill of language production. Therefore, one faces the unwanted possibility of better comprehenders failing to indicate their comprehension as a result of their poor production skills. Acceptability of the reproduced points is not expected to be so clear-cut, and subjective judgment of the raters involved will make this measurement format less reliable. To a lesser degree, SAQs seem to face basically the same difficulty. Reproduction measures also face the dilemma of introducing the unwanted memory factor, if the text is removed before the reproduction, or introducing the confounding of text comprehension with rewriting skill, if the text is left available during the reproduction. Thus, SAQ and MCQ formats emerge as potentially more reliable and valid measures of reading comprehension if the dangers inherent in them are carefully dealt with. In terms of the objectivity of measurement, MCQ format is the most appropriate. It is also the format adopted much more frequently in the reading literature to tap the reading comprehension among both L1 and

L2 readers, which will facilitate the comparison of the results from this study with those from such previous studies. Therefore, the MCQ format was considered further as the measure of careful reading comprehension ability.

A number of reading comprehension measures using MCQs are available. However, not all of them are constructed with the specific intent of asking the reader to read beyond the sentence boundaries. Without such an intent, it is quite possible for a test with passages, appearing to measure the reader's passage reading comprehension, to be actually measuring extremely local reading or knowledge of particular lexical items alone. Such items should be better labelled vocabulary items instead of passage reading comprehension items. Since one of the interests of the current study has been the extent to which vocabulary breadth explains the individual differences in passage reading comprehension, it is crucial for the reading comprehension test to avoid items focusing on extremely local linguistic information. The test items should ideally require the reader to synthesise information from across the entire passage. However, it is often difficult to include many such items in a given passage, as it would result in the need for more passages and longer test time to obtain an acceptable level of test reliability. A realistic goal to be set for time-constrained data collection, then, is for each item to require synthesis of information from across several sentences and that a correct answer cannot be identified by simply locating or understanding a particular word or phrase. Passage reading tests satisfying this condition that have also indicated empirical evidence of reliability are not as easy to find. As a starting point, items from the reading comprehension portion of the College English Test (Yang and Weir 1998), a standardised test for Chinese university populations with known reliability data, were chosen for trialling with a Japanese sample. Four reading passages, each with five questions, constituted this initial trial version of the test. The reading passages were 251–303 words in length (Table 3.1) and were all expository in nature. Topics discussed in the passages were 'Failure of Skylab', 'Efficiency of Agricultural Production', 'Changing Banking Practices' and 'Genetic and Individual Learning'. This

Table 3.1 Passage information
Reading comprehension test: version 1

Passage	**1**	**2**	**3**	**4**
Topic	**Failure of Skylab**	**Efficiency of Agricultural Production**	**Changing Banking Practices**	**Genetic and Individual Learning**
Length (Words)	251	216	261	303
Flesch Reading Ease Score	61	39	54	47
Flesch Kincaid Grade Level Score	14	17	15	16

initial set of passages and their comprehension questions were trialled with a sample of Japanese university students and the results are detailed in Chapter 4 (pp. 55–59).

L2 reading speed

In previous componential studies that observed reading speed in both L1 (Jackson and McClelland 1979, Palmer et al 1985) and L2 (Haynes 1989), the construct of their interest was the reading speed for printed texts at the passage-level. It was therefore reasonable that they timed their readers as they sped through the printed passages. It was also appropriate that all three of the studies included post-reading comprehension questions so that the instrument would elicit speed of reading for comprehension (rather than, e.g. extreme scanning), a point stressed by Davies as a vital requirement in a test of reading speed (1990).

However, it is not clear how accurate their measures of reading time might have been. Jackson and McClelland (1979) do not detail how they ensured accuracy in recording the reading time of their subjects. Palmer et al (1985) had their readers start reading simultaneously in groups and indicate how much they read in a given time by marking the line number on the text when the time was called. Compared to the precision they pursued in recording their readers' word recognition latencies on computer, their procedures for measuring passage reading speed appear rather crude. Haynes (1989) took a slightly safer approach; she had individual sessions with her Taiwanese EFL readers and clocked each reader as they finished reading each of the target passages. Nassaji and Geva (1999) gave their ESL readers one minute in individual sessions to obtain their reading speed in words per minute. Since multiple passages are necessary for reliability, Jackson and McClelland's (1979) and Palmer et al's (1985) approach seems to represent a realistic compromise.

One more difficulty involved in measuring reading speed at the passage level is the effect of memory. Haynes (1989) had the reading text removed from the readers' sight when they finished reading it and had them answer five MCQs, mostly on details. Remembering details in a text is an important product of reading and is presumably possible only after successful comprehension. However, there can also be a case where one comprehends several points in a passage but loses some of them from memory rather quickly.

An alternative which could minimise this memory effect would be to measure the reading time in smaller units of text, such as individual sentences, and test comprehension after each. Since the latencies in identifying word-level stimuli displayed on the computer screen are considered to exhibit important individual differences in perceptual and psycholinguistic processing (Cunningham et al 1990, Jackson and McClelland 1979, Palmer et al 1985), latencies in reading a sentence on computer screen should also record important individual differences. Unless empirically confirmed, computer-based (CB) reading time data should be treated as separate from reading time for passages printed on paper. Nevertheless, as the population of CB readers grows in the real world, their reading behaviour will likely constitute a significant source of research (cf. Chalhoub-Deville 1999). The present research thus pursued a method for recording the Japanese university EFL readers' CB sentence reading speed.

Shizuka (2000) extensively tested sentence-based reading test items to be used with a Japanese sample on computer and compiled a final set of 40 sentence-based reading items. In Shizuka's CB sentence reading test, each target English sentence was presented on a computer screen at the first subject-paced key press and disappeared at the second, which also prompted four Japanese answer choices. Among them, only one matched the target sentence in meaning. A decision was made to trial Shizuka's 40 items. Since the interest has been on reading speed during comprehension rather than individual differences in comprehension during speeded CB sentence reading, easier items which would create time variance were sought. The 40 items were first piloted on paper to obtain information on their difficulty. The results of this piloting will be presented in Chapter 5.

Operationalising the explanatory variables

L2 vocabulary breadth

L2 vocabulary researchers have developed ways of estimating the vocabulary size of the L2 learners by using a list of words sampled at a certain ratio from different word frequency groups and collecting information on the learner's knowledge of the sampled words. Read's (1997) survey of vocabulary testing identifies the Vocabulary Levels Test (VLT) (Nation 1990) and the Vocabulary Checklist Test (Meara and Jones 1990) as two such testing methods that are better established. The former was considered further, since the latter is a self-report knowledge index and, despite efforts to control for the test takers' over-estimation of their knowledge, its limitations do not seem to be fully resolved (Meara 1996, Read 1997). Nation's VLT has target words sampled from particular frequency levels (e.g. most frequently appearing 2,000) and presents the test taker with a set of decontextualised MCQ items for testing knowledge of word-for-word or word-for-short-phrase correspondences in meaning (Figure 3.1).

Figure 3.1 Sample VLT task

1 business	
2 clock	6 part of a house
3 horse	3 animal with four legs
4 pencil	4 something used for writing
5 shoe	
6 wall	

Since it tests what may be termed ‘sight vocabulary’, which is the person’s knowledge of the target word’s form and its common meaning when only the target word is visually available without contextual clues, it can be regarded as distinct from a test of sentence comprehension ability. VLT remains one of the most popular measures of sight vocabulary size (Nation 2001), and its objective data appear to be growing (Beglar and Hunt 1999, Schmitt, Schmitt and Clapham 2001). At the time of initial instrument preparation for the present study, one version of VLT prepared by Norbert Schmitt (later published in Schmitt, Schmitt and Clapham 2001) was undergoing validation and revision with a large sample of L2 learners including an L1-Japanese subgroup. One of his interim versions was made available to the researcher and contained 30 items for each level of the 2,000 word, the 3,000 word, the 5,000 word, and the 10,000 word, as well as 36 items for the Academic Word level. This test was employed as an initial instrument for measuring the vocabulary breadth of a Japanese sample group.

Since it has not been the primary interest of the present study to estimate our readers’ sight vocabulary size in any absolute sense, and since there has been a practical need for a reasonably compact test which records individual differences in the breadth dimension of the learners’ lexical competence, use of these 156 items in entirety was neither required nor advisable. A decision had to be made regarding the selection of items from the five levels. Through the researcher’s subjective judgment and informal interviews with a small group of university students, the difficulty of the low-frequency words in the 10,000 word level appeared to cause a floor effect and contribute little, if at all, to differentiating among the target Japanese sample group in their breadth of English vocabulary. Thus the 10,000 word level was excluded from the candidate list, which left us with four other levels consisting of a total of 126 items.

On the acceptable number of items for the VLT in norm-referenced testing, the literature shows at least one 60-item version of the VLT, which targeted Japanese corporate employees, has obtained an alpha of .95 and .97 on Rasch case estimate (Beglar and Hunt 1999) and suggests sixty items as a useful point of reference. Since Schmitt’s VLT consisted of sets of 30 items (except for the 36-item Academic Word List set included in this interim version), two existing levels could be conveniently joined together to form a 60-item test (66 if the Academic Word List became a part of it). An alternative would be to select better performing items from among all of the four levels based on an item analysis with a sample resembling the final target population.

A preliminary study was designed to compare all possible combinations of the four VLT levels in terms of their overall reliability as a set. The alternative set of items with better item data was to be compiled as well.

The procedure of this initial analysis will appear in Chapter 4.

L2 word recognition efficiency

Since one aspect of lexical competence separated as a variable in the present study is vocabulary breadth, items for measuring the recognition efficiency dimension of lexical competence should minimise any obvious influence from individual differences in vocabulary breadth. One way of reducing such influence can be attempted by restricting the stimulus words to only high-frequency words. Fortunately, there has been an instrument developed for measuring the speed of processing high-frequency words among a subgroup of EFL readers with a non-alphabetic L1 writing system background. In one of the rare empirical component skills analyses of L2 reading which addressed the speed aspect of orthographic and lexical processing, Haynes (1989) measured her Taiwanese EFL students' visual processing time of words, pseudowords and irregular letter strings and their lexical-semantic processing time of words through speeded stimulus matching, which required the students' speeded decisions on whether a visually presented stimulus pair consisted of two same items or not.

Her visual processing target stimuli (originally developed by Brown, Carr and Chaderjian 1987) were all four letters in length, and the base real words were high-frequency words with a mean frequency of 670 occurrences/million according to Kučera and Francis (1967). The real words formed pairs in which one differed from the other in only one of the four letter-positions, and one fourth of the pairs differed at each of the four serial positions (e.g. line vs. fine, list vs. last, step vs. stop, real vs. read). The pool of real word matching items thus consisted of half of such pairs and half of their counterpart pairs in which the constituent words were identical. The pseudowords were created by replacing one or more of the letters at a position not at the location of difference between two real words of a matching pair in such a way that the resulting strings are meaningless and unfamiliar but orthographically regular and presumably pronounceable (e.g. lize vs. fize). The irregular letter strings derived from permuting the non-critical (thus common) letters in real word pairs so as to form letter strings which violate the orthographic regularities found in the English words (e.g. lnei vs. fnei). The synonym/antonym matching stimuli consisted of high-frequency words taken from West's *General Service List of English Words* (1953) and from Longman's *Dictionary of Contemporary English* that formed either synonym or antonym pairs. To control for the unwanted effect of the familiarity with the matching task, pairs of three-digit numerals were prepared in such a way that each number pair differed in only one of the three digit positions, with each of the three digit positions receiving the distinguishing numeral at equal frequency.

Haynes (1989) prepared 48 stimulus pairs for each of the five stimulus types for paper-and-pencil-based administration. For each stimulus type, the stimulus pairs were printed across three pages with 16 pairs on each that were split into two columns. Symbols 's' (for same) and 'd' (for different) were printed beside each pair for the participant to indicate their judgment by circling one or the other. For each participant, the time required to judge all of the 16 pairs on the page was individually recorded using a stopwatch, and the accumulated total time from the three pages was treated as their processing time for the pertinent stimulus type. The present study employs the stimulus items from Haynes as a source from which to prepare an instrument for measuring the word recognition efficiency of a Japanese sample group.

However, the paper-and-pencil-based methods such as Haynes's (1989) clearly offer less accurate measurement compared to CB methods, which have often been adopted for L1 readers' latencies associated with word recognition in L1 componential studies (e.g. Cunningham et al 1990, Jackson and McClelland 1979, Palmer et al 1985). Few componential studies of L2 reading ability (except perhaps Brown and Haynes 1985, which does not detail the methodological procedure) have attempted the same degree of rigour in recording the latencies in processing various types of Roman letter strings. The current study attempted to improve the accuracy of Haynes's latency measurement by the use of CB methodology.

Computer-based measurement of word recognition time enables presentation and latency measurement of each target item for each test taker in the unit of milliseconds. This in turn significantly reduces the necessary number of items compared to the paper-based (PB) counterparts like Haynes's (1989), which required 48 items in each stimulus category to record meaningful variance in the sample. Also, software programs for the CB measurement give the researcher an option of the random sequencing of the items to be presented from across any stimulus category for each test taker. This enables mixed sequencing of the three visual letter string items, thus merging of the three separate tests into one, as well as further reduction in the k size. To illustrate the test taking procedure, the first subject-paced key press would present the first target item on the computer screen which might happen to be 'lnei fnei' (unmatched irregular strings), to which the subject should respond by pressing a designated key for 'different'. The second item might happen to be 'step step' (matched real words), to which the subject should press another key for 'same'. The third one might be 'lize lize' (matched pseudowords), and the fourth one 'real read' (unmatched real words), etc. The computer records the sequence of item presentation and latencies for the individual items as each test taker proceeds through the test. The random sequencing of the items should help isolate the order effect as well.

For the present study, an initial pilot test was prepared using the software PsyScope (Cohen, MacWhinney, Flatt and Provost 1993). Twelve each of real word, pseudoword, and irregular letter string items (a total of 36 visual processing items) and 36 synonym/antonym pairs were chosen from Haynes

(1989) to form an initial set of word recognition items. Attempts were made to exclude real word items that the researcher judged as less familiar to the Japanese students and to vary the positions of the distinguishing letters in visually unmatched pairs. A number matching task was also prepared to obtain data on their latencies that are unrelated to alphabetic processing, which can be treated as a covariate. The preparation of the materials followed the same principle by which each number pair differed in only one of the three digit positions and each digit position received the distinguishing numeral an equal number of times.

It should be noted that under the broad category of word recognition efficiency are in fact the separate variables of (1) real word recognition, (2) pseudoword recognition, (3) irregular letter string recognition, (4) lexical familiarity effect, (5) orthographic regularity effect, and (6) synonym/antonym recognition. Unless specific references are made to these sub-categories, word recognition efficiency will refer to all of the above.

L2 phonological awareness

One of the most comprehensive lists of phonological awareness measures is found in Stanovich et al (1984). To a group of child L1 readers, all of their ten tasks orally presented the key words and/or some word choices. The tasks tested the children's ability: to supply the words rhyming with the key word; to choose among a list of words the one that rhymes with the key word; to repeat the key word and choose among a list the word that had the identical word-initial consonant with the key word; to choose among a list the word that had the identical word-final consonant with the key word; to pronounce the key word without the word-initial consonant; to pronounce the key word substituting the word-initial consonant; to choose among a list one which has a different word-initial consonant from the rest; to choose among a list the word that does not share the word-initial consonant with the key word; to choose among a list the word that has a different word-final consonant from the rest; and finally to identify in a word pair the word-initial consonant which is present in one word but missing in the other. Most tasks correlated moderately with each other and with a reading ability measure, though the rhyme tasks and the initial consonant substitution task showed weaker correlations within themselves and with the other measures. Among some recent developmental reading studies, Cormier and Dea (1997) adopted the Test of Auditory Analysis (Rosner and Simon 1971), whose tasks are similar to some tasks in Stanovich et al, while Leather and Henry's (1994) were almost directly modelled after some tasks by Stanovich et al. With Spanish–English bilingual children, Durgunoğlu et al (1993) used segmenting, blending, and matching tasks in Spanish: the segmenting task involved dividing syllables into phonemes and words into syllables; the blending task involved blending phoneme pairs into syllables, syllable pairs into words, onsets and rhymes into words, and phoneme triples or quadruples into words; and the matching task involved choosing among the list the word which shares the initial phonemes with the key word. With adult speakers of Chinese, which uses a non-alphabetic script, Read, Zhang, Nie and Ding (1986) adopted a task requiring their subjects to delete or add a single consonant at the beginning of a spoken syllable. Koda's (1998) study exploring the relationship between phonological awareness and ESL reading comprehension ability among the L1-Chinese (non-alphabetic) and L1-Korean (partially alphabetic) readers adopted the Wepman auditory discrimination task, the phonemic substitution task, the phonemic deletion task, and the phonemic insertion task. The Wepman auditory discrimination task asked the subjects to determine whether each of the auditorily presented word pairs are phonologically identical or only partially overlapping. The phoneme substitution task involved listening to pseudowords with one wrong phoneme in each and finding in the printed list of correct words the substituted phoneme. The phoneme deletion task involved listening, repeating, then pronouncing the given words removing a specified phoneme. The phoneme insertion task involved listening to pseudowords and finding in the list of printed correct words where the unnecessary phonemes were inserted. Wade-Woolley (1999) adopted a phoneme deletion task similar to Koda's.

Since phonological awareness is considered to be 'a constellation of skills' (Stanovich et al 1984) it should ideally involve as many of the identified subskills as possible. In researching the relatively new field of L2 readers' phonological awareness, however, such multi-skill measurement has been rarely achieved. The largest found in the literature of four tasks adopted in Koda (1998) came closest to the ideal and would have qualified as the initial set on which to perform trialling with a pilot group. However, her items did not become available in the end.

In view of the results in Wade-Woolley's (1999) study that the Japanese ESL readers seemed to depend less on phonological coding than orthographic processing and the dire fact of having to start developing multiple measures of phonological awareness, a decision was made to drop this variable from further consideration in this study.

L2 working memory

Verbal working memory, which is conceptualised as the dual functions of verbal information processing and short-term memory storage, has been operationalised through the Sentence Span Task developed by Daneman and Carpenter (1980). The task is considered to tap one's ability to process multiple sentences while maintaining parts of them in short-term storage, and it presents the subjects with sets of sentences to process in sequence with the concurrent requirement of remembering all the sentence-final words for a set-final recall. Depending on the version of the Sentence Span Task, the sentences are presented either visually (Reading Span) or auditorily (Listening Span). To ensure processing rather than simply concentrating on the storage of the sentence-final words, some versions asked the subjects to somehow indicate their sentence comprehension or perform a sort of sentence verification, e.g. by mixing some untrue statements or permuting some constituent words. The size of the sentence set gradually increases, from two to six in many studies, and Daneman, Carpenter, Just, and colleagues (Daneman and Carpenter 1980, 1983, Just and Carpenter 1992, King and Just 1991, MacDonald, Just and Carpenter 1992, Miyake, Just and Carpenter 1994) have operationalised the working memory span as the largest size of the sentence set for which the subject was able to recall all the sentence-final words. Others (Baddeley et al 1985, Cunningham et al 1990, Dixon et al 1988, Harrington and Sawyer 1992, Masson and Miller 1983, Rankin 1993, Turner and Engle 1989) have summed the number of words recalled, often requiring the correct sequence within each set. A slightly modified version of the Sentence Span Task (Siegel and Ryan 1989) requires the subjects to listen to the sentences with the final words left out for them to supply. The subjects thus remembered, for later recall, all the sentence-final words which they supplied themselves for sentence completion. This format has been employed in such studies as Geva and Ryan (1993), Leather and Henry (1994), Siegel (1994), and Da Fontoura and Siegel (1995). The Sentence Span Task for the Japanese EFL learners will have to take account of their incomplete linguistic resources and limited processing efficiency. Complex target sentences, low-frequency words, and demanding processing load are likely to cause floor effects (invariantly poor group performance due to difficulties of the task, which provides little information about the participants' individual differences). Siegel and Ryan's (1989) task required contextual word prediction and production skills and can be considered to be comparatively more demanding. Studies exploring the verbal working memory of adult L2 learners (Harrington and Sawyer 1992, Osaka and Osaka 1992) have used somewhat shorter and simpler sentences than Daneman and Carpenter's (1980) original version, which was intended for an American university student population. Harrington and Sawyer's ESL version of the Reading Span Test (RST) included a sentence verification task requiring the test taker to indicate whether the sentence was grammatical or not. Although their RST did correlate with reading abilities of their advanced learners of English, it has been the researcher's judgment that their sentences can be challenging and, together with their inclusion of concurrent verification task, can present problems to many less advanced Japanese EFL readers and result in a floor effect. Osaka and Osaka developed their ESL version of the RST, which correlated significantly with Daneman and Carpenter's original version when their advanced

Japanese learners of English took both tests. Osaka and Osaka's ESL version was made available to the researcher and judged to be less challenging than Harrington and Sawyer's version. Therefore, a decision was made to include this test as one of the pilot instruments. The procedure for this test will be detailed in 'Preliminary Study E' (pp. 90–94).

L2 syntactic knowledge

As mentioned in the section on vocabulary knowledge, one's lexicon has to contain information on the syntactic behaviours of each of the words, making it difficult to completely dissociate between the two constructs of lexical knowledge and syntactic knowledge. It is, in fact, simply impossible to completely remove the effects of lexical knowledge if one is to measure someone's syntactic knowledge, since it presupposes at least some awareness of word categories which depends on knowledge of word forms, a part of the lexical knowledge.

In operationalising syntactic knowledge for the purpose of the current research, it is vital that, whatever category of words is involved as part of the measurement process, the syntactic rules in question be as independent of extremely word-specific knowledge as possible, since in as much as the test items focus on word-specific aspects of syntax, they deviate from sampling the more pervasive features of syntax and measuring the learner's general syntactic knowledge, which is of more interest to the current study.

One more rather complex issue to be considered concerns the elements of reading involved in a measure of syntactic knowledge. As Urquhart and Weir (1998) point out, what claims to be a test of grammar or syntax should be precisely that and as separate from any tests of other constructs as possible, and the overlap in the behaviours sampled in what is meant to be a test of syntax with those in the test of reading must be minimised. This is especially important if the research is concerned with any relationship between the two variables (little would be revealed from a significant correlation between the scores on a reading comprehension test and those on a syntax test which required more semantic than syntactic processing of the test sentences). Such a requirement is not easy to meet (Alderson 1993). Indeed, most measures of L2 syntactic knowledge would involve processing of visually presented text, which may in itself be judged as a kind of reading. Also, the more meaning extraction is integrated in the task and the more contextualised the task becomes, the more it seems to include the characteristics of reading. It appears then that to achieve independence from a test of reading, a test of syntactic knowledge should minimise the elements of semantic processing and contextualised discourse or sentences. Two types of testing of syntactic knowledge appear to meet such a requirement better than the others.

One is the type which has been adopted in the Structure and Written

Expression section of TOEFL. Each item visually presents a sentence with a part of it replaced by a blank, and answer choices are offered which may have similar semantic contents but only one satisfies the syntactic constraints imposed by the structure of the rest of the sentence as in the example below (Figure 3.2).

Figure 3.2 Sample Structure and Written Expression item

Conifers first appeared on the Earth ______ the early Permian period, some 270 million years ago.

A. when
B. or
C. and
D. during

In the context of evaluating TOEFL as a test of communicative competence, Bachman judged the Structure and Written Expression items of this particular version of TOEFL to 'require syntactic competence almost exclusively' and to be 'largely context-reduced' (1986:78; the items are reproduced in Duran, Canale, Penfield, Stansfield and Liskin-Gasparro 1985). All 15 items from this version of TOEFL are further considered as possible candidates.

Another method of testing syntactic knowledge which may satisfy the condition above and is sometimes used in L2 reading studies (e.g., Lee and Schallert 1997) is the grammaticality judgment task, in which the learner judges whether or not each target sentence conforms to the grammar of the language. In spite of its content validity, unfortunately, this method generally suffers lower reliability compared to the Structure and Written Expression type method, due to the much higher likelihood of accidentally guessing the answer correctly ('Grammatical' or 'Ungrammatical'). The results would have to be interpreted only in terms of whether the performance was significantly better than chance (50%). It was therefore determined that an initial instrument for measuring Japanese EFL groups' syntactic knowledge be prepared based on the 15 Structure and Written Expression items from the version of TOEFL quoted above (Duran, Canale, Penfield, Stansfield and Liskin-Gasparro 1985).

For reliability, at least twice as many items seemed necessary. Since many average Japanese students find TOEFL items relatively difficult, merging an additional 15 items from another TOEFL paper was judged inappropriate. One test developed in the British context which includes grammar items of the same format is the Test of English for Educational Purposes (TEEP) (Weir 1983). Since TEEP's entire grammar items and their facility and corrected item-total correlation values were available and, from the researcher's initial content observation, they appeared somewhat less difficult for the

Japanese university population than the 15 TOEFL items above, a decision was made to examine its 60 items further for possible inclusion in the syntax instrument.

TEEP's 60 grammar items were reviewed first in terms of their content. As discussed above, testing of extremely word-specific knowledge should be avoided since such items will make the test overlap in content with a vocabulary test. An initial inspection of the original TEEP by the researcher led to identification of items which could be recategorised into a vocabulary section of a test rather than in a grammar section such as below (Figure 3.3).

Figure 3.3 Sample unsuitable syntax item

The library (   ) nearly two million volumes.

A. consists
B. compares
C. composes
D. contains

This item seems to be testing knowledge of four (or possibly even fewer) particular verbs' syntactic behaviour. Although there is a focus on the syntax, one can also argue that the success on this item depends more on one's knowledge of the specific lexical item 'contain' and of the distractors. This illustrates the difficulty of separating the tests of syntax and lexis. Items of this nature were excluded from the item candidate pool for the syntax instrument for the present study. Another danger, also aforementioned, of an item actually tapping semantic comprehension rather than syntax, was not obvious among the given TEEP items.

Since one of the practical considerations in the selection of TEEP syntax items was the inclusion of items that were somewhat easier than the 15 TOEFL items for the Japanese EFL population, the facility values of the TEEP items were consulted (Weir 1983), with the fact in mind that the recorded figures must be attributed not only to the test items themselves but to the specific test population. The item facility values ranged between .32 to .94 and the corrected item-total correlation between .11 to .64. Items that were below .25 on corrected item-total correlation were removed first. Since relatively easier items were sought at this point, a strict value of .40 on item facility was set as the criterion for rejecting difficult items. From among the items that remained after the screening process described here, 20 items were selected as an addition to the 15 TOEFL items to form an initial 35-item syntax knowledge measure. To take the matter of content validity further and test the researcher's personal judgment objectively, a small validation study was conducted on these items, and its results will be presented and discussed in the section on 'Preliminary Study B' (pp. 59–64).

L2 listening comprehension

Although two studies (Brown and Haynes 1985, Haynes 1989) have demonstrated that L2 listening comprehension ability can account for L2 reading comprehension ability, when the results were partialled out for the Japanese learners of English, this did not apply at all. When Brown and Haynes calculated the listening–reading correlation for each of the Spanish, Arabic, and Japanese L1 subgroups, they found that significant correlations emerged only within each of the Spanish ($r=.56$, $p<.01$) and the Arabic ($r=.57$, $p<.01$) L1 subgroups. The L1-Japanese group showed much lower correlation ($r=.12$, n.s.). Since their Japanese subgroup had outperformed the Spanish and Arabic speakers in reading-related skills, Brown and Haynes speculate that the traditional emphasis in secondary EFL education in Japan on literacy skills might have caused their reading ability to outpace their listening ability. This view is thoroughly shared by the researcher and possibly by many of those involved in ELT in Japan. Whatever the underlying factors, the Japanese readers' data above prevent one from assuming that the primacy of listening observed with Haynes's Taiwanese group would recur with a Japanese population. In fact, Brown and Haynes's Japanese data cancel out the interest created by Haynes (1989) in including listening comprehension as an explanatory variable with a study of Japanese EFL readers. Given the outlook that there would be already a number of significant variables and a lengthy battery of individualised and group-administered testing, listening comprehension was dropped from further preparation of the study at this point.

Language-independent metacognitive knowledge of text and reading

Since the present study is most concerned with knowledge and skill areas in L2 that may account for the individual differences in the Japanese EFL readers, obtaining information about the participants' text processing skills in L1 is treated as rather supplementary. As is the case for listening comprehension, including another reading comprehension test would adversely affect the feasibility of data collection on other significant L2 variables. Thus, a possibility was explored of making use of an existing questionnaire which would not lengthen the total 'test' battery but would capture individual differences in knowledge about or skills in reading in L1.

Although it was initially difficult to find any such instrument, a working instrument became available in a later stage of this study. Research into the abilities of L1 Dutch and L2 English reading and writing in the Netherlands conducted by Schoonen and colleagues (e.g., Schoonen, van Gelderen, de Glopper, Hulstijn, Simis, Snellings and Stevenson 2003) employed a questionnaire to assess the students' language-independent metacognitive knowledge about the text, reading, and writing in general, which predicted their learners' reading comprehension abilities (van Gelderen, Schoonen, de Glopper, Hulstijn, Snellings, Simis and Stevenson 2003). From an English translation of their set of 80 yes/no questions, items clearly focusing on writing skill and those reported to show poor psychometric characteristics by the questionnaire developer (Rob Schoonen, personal communication, 2001) were removed. The remaining items on metacognitive knowledge of the text in general (k=14) and on reading strategies (k=24) were translated into Japanese and adopted as a single measure in the main study. Time constraint due to the relatively recent appearance of this instrument prohibited preliminary testing prior to its use in the main multivariate study.

Revised research questions

The initial set of research questions (cf. Chapter 2) had been formed purely from the theoretical perspectives gained through the review of the previous research on reading and related skills of L2 learners in general. A survey on the research instruments for the specified sample population with some key considerations on methodological feasibility helped to define the domain of current research more clearly. The revised research questions resulting from it are thus stated as follows:

Q1: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills − L2 vocabulary breadth, L2 word recognition efficiency, L2 working memory span, L2 syntactic knowledge, and language-independent metacognitive knowledge about the text and reading − account for the individual differences in L2 careful passage reading comprehension ability of Japanese EFL learners?

Q2: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in L2 careful passage reading comprehension ability of Japanese EFL learners as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

Q3: To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners?

Q4: Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

Approach to the research questions

The research questions revised above are approached along the general methodological direction outlined in the beginning of this chapter. Namely, the degree to which the specified skill areas account for individual differences in the reading abilities of Japanese EFL readers are examined by the use of multiple regression analysis. Calculations of Pearson product-moment correlation for all the combinations of variables address the subsidiary questions under the first research question (cf. Chapter 2), e.g., about the relationships between vocabulary breadth and lexical access speed or between lexical knowledge and syntactic knowledge. The explanatory power of the predictors is evaluated in terms of the amount of variance they account for in reading in relation to the other variables. The primary focus of this study is careful passage reading comprehension, and it is analysed first as the criterion variable with the explanatory variables.

The second research question – the differences in the extent to which the independent variables explain the L2 careful reading variance as a function of the L2 careful reading skill level – is explored after dividing the entire sample into two subsections of readers with different scores on the criterion reading comprehension measure.

This is followed by a separate multiple regression analysis with CB sentence reading speed as the criterion variable to answer the third research question – the degree to which the subskill areas account for individual differences in the CB sentence reading speed of the Japanese EFL readers.

Finally, separate analyses are performed on the two subsections to compare the results between the two.

Additionally, to explore the possibility of identifying latent variables underlying the intercorrelations among the observed variables, the data are subjected to exploratory factor analyses. Consulting the resulting factor structure, plausible models of the relationships among both observed and latent variables are specified and tested through confirmatory factor analysis.

Preparation for the main study

The current research started with a series of small scale preliminary studies, each involving a small number of variables prior to the full multivariate study. These preliminary studies were conducted with groups of Japanese EFL students at the tertiary level and served two purposes.

The first purpose was to evaluate the suitability of the measurement instruments for the Japanese university sample and to modify them based on the item analyses as necessary. To reduce the test battery for the final multivariate study down to a manageable size without compromising the reliability of each measure, the items which would clearly undermine the overall reliability of the measure were to be removed or replaced. Whenever appropriate, the measurement instruments were examined in terms of their internal consistency as estimated via Cronbach's alpha. The individual items were also subjected to analysis on item facility as measured via item score mean and on item discrimination as measured via point-biserial item-to-total correlation. Rasch based analyses were also performed using Quest (Adams and Khoo 1993) which yielded the item and case reliability estimates as well as the fit statistics for the items and the cases. Misfitting cases were to be checked first as signs of unusual test taking behaviour. Items with extreme facility, discrimination, and fit values are considered problematic (Henning 1987, McNamara 1996). Since high homogeneity of the sample can also lower the reliability coefficients and each Japanese university tends to attract students with similar English proficiencies, reliability of a measure must be carefully judged in light of the sample distribution. In addition to the reliability analysis of the measures, one instrument, namely the syntactic knowledge test, also underwent a small content validation study. This is because the separation of the syntactic knowledge construct was vital and the initial syntax items originally derived from loosely defined structure and grammar sections of existing tests.

The second purpose of the preliminary studies was to informally test the research value of each predictor variable separately, based on its correlation with reading. For this reason, every attempt was made to administer the careful reading test with one predictor instrument at this preliminary stage of the project. A Pearson product-moment correlation for each subgroup between reading comprehension performance and subskill measure was calculated. Considering the possible demand a large test battery would impose on research participants, the final set of predictor instruments had to be as parsimonious as possible. Therefore, if a predictor variable did not exhibit meaningful correlation with reading after repeated preliminary studies, it was to be reconsidered and possibly excluded from the final test battery. However, since low correlation can result not only from lack of relationship but from imperfect measurement instruments, and since the measurement instruments were still undergoing refinement at this preliminary stage of the project, conservative decisions were recommended. Each of the preliminary studies will be described in more detail in Chapters 4 and 5.

Data collection

In both the preliminary and the main studies, data were collected at Japanese educational institutions at the tertiary level by the researcher himself and by co-operating university instructors who found participants from among their students. Data for the preliminary studies and the main study come from students at universities in western Japan. The selection of the universities was based on availability of participating EFL instructors rather than on principled random sampling.

Most paper-based (PB) tests were group-administered as part of English language class activity although, in some cases, group or individualised sessions of tests were scheduled outside of class hours. Whenever possible, the researcher was present at the site of PB testing to ensure equal administration conditions across different institutions and classes. Where this was not possible, a printed proctor procedure was included with the test materials and any unclear points were clarified prior to the administration. Information on late arrivals of participants and unexpected conditions, such as a participant falling ill during the test administration, were reported so that their papers could be treated separately or removed from further processing.

The instruments for word recognition efficiency, working memory span, and CB sentence reading speed required individual sessions with the participants. These sessions were thus scheduled outside of the class hours. Students who participated in the PB testing above were invited to attend these additional individual sessions for which payment was offered. The sessions were individually scheduled with those who responded to this call for participation and were held in available university classrooms or office rooms. The researcher himself conducted most of the individual sessions as the proctor. However, when concurrent sessions had to be scheduled or the researcher could not be available, trained and paid proctors conducted the sessions as well. The procedure for each of the individualised tests was illustrated on paper for the test taker and oral explanations were added at the beginning of each individual session.

Regarding the individualised tests, the CB tests were prepared and implemented using the software PsyScope (Cohen et al 1993) on Apple Macintosh PowerBook computers. Response time and decision data were recorded on the computer as programmed on PsyScope. The working memory recall performances were recorded by the proctor on the recall check sheet. Both correct and incorrect recalls were recorded. Any unexpected events during the individual sessions were recorded.

The main study also employed a questionnaire to elicit information about the participants, including their exposure to English in terms of the length of stay in an English speaking country or region which might have led to misfitting data. The questionnaire was usually filled out at the participant's leisure between the weekly classes but sometimes group-administered during the class.

The participants were presented with a consent form which they were to sign if they accepted the use of the information they provided on the tests and the questionnaire for the present research.

Data processing

Along with the participants' identification numbers, their responses to the PB MCQ items were manually entered into spreadsheet files either directly on a spreadsheet application or via a database application, after which the entered responses were checked against the answer key to generate a table of 1/0 results, where 1 represented success on the item and 0 failure. The reading span recall was also manually entered into spreadsheet files from the recall check sheets. Each recall item was represented by 1/0 notation as well. Data from CB tests were first recorded on text format but were converted into spreadsheet files. Questionnaire data were initially entered into a database file but were converted into a spreadsheet format as well.

After this initial preparation, analyses of the data were performed on the applications SPSS for Windows version 10, Quest (Adams and Khoo 1993), and AMOS version 4 (Arbuckle and Wothke 1999) depending on the types of analysis required.

Chapter review

This chapter has presented revised research questions after specifying the prospective sample population and operationalising the research variables on the basis of both theoretical significance and research feasibility. The study will employ ability measures relying on both paper- and computer-based methods and explore the extent to which various subskills account for the individual differences in the PB careful passage reading comprehension and CB sentence reading speed of the Japanese EFL students at the tertiary level. It will also investigate whether different patterns of results would emerge depending on different reading ability levels.

4 Preliminary studies: phase 1

Chapter overview

As explained previously, the present research programme includes a series of preliminary studies conducted prior to the main study. They serve the purposes of determining the instruments that are usefully employed for the measurement of the Japanese EFL readers' abilities and refining such instruments. This chapter reports on three such studies conducted in the initial stage of the project. The first study (Preliminary Study A) examined the appropriateness of the initial version of the passage reading comprehension measure, whereas the second study (Preliminary Study B) is based on the content analysis of the measure of syntactic knowledge. The third (Preliminary Study C) analysed a relatively long vocabulary test to generate a more concise but equally effective measure of Japanese EFL learners' vocabulary breadth. The following sections will describe the three preliminary studies in sequence.

Preliminary Study A: testing passage reading comprehension (version 1)

Introduction/purpose

The first of this series of preliminary studies aimed to obtain the psychometric qualities of the main criterion instrument of careful passage reading comprehension so that it could be evaluated for its adequacy for a Japanese university sample.

Method

The initial version of the passage reading comprehension test was subjected to reliability analysis using Cronbach's alpha. Item facility and discrimination values of the individual items were examined. Rasch analysis using Quest (Adams and Khoo 1993) yielded fit values for both the individuals and the items for checking if they fit a model of response pattern. Logit values obtained from the Rasch analysis were used to see how well the levels of item difficulty and the candidate ability matched.

Participants

A total of 101 students from two universities and a junior college took the passage reading comprehension test. They were first to third year English majors mostly belonging to the age range 18 to 21.

Procedure

The test was group-administered by the students' instructors in their EFL and related classes. The test rubric was printed in the students' L1 on the beginning page of the test paper. Forty minutes were allowed to complete the answers, and no passage-specific time limit was imposed. The students were told they should not attempt to guess blindly when they have no idea what the correct answer might be.

Results

The mean raw score of this sample group for this 20-item passage reading test was a low 7.54 with an SD of 2.84. The overall reliability coefficient as estimated via Cronbach's alpha was also a low .47. The item facility value ranged from .17 to .68 with a mean of .38, and point-biserial corrected item-total correlation ranged between .02 to .27 with a mean of .14.

Next, the mean of the facility and the discrimination indices were separately obtained for each passage and summarised in Table 4.1. The mean facility value for the passages ranged from .26 to .43 and item-total correlation mean from .08 to .15. Three of the passages had items below a success rate of .33, an indication of excessive difficulty. Passage 1 had one, Passage 2 had two, and Passage 4 had four. Regarding item-total correlation, only one item each from the first three passages had the recommended value of over .25 (Henning 1987).

Table 4.1 Average item statistics for the test passages (n=101)

	Facility	Discrimination	Alpha if deleted
Passage 1	.43	.15	.46
Passage 2	.36	.14	.46
Passage 3	.42	.19	.45
Passage 4	.26	.08	.47
Overall	.38	.14	.46

Rasch analysis yielded an overall reliability estimate of .86 for the items and .53 for the cases. Infit Mean Square values ranged from 0.92 to 1.08 for the items and from 0.79 to 1.38 for the cases. A total of four cases, out of

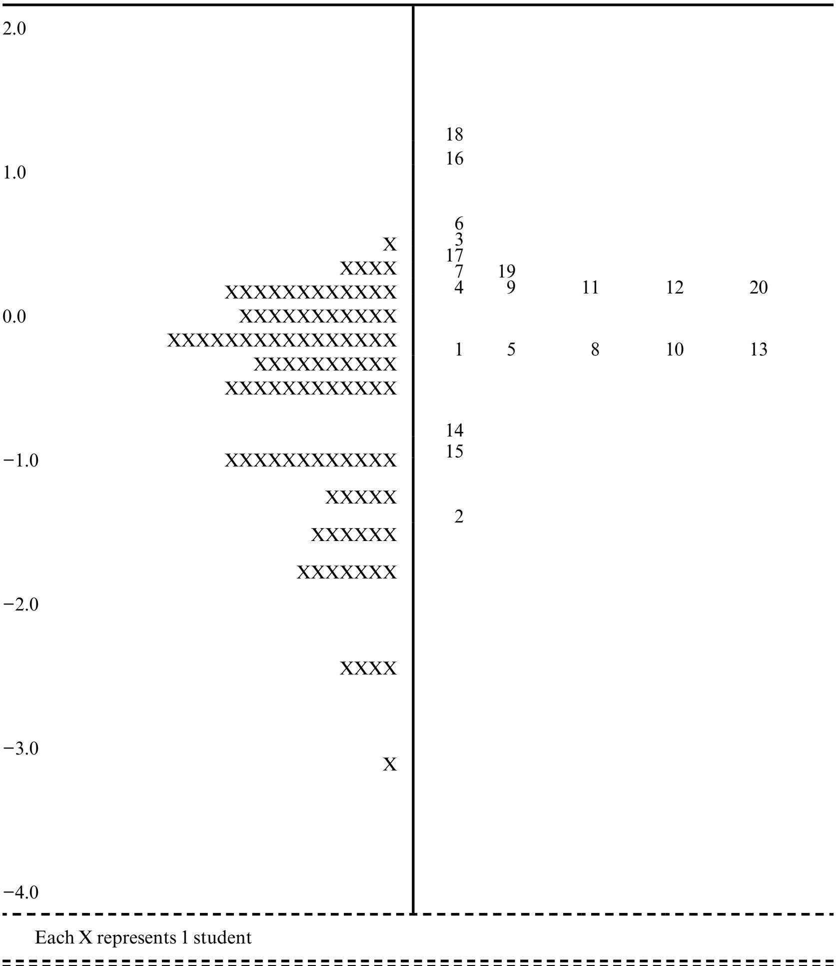
101, received an Infit Mean Square value exceeding 1.3, possibly representing cases of test-person misfit (McNamara 1996). The logit mean for item difficulty was 0.00 with an SD of 0.55 and a range of between −1.39 to +1.10. Person ability estimates yielded a lower logit mean of −0.61 with an SD of 0.76 and a range of −3.11 to +0.67.

Discussion

Overall, this passage reading comprehension test was found to be quite difficult for the Japanese sample group that took part in this initial piloting and failed to record a sufficient level of reliability. Item facility was low on many of the items and item discrimination was also low on most. The item-to-person mean discrepancy in logit estimates indicates a level mismatch which can be confirmed graphically in the item-to-person map in Figure 4.1. More of the test takers on the left-hand side are placed lower than the zero logit line indicating relatively low person ability while many of the items are plotted above the same line on the right-hand side indicating relatively high item difficulty. The low level of reliability is probably associated with the lack of variance caused by the passages and/or the items being too difficult. This pilot sample group consisted of English majors and included a third year subgroup and yet this reading test presented problems to many of these participants. Since the final target sample in the main study would additionally include less proficient non-English majors, continued use of this version of the passage reading test would simply result in a floor effect and fail to record any meaningful correlations with other variables. An alternative had to be pursued. Use of less difficult passages should result in a wider spread of performance and increased reliability.

Instead of replacing the reading test in its entirety, however, replacing parts of the initial version appeared a more economical but plausible option. Since none of the items had a problematic Infit Mean Square value and most had a problematic item-total correlation, the passages were primarily compared in terms of their item facility values. Passage 4 had the lowest facility mean of .26 with four out of five items below 33% in success rate and was clearly the most difficult of the four passages. Passage 2 followed with a mean of .36 and two items falling below 33% in facility. The other two passages, Passage 1 and Passage 3, had facility means of .43 and .42 respectively and appeared to have caused less difficulty with the Japanese group. Since the overall facility needed to improve to match the ability of the target sample, it was decided that the two most difficult passages be replaced by two new passages.

Figure 4.1 Plot of person ability and item difficulty estimates from reading comprehension test, version 1



Conclusion: toward a second version of passage reading comprehension instrument

A search for alternative passages identified an existing reading comprehension test with a known record of having functioned with an Asian EFL population sufficiently well. Lee and Schallert (1997) reported a use of a 4-passage reading comprehension test in their investigation of L2 reading threshold among a Korean secondary school population. Their question format was the same as the present study's in that five MCQs followed each of the reading passages. All of their items required an inferential understanding of the passages rather than a mere search for particular lexical items or simple word level paraphrase, and they had been selected from among a larger pool of passages based on item discrimination in their pilot studies. They reported a split-half reliability of .84 for this 4-passage test. Since Lee and Schallert's was intended for use with a secondary school population and the present study targets a tertiary level population, the two longest and most challenging of their four passages were selected and combined with the two passages from the initial version of the reading test. One passage discussed a species of ant known as Amazon ants and the other passage described the discovery of penicillin. This formed a second version of the passage reading comprehension test for the present study. Since the added passages were shorter and expected to be easier than the existing two passages, they were placed in the beginning of the test and named Passage 1 and Passage 2, while the two passages from the initial version of the reading test were renamed as Passage 3 and Passage 4 from their original passage numbers of 1 and 3 respectively. Table 4.2 summarises the four passages in the second version of the reading comprehension instrument.

Table 4.2 Passage information reading comprehension test: version 2

Passage	**1**	**2**	**3**	**4**
Topic	Amazon ants	Discovery of penicillin	Failure of Skylab	Changing banking practices
Length (Words)	82	79	251	261
Flesch Reading Ease Score	81	54	61	54
Flesch Kincaid Grade Level Score	8	10	14	15

Preliminary Study B: analysing the content of the syntactic knowledge measure (version 1)

Introduction/purpose

The rationale and the initial procedure for preparing the syntactic knowledge measure by combining 15 past TOEFL items and 20 past TEEP items were explained in Chapter 3, and the difficulty of distinguishing the syntactic knowledge from reading at the level of instrumentation has been discussed as a source of concern for research exploring the relationship between syntactic knowledge and reading. This second preliminary study was designed to subject the initial 35-item version of the syntactic knowledge measure to judgment by multiple raters to determine if it is suitable for use as a measure of syntactic knowledge in the present study.

Method

In preparing the instruments for this second preliminary investigation, a rating instrument developed and employed in Bachman, Davidson, Ryan and Choi's (1995) content analysis of TOEFL and Cambridge examinations was first considered. Bachman and colleagues were attempting to explore, among other things, which of the 12 components of Communicative Language Ability (CLA) were involved in correctly responding to the items in question, and the target of their analysis subsumed the Structure section of TOEFL, from which the syntax measure for the present study is adopted.

Their rating scale was designed to elicit from expert judges at least two types of information on each item simultaneously: (1) the degree to which a given ability component is involved and (2) the level required in that ability. Combining these types of information, they produced the scale format shown in Table 4.3.

Table 4.3 Bachman et al's rating scale for content analysis (1995:102)

Not Required	Somewhat Involved	Critical Basic	Critical Intermediate	Critical Advanced
0	1	2	3	4

When the 40 Structure and Written Expression items of the TOEFL were judged against this scale, the mean ratings on the 12 CLA components were as reproduced in Table 4.4.

It is obvious from Bachman and colleagues' five ability components receiving zero means and zero variance that these abilities were regarded as unrelated to the success of the items in this section. In fact, it can be inferred from the other means and SDs that only their Lexicon, Syntax, and Phonology/Graphology have been judged 'Critical' more often than 'Not Required' or 'Somewhat Involved'. Among these three ability components, Phonology/Graphology had a mean of 2.00 with no variance suggesting that a basic ability in this component was judged to be required for all of the 40 items by both of the two experts, but Phonology/Graphology received the same rating of 2.00 with no variance throughout their analysis of TOEFL and Cambridge FCE including Listening, Vocabulary, and Reading. A very plausible explanation of their judgment is that basic phonological and graphological competence is invariantly necessary regardless of the section of the tests. Since our interest

Table 4.4 Bachman et al's rater responses on 40 TOEFL Structure and Written Expression items (1995:107)

CLA Component	Mean	SD
Lexicon	2.20	1.05
Morphology	1.05	1.21
Syntax	2.84	0.51
Phonology/graphology	2.00	0.00
Cohesion	0.46	0.89
Ideational functions	0.50	0.50
Manipulative functions	0.00	0.00
Heuristic functions	0.00	0.00
Imaginative functions	0.00	0.00
Sensitivity to dialect/variety	0.00	0.00
Sensitivity to register	0.00	0.00
Strategic competence	1.00	1.01

is in the ability components that may or may not be involved in a test item, among Bachman and colleagues' abilities only Lexicon and Syntax emerge as worthy of rating. In addition to these two abilities, as discussed earlier, the involvement of sentence semantic processing required judgment in our analysis especially because the present study involves an investigation into the role of syntactic knowledge for reading comprehension, which might be argued by some to be synonymous with sentence semantic processing.

The content analysis thus focused on the three ability components of syntactic knowledge, lexical knowledge, and sentence reading comprehension. Bachman and colleagues' format of 5-point scale rating needed modification to avoid the interpretation problem which was caused by treating separate types of information on a single continuum. Instead of having two judges give ratings on a 5-point scale to each of the abilities for each item as they did, it was judged more appropriate, at least for the purpose of the present study, to have more judges give categorical responses on whether an item primarily tests the candidate's knowledge of syntax, lexical semantics, or sentence semantics. Items judged by more raters to be testing the two abilities other than syntax would have to be considered problematic. The instrument thus asked the raters to indicate their judgment, on each item, of which of the following three it measures, and the raters were asked to restrict their answers to a single choice on each item:

(1) the candidate's 'knowledge of the meanings of certain words and phrases' (lexical-semantic knowledge)
(2) the candidate's 'knowledge of sentence structures and that of acceptable sequences and forms of words in terms of syntax' (syntactic knowledge)
(3) the candidate's 'understanding of the meaning of the overall sentence' (sentence reading comprehension)

Participants

A group of 11 L1-English ELT experts with at least a master's degree in applied or theoretical linguistics or TEFL participated in the study at a British university, while three Japanese lecturers of English syntax with at least a master's degree in linguistics from three universities in Japan participated in the study in Japan.

Procedure

At the British site of data collection, ELT experts were contacted through the co-ordinator of a university EAP course in which they worked as EAP instructors, and they were invited to participate in the study by completing the ratings at their leisure. Eleven anonymous ratings were collected from this group. A group of Japanese scholars belonging to a circle of English linguistics were contacted through one of its members and three ratings, completed at the judges' leisure, were returned via email. The judgments of the raters from the two groups were combined and processed for frequency counts of the three ability components for each item.

Results

Table 4.5 presents a summary of the judges' responses. Each figure represents the number of 'votes' for the ability regarding each item. After examining Item 1, for instance, 12 judges thought it was testing syntactic knowledge more than the other two abilities, while two judges thought the same item was testing lexical-semantic knowledge more than the other abilities. In the cases where one or more judges gave no response on an item, the figure in the total column for the item is less than 14. However, as the many 14s in the column indicate, most items received a response from each judge.

A total of 483 votes from across 14 judges were given on the 35 items. Of these, 331 were for syntactic knowledge (Mean=9.46, SD=2.39), 83 were for lexical-semantic knowledge (Mean=2.37, SD=2.24), and 69 were for sentence comprehension (Mean=1.97, SD=1.76). Therefore, in general, the items tended to be judged more as syntax items rather than lexical-semantic or sentence comprehension items.

Examining each line for the specific item, it can be confirmed that most items were judged by more individuals to be testing syntax than the other two abilities. There were, however, three exceptions. Namely, Items 12, 18, and 21 (with italicised item numbers), were judged by more individuals to require either lexical-semantic knowledge or sentence comprehension ability.

Table 4.5 Rater judgment on 35 trial syntax items

Item	Syntactic	Lexical-semantic	Sentence comprehension	Total
1	12	2	0	14
2	12	1	1	14
3	11	0	3	14
4	7	5	2	14
5	13	1	0	14
6	7	3	4	14
7	11	2	1	14
8	9	1	4	14
9	9	4	1	14
10	6	4	4	14
11	12	2	0	14
12	5	8	1	14
13	13	1	0	14
14	7	5	2	14
15	11	1	2	14
16	9	3	2	14
17	12	1	1	14
18	5	9	0	14
19	10	0	4	14
20	10	0	4	14
21	4	0	8	12
22	11	0	2	13
23	9	4	1	14
24	9	1	3	13
25	8	3	3	14
26	10	1	3	14
27	12	1	1	14
28	8	0	5	13
29	10	3	1	14
30	10	1	2	13
31	7	6	1	14
32	13	1	0	14
33	9	4	1	14
34	11	2	1	14
35	9	3	1	13
Total	331	83	69	483
Mean	9.46	2.37	1.97	13.80
SD	2.39	2.24	1.76	0.47

The actual items on which more judges gave non-syntax responses were the following:

12. __________ how hard he worked, his tutor never commented on it.
   A. Of no account
   B. No matter
   C. Without regard
   D. Mindless

18. I am taller than you ________ three inches.
    A. with
    B. by
    C. of
    D. in

21. At thirteen ________ at a district school near her home, and when she was fifteen, she saw her first article in print.
    A. the first teaching position that Mary Jane Hawes had
    B. the teaching position was Mary Jane Hawes' first
    C. when Mary Jane Hawes had her first teaching position
    D. Mary Jane Hawes had her first teaching position

Item 12 was judged by eight individuals to be testing lexical-semantic knowledge more than the other two. Item 18 was also labelled a lexical-semantic item by nine judges, while Item 21 was called a sentence comprehension item by eight judges.

Discussion and conclusion

The items to which more judges gave the syntax response should remain as syntax items, but the three items to which they did not must be reviewed.

Item 12 appears to require the knowledge of the behaviour of a particular lexical phrase 'no matter', which is frequently followed by a relative adverb such as 'how' or a relative pronoun. Item 18 seems to require the knowledge of meanings expressed by certain prepositions. And, for Item 21, although it does appear to tap the candidate's awareness of the grammatical structure of the target compound-complex sentence, it appears to involve the candidate in semantic processing at the same time, and accurate semantic processing might be more important for some in finding the correct answer. By a somewhat conservative approach, it was determined that the three items above be excluded from our test of syntactic knowledge. Since the decrease in the number of items was relatively small, no additional items were added, and the remaining 32 items formed the syntax measure for the rest of the present study.

Preliminary Study C: testing vocabulary breadth (version 1)

Introduction/purpose

The main study for the present project required a reasonably compact test of Japanese university learners' breadth of English vocabulary. This preliminary study aimed to obtain empirical data on the 126 items of an existing VLT (cf. Schmitt, Schmitt and Clapham 2001) to determine which combination of two levels as a set generates the highest overall reliability and whether a new set consisting of items from across all of the four levels would exhibit more desirable psychometric characteristics when used with a Japanese university sample group.

Method

The 2,000 word (k=30), 3,000 word (k=30), 5,000 word (k=30), and Academic Word (k=36) levels of Schmitt's VLT were subjected to reliability analysis. As in the analysis of the initial version of the passage reading test in Preliminary Study 1, both classical and Rasch item analyses were employed. Separate analyses were produced for all possible combinations of two levels as a set. The best of these combinations were then compared with an alternative set made up of items chosen from across the entire 126 items.

Participants

A total of 40 L1-Japanese students from a university (n=21) and a junior college (n=19) participated in this preliminary study. They were first to third year English majors and mostly between the ages of 18 to 21. The sample size of 40 is usually considered insufficient for the use of Rasch analysis, thus the results must be treated as tentative and require a follow-up study to verify them with a larger sample.

Procedure

The test was group-administered by the students' instructors in their EFL and related classes. The test rubric in the students' L1 and an example task with answers were printed at the beginning of the test. Thirty minutes were allowed to complete the answers, and no separate time limit for each word frequency level was imposed. The students were told they should not attempt to guess blindly when they have no idea what the correct answer might be.

Results

Table 4.6 summarises the main descriptive statistics for the VLT as a 126-item test and as four separate level tests. The mean raw score of this sample group for the 126 items as a single large test was 58.03 (46% correct) with an SD of 22.48. Among the four levels, the percentage of the correct answers ranged from 76% for the 2,000 word level to 29% for the 5,000 word and Academic Word List levels.

Table 4.6 Descriptive statistics: VLT by level (n=40)

Word level	*k*	Mean	SD
2,000	30	22.85 (76%)	6.43
3,000	30	16.20 (54%)	6.36
5,000	30	8.55 (29%)	4.98
AWL	36	10.43 (29%)	7.53
4 levels combined	126	58.03 (46%)	22.48

When the 126 items from all four levels were treated as a single test, an alpha of .97 was obtained. Point-biserial corrected item-total correlation ranged between −.15 and .78 with a mean of .40. Of the 126 items, 92 items (73%) had a recommended value of above .25 (Henning 1987). The item facility value ranged from 0 to 1 with a mean of .46. Four items were answered correctly by all 126 persons and one item by none. The number of items falling between the recommended facility value range of .33 and .67 (Henning 1987) was 37 (29%). The ratio of the items obtaining the recommended values in facility and discrimination is summarised in Table 4.7.

Table 4.7 Ratio of items achieving recommended values: 126-item VLT (n=40)

Item facility between 33−67%	pbi over .25
37/126 (29%)	92/126 (73%)

Rasch reliability estimates were .94 for the items and .97 for the persons. Infit Mean Square value for the cases ranged from 0.64 to 1.53, and three cases each were lower and higher than the recommended range of 0.75 to 1.3 (McNamara 1996). The items ranged between 0.54 and 1.64 in Infit Mean Square with 14 and 16 items each falling above and below the recommended range respectively (see Table 4.8 for summary).

Table 4.8 Problematic VLT items − all combined (n=40)

Zero or perfect	Infit MS >1.3	Infit MS <0.75	Total
5 items (4%)	14 items (11%)	16 items (13%)	35 items (28%)

The logit mean for item difficulty was 0.00 with an SD of 1.95 and a range of between −5.03 and +3.97. The person ability estimate in logit had a mean of −0.41 with an SD of 1.43 and ranged between −3.7 and +2.12.

As separate tests, the four levels ranged between .83 and .92 on alpha and between .81 and .90 on Rasch case estimate. The overall reliabilities of these separate levels are shown in Table 4.9.

Table 4.9 Reliability data: original VLT levels (n=40)

Word level	*k*	Reliability estimate		
		Alpha	Rasch	
			item	case
2,000	30	.91	.83	.84
3,000	30	.90	.93	.90
5,000	30	.83	.89	.81
AWL	36	.92	.89	.88
All 4 levels combined	126	.97	.94	.97

The next set of analyses focused on the overall reliability coefficients obtained when two of the four levels of Schmitt's VLT were merged into a new test form with around 60 items and attempted to determine the best level pair to use with the Japanese university population. Table 4.10 describes the means and the SDs for all of the combinations.

Table 4.10 Descriptive statistics: VLT two level combinations (n=40)

Word level	*k*	Mean	Mean%	SD
2,000 & 3,000	60	39.05	65%	12.18
2,000 & 5,000	60	31.40	52%	10.69
2,000 & AWL	66	33.28	50%	12.57
3,000 & 5,000	60	24.75	41%	10.85
3,000 & AWL	66	26.63	40%	12.76
5,000 & AWL	66	18.98	29%	11.44

Overall reliability estimates for the six separate combinations are found in Table 4.11. The reliability coefficients were in the narrow ranges of between .93 and .95 on alpha and between .92 and .95 on a Rasch case estimate.

In terms of the ratio of the items which exhibited desirable psychometric characteristics on the basis of classical item analysis, no particular combination could be singled out as outstanding (Table 4.12). The ratio of the items within the recommended item facility range was generally low, from

Table 4.11 Reliability data: VLT 2-level combinations (n=40)

Word level	*k*	Reliability estimate		
		Alpha	Rasch	
			item	case
2,000 & 3,000	60	.95	.92	.94
2,000 & 5,000	60	.93	.94	.95
2,000 & AWL	66	.94	.94	.95
3,000 & 5,000	60	.93	.93	.93
3,000 & AWL	66	.94	.93	.94
5,000 & AWL	66	.93	.88	.92

25 to 33%. On the ratio of items achieving over .25 on item-total correlation, the 2,000−5,000 level combination had the lowest ratio of 68% while the 3,000-AWL level had the highest at 83%.

Table 4.12 Ratio of items achieving recommended values: VLT 2-level combinations (n=40)

Word level	Item facility between 33–67%	pbi over .25
2,000 & 3,000	18/60 (30%)	49/60 (82%)
2,000 & 5,000	15/60 (25%)	41/60 (68%)
2,000 & AWL	18/66 (27%)	53/66 (80%)
3,000 & 5,000	19/60 (32%)	46/60 (77%)
3,000 & AWL	22/66 (33%)	55/66 (83%)
5,000 & AWL	19/66 (29%)	47/66 (71%)

Fit statistics for each of the six paired-level tests were analysed for problematic items in the same way the whole 126-item set was analysed. The number and ratio of the unsatisfactory items for each paired set were as shown in Table 4.13. The 5,000-AWL combination had the fewest items with obvious problems (7 out of 66 items; 11%), while the 2,000−3,000 combination had the most (24 out of 60 items; 40%).

The final section of this preliminary study considered constructing a new vocabulary breadth measure by selecting the items from the 126-item set which demonstrated better item statistics. The analysis started by removing from the 126-item set those items that had zero or perfect scores and then considering the remaining items on the basis of item discrimination and Rasch model fit.

Since Rasch estimates provide more sample-free information than classical analysis, the screening of the items was based on the Infit Mean Square values. Another reason for adopting the Rasch estimates was related to the structure of VLT. Since each VLT task subsumes three items whose answers

Table 4.13 Problematic items – VLT 2-level combinations (n=40)

Word level	Zero or perfect	Infit MS >1.3	Infit MS <0.75	Total
2,000 & 3,000	4 items (7%)	7 items (12%)	13 items (22%)	24 items (40%)
2,000 & 5,000	2 items (3%)	6 items (10%)	10 items (17%)	18 items (30%)
2,000 & AWL	3 items (5%)	7 items (11%)	11 items (17%)	21 items (32%)
3,000 & 5,000	2 items (3%)	2 items (3%)	7 items (12%)	11 items (18%)
3,000 & AWL	3 items (5%)	7 items (11%)	11 items (17%)	21 items (32%)
5,000 & AWL	1 item (2%)	2 items (3%)	4 items (6%)	7 items (11%)

are to be found from a common pool of six synonym answer choices, it could be argued that item independence may not be completely ensured. Infit Mean Square can be consulted since low values might indicate a problem in item independence (McNamara 1996).

The structure of a VLT task imposes a restriction on the manner of item selection. Namely, any deletion or substitution of an item entails that of its adjacent item(s) in the same 3-item task. The process of removing the 35 problematic items and the adjacent items in the same task sets left the new vocabulary measure with 57 items, of which nine were from the 2,000 word level, 15 from the 3,000 word level, 21 from the 5,000 word level, and 12 from the AWL level. To make the item total 60 and to have more balance across the frequency levels, one task with only one problematic item was retrieved from the list of rejected task sets at the 2,000 word level.

If the students had taken a vocabulary test consisting of only these 60 items, they would have had a raw score mean of 23.85 (40% correct) with an SD of 10.86 (Table 4.14).

Table 4.14 Descriptive statistics: new vocabulary breadth test based on item analysis (n=40)

k	Mean	Mean%	SD
60	23.85	40%	10.86

The overall reliability expressed in alpha for this 60-item test was .93, and the item-total correlations were between −.01 and +.65 with a mean of +.41. Item facility mean was .40 and ranged between .03 and .96. As summarised in Table 4.15, 15 items (25%) had facility values above .33 and below .67, and 50 (83%) had an item-total correlation above .25.

Table 4.15 Ratio of items achieving recommended values: new vocabulary breadth test (n=40)

Item facility between 33–67%	pbi over .25
15/60 (25%)	50/60 (83%)

The 60-item set was also subjected to a Rasch analysis, through which both the item and case reliabilities were estimated to be .94. On Infit Mean Square, the range was between 0.53 and 1.56 with a mean of 1 and an SD of 0.28. Six cases were above 1.3 and 10 cases were below 0.75. Six items had values higher than 1.3 and one failed to reach 0.75 by 0.01. The item Infit Square Mean ranged between 0.74 and 1.46 with a mean of 1 and an SD of 0.20. Table 4.16 provides a summary of the numbers and the ratios of problematic items based on Infit Mean Square.

Table 4.16 Problematic items: new vocabulary breadth test based on item analysis (n=40)

Zero or perfect	Misfitting	Overfitting	Total
0 items (0%)	6 items (10%)	1 item (2%)	7 items (12%)

Seven items showed signs of misfit and this accounted for 12% of the 60 items, a ratio comparable to the best of the 2-level combinations analysed above (11%, of 5,000-AWL level).

Discussion and conclusion

This preliminary study has analysed 126 VLT items on their psychometric qualities as they are placed among other items and compared the overall reliabilities of item sets that were produced by merging two different word frequency levels of the test as well as producing a new vocabulary test by selecting the better functioning items based on the item analysis.

Since the overall reliability of a test is directly influenced by the number of items contained in it, it had to be expected that the extremely high reliability of .97 obtained for the whole 126-item test would decrease when the test was reduced to 60 or 66 items in length. The 2-level combination tests did not differ much among each other in the alpha and the Rasch case reliability estimates with the smallest figure being .92 and the largest .95. The new item set produced through item analysis also obtained a comparable overall reliability of .93 in alpha and .94 in the Rasch case estimate.

An item facility range of between .33 and .67 as recommended in the literature (Henning 1987) was quite difficult to achieve, with only 25% to 33% of the items passing that criterion for each of the 2-level combinations. The new test based on item data also suffered in this category with only 25% of its items falling within the ideal range.

In item discrimination, which is expressed in point-biserial corrected item-to-total correlation, the 3,000-AWL pair had the highest percentage of items (83%) showing desirable statistics while the 2,000–3,000 pair obtained a very close figure (82%). The 60-item set based on item analysis also had 83% of its items above the .25 threshold.

When each 2-level combination was subjected to Rasch analysis, the 5,000-AWL pair had the smallest ratio of problematic items (11%), followed by the 3,000–5,000 pair (18%). Here again, the new set of 60 items compiled from the original 126-item test data produced a comparable result, with as few as 12% of its items having weaknesses.

The overall reliability and item facility did not help single out any particular 2-level pairing, thus item discrimination and the model fit had to be the main sources of decision. Depending on the criterion, however, completely different level-pairs would have to be recommended: 3,000-AWL and 2,000–3,000 pairs on item discrimination and 5,000-AWL and 3,000–5,000 pairs on model fit. The set of 60 items prepared by removing items with fit problems from the original 126 item set exhibited equivalent discrimination results and comparable model fit. Therefore, it was tentatively selected for use as a new version of the vocabulary breadth test (Version 2) for the subsequent parts of this research on the ability of the Japanese EFL readers at the tertiary level.

Chapter review

Three preliminary studies have been detailed and their results reported in this chapter. The first study analysed the initial version of the test for measuring the passage reading comprehension ability of our sample population, and the data helped identify a general mismatch between the participant ability and item difficulty. Based on such a finding, replacing the two most difficult of the four passages on the test was proposed. The second study asked EAP professionals and linguists to examine the items in the initial version of the test which aimed to measure the breadth of syntactic knowledge, and the results indicated three of the 35 items to be focusing less on syntactic knowledge than other areas. After the removal of the three items, the updated syntax measure thus consists of 32 items rather than 35. Finally, the vocabulary breadth measure underwent an item analysis which resulted in a compilation of a new set of 60 well-functioning items to be used in the subsequent analyses.

5 Preliminary studies: phase 2

Chapter overview

After the three preliminary studies conducted in the first phase of the research programme, three more are reported in this chapter. The first of the three (Preliminary Study D) included the PB measures of passage reading and syntactic and vocabulary breadth as well as the CB measures of word recognition latency. This is therefore the second opportunity for piloting the three PB measures and the first for the CB measures. The correlations among the measured variables are also analysed in this study. The second of the three studies reported in this chapter (Preliminary Study E) analysed an ESL version of the Reading Span Test (RST), which purports to measure the individual's working memory span. The participants also provided performance data on the passage reading comprehension so that its correlation with the Reading Span measure can be assessed. The last of the preliminary studies (Preliminary Study F) attempted to identify a group of sentence comprehension items that may be utilised in the CB sentence reading speed measure. The following sections will report on the three studies in sequence.

Preliminary Study D: passage reading comprehension, syntactic knowledge, vocabulary breadth, and word recognition efficiency

Introduction/purpose

The first three preliminary studies reported above represented some initial attempts at evaluating and revising the group-administered portion of the research instruments for use with a sample of Japanese university students. This fourth preliminary study had two main objectives. First, it sought to obtain reliability information on our initial set of word recognition efficiency measures as well as on our second versions of passage reading comprehension, syntactic knowledge, and vocabulary breadth. The word recognition item data will be consulted with a view to reducing the test item size for the subsequent main study. Second, on the assumption that the instruments would be found reasonably reliable, the study also aimed to explore the degree to which syntactic knowledge, vocabulary breadth and word recognition efficiency account for the individual differences in passage reading performance. Since the research instruments themselves are also the primary foci of this preliminary study, their details are discussed separately, with the PB tests and the CB tests having their own sections on methods and results. However, since the data were obtained from a common sample of research participants, the information on the sample and the general research procedure will precede such discussions. After the description of the general procedures, the reliability results of the PB tests are first reported, followed by the details of the CB latency measures. After these instrument analyses, the method and the results of the analysis of the relationships among reading, syntax, vocabulary, and word recognition efficiency will be detailed.

Participants

The sample for this preliminary study consisted of L1-Japanese EFL students from three universities in northern Kyushu. Their academic areas of concentration were American and European studies, business, economics, English, law, mechanical engineering, and sports science. Most belonged to the first three years of university study and the age range was between 18 and 21 years, though the sample also included some fourth-year and postgraduate students. A total of 189 students participated in the study, but 75 of them attended both the PB and the CB test sessions. A total of 182 students took the three PB measures of passage reading comprehension, syntactic knowledge, and vocabulary breadth, and 82 of them also provided word recognition latency data on CB testing. The reliability studies are reported on the basis of the data from these sets of students.

General procedure

The PB tests were administered as part of the students' coursework in their EFL or related classes. The three tests were completed in one session during a 90-minute class. The passage reading comprehension test was administered first, followed by the syntax test and the vocabulary test. The time allowed to complete each of the tests was 35 minutes, 20 minutes, and 15 minutes, respectively. The test rubric was printed in the students' L1 at the beginning of each test, and the vocabulary test also had one sample test task. The students were told they should leave an item unanswered if they had no idea what the correct answer might be, rather than guessing randomly. However, educated guesses were not discouraged. Even when they were not certain of the correct answer, they were encouraged to guess if they had some idea of it. The details of each test are discussed in their own sections below.

The students who took the PB tests were invited to participate in the CB individual sessions, which were scheduled outside of their class time and for which payment was made. An effort was made to schedule the individual session for each student so that it took place as close to the PB testing as possible. Every student took both the PB and the CB tests within two weeks of each other, many of them on the same day or within a few days. The CB tests took 10 to 20 minutes per student.

Testing passage reading comprehension, syntactic knowledge, and vocabulary breadth (version 2)

Method

In the same manner in which the initial versions of the passage reading and vocabulary tests were analysed in Preliminary Study 1 and 3, each of the three measures of passage reading ability, syntactic knowledge, and vocabulary breadth was subjected to a series of analyses including its overall internal consistency and a Rasch model fit as well as its individual item statistics. Person ability and item difficulty estimates were also compared to evaluate how well they matched.

Results

As summarised in Table 5.1, the raw score means for the three tests of passage reading, syntactic knowledge, and vocabulary breadth were 10.07 (SD=4.38), 15.41 (SD=5.69), and 24.14 (SD=11.50), respectively. Their reliabilities, as estimated using Cronbach's alpha were .80 for passage reading, .82 for syntax, and .94 for vocabulary breadth. Their raw means translate to 50%, 48%, and 40% of their respective maximum possible scores.

Table 5.1 Descriptive statistics: paper-based tests – version 2 (n=182)

Word level	*k*	Mean	SD
Passage reading	20	10.07 (50%)	4.38
Syntactic knowledge	32	15.41 (48%)	5.69
Vocabulary breadth	60	24.14 (40%)	11.50

Item statistics were also computed for the three measures. Regarding the passage reading items, the item-to-total correlations ranged between .13 and .60 with 19 (95%) of the items achieving higher than .25. Their item success rate ranged between .21 and .81 with 15 items (75%) falling in the range between .33 and .67. The item-total correlation for the syntax items ranged between .06 and .58 with 22 items (69%) falling above .25. The item success rates for the syntax items were between .11 and .81 with 18 items belonging to the recommended range from .33 to .67. Vocabulary items had a range in item-total correlation from .06 to .62 and 57 of them (95%) had a value above

.25. In item success rate, the vocabulary items ranged from .06 to .92 and 19 (32%) were in the range between .33 and .67.

Rasch overall item reliability estimates were .96 for passage reading, .97 for syntax, and .98 for vocabulary, while the case reliability estimates were .79 for passage reading, .82 for syntax, and .93 for vocabulary (Table 5.2).

Table 5.2 Reliability data on the paper-based tests – version 2 (n=182)

	k	**Reliability estimate**		
		Alpha	**Rasch**	
			item	**case**
Passage reading	20	.80	.96	.79
Syntactic knowledge	32	.82	.97	.82
Vocabulary breadth	60	.94	.98	.93

Infit Mean Square for the passage reading ranged from 0.76 to 1.29 for the items and from 0.62 to 1.44 for the cases. All items were within the recommended range but 20 persons fell outside of it. The logit mean for item difficulty was 0.00 (SD=0.87) and a range of between −1.70 and +1.66. Person ability logit mean was 0.04 (SD=1.24) and ranged from −3.28 to +3.23.

Infit Mean Square for the syntax test had a range between 0.78 and 1.26 for the items and between 0.63 and 1.43 for the persons with 18 persons placed outside of the recommended range. The item difficulty, in logits, ranged between −1.79 and +2.29 with a mean of 0.00 (SD=1.08) and person ability ranged between −2.67 and +3.21 with a mean of −0.1 (SD=0.99).

For the vocabulary test, Infit Mean Square ranged from 0.77 to 1.36 for the items and from 0.58 to 1.56 for the persons. Two items and 27 persons had Infit Mean Square values below or above the recommended range. In logits, the item difficulty estimates ranged from −3.82 to +2.95 around a mean of 0.00 (SD=1.68) while the person ability estimates were between −3.78 and +5.05 with a mean of −0.61 (SD=1.43).

Discussion and conclusion

The three PB measures analysed above will now be discussed in turn. The second version of the PB instruments for measuring the Japanese university students' EFL passage reading comprehension achieved a much higher level of internal consistency (alpha=.80, Rasch case reliability estimate=.79) compared to the initial version reported in Chapter 4 (alpha=.47, Rasch case reliability estimate=.53). This also coincided with a higher group mean of 10.07 (50% correct) and closer match between the item difficulty and person ability, which can be visually verified by comparing the logit score distributions on the map (Figure 5.1). Individual item statistics also improved for the reading test and 95% of the items achieved the recommended level of item-total correlation (compared to only 15% in the initial version). The substitution of the two texts, thus, seems to have had the desired effect and now, though far from being perfect, our passage reading comprehension measure can be expected to distinguish among our sample of Japanese EFL readers at a satisfactory level.

Figure 5.1 Plot of person ability and item difficulty estimates from reading comprehension test, version 2

4.0		
	XXXX	
3.0		
	XX	
2.0	XXXXXXXXXXX	
		16
	XXXXXXXX	
	XXXXXXXXXX	
1.0	XXXXXXXX	14
	XXXXXXXXXXXXXXXXXX	17
		13 15 20
	XXXXXXX	18
		3
	XXXXXXXXXXXX	5 11 12 19
0.0	XXXXXXXXXXXXXX	
	XXXXXXXXXXXXXXX	
		4 7
	XXXXXXXXXXXXXXXX	8 9
	XXXXXXXXXXXXXXXX	1
−1.0	XXXXXXXXXX	
	XXXXXXXXXXXX	6
	XXXXXXXXXXX	10
		2
−2.0	X	
	XXXXX	
−3.0		
	XX	
−4.0		

Each X represents 1 student

The 32-item syntax knowledge measure also achieved a fairly high level of internal consistency (.82 on both alpha and Rasch case reliability estimate) and the Rasch model fit for the items was within the recommended range, although the ratio of items achieving recommended levels of item-total correlation was a relatively low 69%. Still, the item difficulty and person ability matched well, suggesting suitability for use with our Japanese EFL sample.

Finally, the vocabulary breadth measure maintained exactly the same level of internal consistency compared with the data from the previous sample. The alpha was .94 and the Rasch case reliability estimate was .93 with the latest sample while they were .93 and .94 respectively when the same items were embedded in the full length vocabulary test in Preliminary Study C. Most of the items obtained the recommended level of discriminability (95%) and Rasch model fit (97%), an improvement from the somewhat lower ratios obtained earlier (83% and 88%, respectively) when the same 60 items in the original 126-item test were treated as one autonomous test (see Chapter 4). Interestingly, the mean raw scores were very similar between the two samples (approximately 40% correct in both Preliminary Study C and in this Preliminary Study D). The larger sample size and the wider level of participants available for the current study may have contributed to larger score variance (SD=11.50, compared to 10.86 for Preliminary Study C), which is a factor in improving reliability coefficients. Person ability estimate means in logits were also similar: −0.8 with an SD of 1.55 in Preliminary Study C and −0.61 with an SD of 1.43 in the present preliminary study.

The three PB tests demonstrated a reasonably high degree of reliability when analysed with both classical and Rasch methods, thus it was judged appropriate to adopt these tests as instruments for the measurement of our target group's passage reading comprehension, syntactic knowledge, and vocabulary breadth.

Testing word recognition efficiency (version 1)

Method

As detailed in the 'Methodology' section (Chapter 3), the word recognition efficiency measure employed speeded 'same-or-different' decision tasks. The objective was to record the participants' latency in the visual recognition of five types of stimulus: real word (k=12), pseudoword (k=12), irregular letter string (k=12), synonym/antonym (k=36), and numeral (k=24). Since the first three subsets uniformly required no more than visual recognition of alphabetic character strings, unlike the number matching, which required no alphabetic processing, and unlike the synonym/antonym decision task, which required both visual processing and lexical access, all item pairs from these three visual alphabetic processing types were mixed and randomly sequenced within a single block (k=36). Therefore, the five stimulus types were presented to each student as three blocks rather than five blocks, and they appeared in the test in the order of (1) number pair block, (2) alphabetic string pair block, and (3) synonym/antonym pair block. A practice session consisting of six item pairs preceded each of the actual test blocks. The test-taking procedure was explained in the student's L1, first by a test proctor individually and also on the computer screen as each student proceeded through the instruction screens and responded to practice items. They were first instructed to place their left index finger on the 'F' key, their right index finger on the 'J' key and their thumb(s) on the space bar. The display equipment was a 14-inch active matrix LCD screen with its resolution set at 1024 by 768 pixels. At the student's self-paced key press, each stimulus pair appeared in the centre of the screen in size 36 Times font and remained there until one of two keys was subsequently pressed to indicate either 'same' or 'different'. The latency for each item was the chronometric gap between these two consecutive key presses, which was timed and recorded by the computer at 17-millisecond accuracy. A prompt would appear instructing the student to press the space bar to proceed to the next screen or to initiate a trial, and the 'J' key was to be pressed when the two items in a pair were the same and the 'F' key when different. The actual items were randomly selected by the computer program from within each of the three blocks of stimulus pairs with replacement. Therefore, depending on the student, some items appeared more than once, while others did not at all. The split-half method was employed in estimating the internal consistency of the latency measures since the missing values resulting from random item sampling with replacement made the use of Cronbach's alpha unsuitable. However, since the split-half correlation coefficients can vary depending on how the test is split, it was deemed appropriate to obtain several correlation coefficients based on repeated random splits, and five separate correlations were produced. Before calculating the split-half correlations, the students' same-or-different responses were checked for their accuracy. In particular, the response accuracy of the synonym/antonym items was compared to identify those items which were responded to incorrectly by many. Since this measure targets efficiency of accessing words in memory and not the individual difference in vocabulary breadth, items which tend to test whether a word is known or not, thus likely to obscure the latency results, were to be removed from subsequent studies.

Results

The analysis began by considering the accuracy of the students' responses to all of the stimulus pairs. An initial set of response accuracy data revealed that a small proportion of the students responded to certain types of stimulus with an unusually low level of accuracy, and a total of six students whose mean accuracy rate was either lower than the chance level of 50% or four SDs below the sample mean on any one stimulus type were removed from further analysis. (Three SDs might have been a valid cut-off criterion for normally distributed accuracy scores, but given the narrow SDs for the present sample, a slightly more conservative cut-off point of four SDs was adopted.)

Tables 5.3 and 5.4 list the basic descriptive statistics on response accuracy from the remaining 76 students. Table 5.3 pertains to the person means and Table 5.4 to the item means. The mean accuracies in percentage were in the mid 80s to mid 90s and the SDs were between 3 to 12 for each stimulus type. As expected from their exposure to the numerals through education and daily use, the numeral pairs recorded the highest mean accuracy, while the irregular string matching and the synonym/antonym decision had the lowest.

Table 5.3 Descriptive statistics: response accuracy mean for person by stimulus type (n=76)

	Person			
	Min	**Max**	**Mean**	**SD**
Number	0.79	1.00	0.94	0.05
Real word	0.67	1.00	0.91	0.09
Pseudoword	0.75	1.00	0.91	0.08
Irregular string	0.50	1.00	0.85	0.12
Synonym/antonym	0.58	0.97	0.85	0.09

Table 5.4 Descriptive statistics: response accuracy mean for item by stimulus type (n=76)

	Item			
	Min	**Max**	**Mean**	**SD**
Number	0.88	1.00	0.95	0.03
Real word	0.73	1.00	0.93	0.08
Pseudoword	0.86	0.98	0.93	0.04
Irregular string	0.72	0.96	0.87	0.09
Synonym/antonym	0.61	1.00	0.88	0.10

The student's recognition time for each stimulus type was obtained by averaging their latencies on all of the item pairs of that type including the misjudged ones. This decision to include the latencies of misjudged items was based on the almost perfect correlation (.99) between the mean latencies of all items and those of only the items which were judged correctly.

Table 5.5 summarises the descriptive statistics of the same-or-different decision latencies. Of the four speeded decision tasks involving alphabetic processing, the real word pairs had the shortest mean decision time of 944.18 milliseconds (msecs) while the synonym/antonym decision required the longest 1,621.02 msecs. Between the two were the pseudowords (1,021.53

msecs) and the irregular letter strings (1,170.22 msecs). The latency difference between real word pairs and pseudoword pairs was calculated as the lexicality effect and its mean for the sample group was 77.35 msecs. The mean of the orthography effect, which is the difference between pseudoword and irregular string latencies, was 148.69 msecs for this sample.

Table 5.5 Descriptive statistics: same–different decision latency mean – version 1 (n=76)

	k	Min	Max	Mean	SD
Number	24	538.25	1288.00	800.11	159.49
Real word	12	679.67	1627.00	944.18	184.00
Pseudoword	12	714.00	1639.71	1021.53	195.40
Irregular string	12	721.63	2033.88	1170.22	276.70
Synonym/antonym	36	998.04	2631.88	1621.02	387.99
Lexicality effect		−294.40	362.13	77.35	108.47
Orthography effect		−206.09	707.75	148.69	153.62

Unit=millisecond

Results of a repeated-measures ANOVA (Table 5.6) indicated that there is a significant difference in the latencies between the stimulus types ($F_{(4,\ 300)}$=218.485, p<.001), and post-hoc pairwise comparisons with the Bonferroni adjustment procedure also indicated that the mean difference was statistically significant between any two stimulus types (p<.001).

Table 5.6 ANOVA table: within subject effect of stimulus type

Source	SS	df	MS	F	Sig.
Stimulus type	30104500.34	4	7526125.085	218.485	.000
Error	10334058.91	300	34446.863		

Therefore, the pattern of differences in recognition latencies among the five stimulus types were as follows:

Number < Real Word < Pseudoword < Irregular String < Synonym/Antonym

Since a major purpose of this preliminary study has been to examine the adequacy of the current word recognition latency measure for a Japanese university sample, the reliability data are reported with some detail. The split-half correlation coefficients based on the 76 students' latencies are summarised in Table 5.7. For each stimulus type, the items were randomly sequenced, with the resulting initial half and latter half yielding their own average latencies for that particular iteration, and these random split averages were correlated. This process was repeated five times for each stimulus type to avoid one particular instance of split influencing our decision. As found in the far right column, the median split-half correlation for each stimulus type ranged from .72 to .88, with the number pairs showing the most desirable statistics and the irregular letter strings the least.

Table 5.7 Split-half correlations of latency measures – version 1 (n=76)

		Sampling iteration					
	k	**1st**	**2nd**	**3rd**	**4th**	**5th**	**Median**
Number	24	.84	.88	.89	.88	.87	.88
Real word	12	.79	.80	.76	.79	.76	.79
Pseudoword	12	.77	.72	.70	.74	.76	.74
Irregular string	12	.77	.81	.72	.70	.72	.72
Synonym/antonym	36	.83	.83	.83	.82	.84	.83

Since each speeded response item is recorded in continuous values (i.e., in msecs) unlike the dichotomously scored PB test items (i.e., correct=1, incorrect=0), recognition efficiency tasks should require far fewer items to observe individual differences. The practical requirement of reducing the size of the test battery for the subsequent main study motivated an analysis in which each of the stimulus type means was correlated with a corresponding mean latency that was newly calculated from a smaller number of randomly chosen item subsets. For number matching and synonym/antonym types, subsets of half of the original item sizes were tested (k=12 and 18, respectively), while the three visual alphabetic processing types adopted subsets of two thirds of the original item size (k=8 each). Again, though randomly selected, use of a single instance of item selection may fail to capture reality. Therefore, random compilation of the item subsets was repeated five times, and the results are summarised in Table 5.8.

Table 5.8 Correlations with reduced item sets: latency measures – version 1 (n=76)

	Full *k*	Reduced *k*	Sampling iteration					Median
			1st	**2nd**	**3rd**	**4th**	**5th**	
Number	24	12	.97	.96	.96	.97	.97	.97
Real word	12	8	.98	.99	.99	.98	.99	.99
Pseudoword	12	8	.98	.97	.96	.96	.97	.97
Irregular string	12	8	.97	.97	.97	.97	.95	.97
Synonym/antonym	36	18	.95	.96	.97	.97	.97	.97

The median correlation across the five random sampling iterations was .99 for real words and .97 for the other four stimulus types. In each stimulus type, there was little variation across the sampling iterations, with .95 to .97 for irregular strings and synonym/antonym decisions and .96 to .98 for pseudowords recording the widest variation of the five.

Discussion and conclusion

One of the objectives of this preliminary study was to examine the synonym/antonym items with the view of identifying some of the items causing our sample group problems in responding accurately. To focus on such items, Table 5.9 lists 10 items which had the lowest accuracy rates among the 36 employed in the study.

Table 5.9 Synonym/antonym items with lowest accuracy means

near close	61%
calm quiet	65%
send receive	66%
right correct	68%
full empty	79%
buy sell	82%
never always	82%
forget remember	85%
small little	86%
past future	86%

Since even the purely visual processing stimulus types had items with accuracy rates as low as 72% (irregular strings), it can be assumed that roughly 30 to 35% of inaccurate responses can occur due to mere recognition errors and not necessarily to lack of lexical-semantic knowledge in long-term memory. If synonym/antonym items above this accuracy level are to be employed, the top four pairs in Table 5.9 emerge as problematic. If necessary, these items should be removed before the others.

The pattern of data concerning the differences across the stimulus types within individuals was not surprising and generally replicated that of Haynes (1989), in which the response time of the Chinese EFL readers at a Taiwanese university was found to be negatively affected by the lack of familiar word shapes and of orthographic regularities. In her study, native speakers of English and the Taiwanese EFL learners did not differ significantly in the speed of recognising real words. However, the native speakers were significantly faster at recognising pseudowords, and they processed them as fast as they did real words. Haynes's subjects are described as among the most competent EFL learners at Taiwanese universities and still lacked the ability to benefit from systematic sequencing of alphabetic characters. The English proficiency of the Japanese sample group for the present experiment is expected to vary somewhat more than theirs as it included less competent subgroups of learners as well. It is therefore not surprising that their overall group performance indicated that they had not developed the skill to process regular orthographic sequences at similar speeds whether the sequences formed real words or not. Since these students were also affected by the lack of orthographic regularities (the orthography effect), it is evident that they relied on such regularities rather than employing an entirely different whole word approach. Nevertheless, with non-alphabetic L1 background like the L1-Chinese readers, Japanese EFL readers in the present study also seemed to lack the level of sensitivity to the regular orthographic patterns of English words that is considered to be normally available among native-speaker readers.

The analysis also paid closest attention to the reliabilities of the latency measures. The split half correlation coefficients were in the low seventies to high eighties, indicating a fairly reasonable level of internal consistency for these measures. The number matching was the most consistent, and it constantly surpassed even the synonym/antonym decision, which had a 50% larger k size. The three stimulus types of real word, pseudoword, and irregular letter strings showed somewhat smaller values, which may be associated with smaller k sizes. The lack of uniformity in the actual items and the number of items used across the individuals for each of these stimulus types may have interacted negatively with their smaller k sizes. This initial investigation of the recognition efficiency test prioritised randomness such that not only the sequence of item presentation but the actual item selection was randomised. While there is no reason to abandon random sequencing of the items within the three blocks, it appears more appropriate, for reliability purposes, to fix the set of items across the individuals rather than to have the computer randomly create the set for each individual.

The continuous nature of the values obtained for each stimulus item, as opposed to dichotomously coded items for PB tests, may make the CB latency measurement much less prone to influence from reduced item size. To investigate the effect of reduced item size, the original latency mean for each of the stimulus types was correlated with mean latency of a subset randomly created from either half or two thirds of the original item set. The median correlations from the five separate random samplings of the subset items were .97 or higher, suggesting that the results would have been very similar even if the item size had been half or two thirds of this original.

Taking into account all of the results regarding accuracy rates, split-half correlations, and correlations with reduced subsets, it was determined that for the main study to follow in the current project, the word recognition efficiency test would contain at least 12 number matching, eight real word matching, eight pseudoword matching, eight irregular string matching, and 18 synonym/antonym decision items, all of which would be presented to each student though the sequence of appearance would be randomised, and that the four synonym/antonym items with the lowest response accuracy rates would not be included in the 18.

Syntactic knowledge, vocabulary breadth, and word recognition efficiency as predictors of passage reading comprehension

Method

To explore the extent to which syntactic knowledge, vocabulary breadth, and word recognition efficiency account for the individual differences in passage reading test performance, the test results from 69 students who attended both the PB group testing of passage reading, syntax, and vocabulary and CB individual testing of word recognition efficiency were subjected to a multiple regression analysis.

Results

Prior to the multiple regression analysis, the basic descriptive statistics and correlation results from the 69 cases were obtained.

The descriptive statistics are summarised in Table 5.10. Results from the three PB measures are represented in logit values as well as in bracketed raw values. The CB latency values are all in milliseconds.

Table 5.10 Descriptive statistics (n=69)

	Min	Max	Mean	SD
Reading-logit	−3.28	3.23	0.65	1.37
(Reading-raw)	(1)	(19)	(12.28)	(4.66)
Syntax-logit	−1.36	3.21	0.33	1.08
(Syntax-raw)	(8)	(30)	(17.81)	(6.10)
Vocab breadth-logit	−2.77	5.05	0.01	1.65
(Vocab breadth-raw)	(8)	(59)	(29.28)	(13.35)
Number latency	538.3	1206.1	789.96	140.78
Word latency	679.7	1627.0	933.88	176.14
Pseudoword latency	714.0	1639.7	1014.14	189.78
Irregular string latency	721.6	2033.9	1164.28	274.68
Synonym/antonym latency	998.0	2631.9	1604.29	386.34
Lexicality effect	−78.1	362.1	80.25	102.28
Orthography effect	−206.1	707.8	150.15	158.67

Table 5.11 lists the bivariate correlation coefficients among all of the variables and Table 5.12 the partial-correlation coefficients with the individual differences in the number matching latency held constant. The three variables

Table 5.11 Simple bivariate correlations among the measured variables (n=69)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
1. Reading	--	.70***	.67***	.03	−.02	.09	.19	−.40**	.20	.22
2. Syntax		--	.72***	.09	−.12	.01	.11	−.53***	.22	.18
3. Vocabulary breadth			--	.18	.00	.15	.15	−.42***	.28*	.08
4. Number latency				--	.56***	.59***	.55***	.16	.14	.24
5. Word latency					--	.85***	.83***	.45***	−.15	.42***
6. Pseudoword latency						--	.83***	.44***	.40**	.24
7. Irregular string latency							--	.32**	.11	.74***
8. Synonym/antonym latency								--	.03	.03
9. Lexicality effect									--	−.28*
10. Orthography effect										--

Note: *** $p<.001$, ** $p<.01$, * $p<.05$

Table 5.12 Partial correlations controlling for number matching latency (n=69)

	1.	2.	3.	5.	6.	7.	8.	9.	10.
1. Reading	--	.70***	.67***	−.05	.09	.20	−.41***	.20	.22
2. Syntax		--	.71***	−.21	−.06	.08	−.55***	.21	.17
3. Vocabulary breadth			--	−.13	.05	.06	−.46***	.26 *	.04
5. Word latency				--	.77***	.75***	.44***	−.28 *	.36**
6. Pseudoword latency					--	.75***	.43***	.40 **	.12
7. Irregular string latency						--	.28*	.05	.75***
8. Synonym/antonym latency							--	.01	.00
9. Lexicality effect								--	−.33**
10. Orthography effect									--

Note: *** $p<.001$, ** $p<.01$, * $p<.05$

measured via PB testing are based on their logit scores obtained from Rasch analysis. Correlation between the logit and raw scores for any of these three variables was .99 or above.

The strongest correlations were found among the alphabetic processing latencies of words, pseudowords, and irregular letter strings ($r=.83$ to $.85$, $p<.001$) and among the three PB tests of passage reading, syntax, and vocabulary breadth ($r=.67$ to $.72$, $p<.001$). Additionally, the orthography effect correlated significantly with irregular letter string latency ($r=.74$, $p<.001$). The number matching latency correlated significantly only with visual processing latencies of words ($r=.56$, $p<.001$), pseudowords ($r=.59$, $p<.001$), and irregular letter strings ($r=.55$, $p<.001$) and marginally with the orthography effect ($r=.24$, $p=.05$). The synonym/antonym decision latency correlated moderately with both the PB test variables ($r=-.40$ to $-.53$, $p<.001$) and alphabetic processing latencies ($r=.32$ to $.45$, $p<.01$) but not with number matching ($r=.16$, n.s.).

When the individual differences in number matching latency were statistically held constant, correlations among the alphabetic processing latencies were somewhat reduced although they still recorded the strongest partial correlations ($r=.75$ to $.77$, $p<.001$). As was expected from the lack of relationship between the PB tests and the number matching latency, the partial correlations among the results of PB tests were little affected when the number matching speed variance was controlled for ($r=.67$ to $.71$, $p<.001$). The orthography effect maintained its strong relationship with irregular string matching latency ($r=.75$, $p<.001$), and the synonym/antonym decision latency marginally gained its strength of relationship with the PB test performances ($r=-.41$ to $-.55$, $p<.001$) while slightly losing that with alphabetic processing latencies ($r=.28$ to $.44$, $p<.05$).

Multiple regression analysis was performed with the passage reading logit score as the criterion variable and the other PB-test logit scores and the recognition latencies as the predictor variables. Since the number matching latency was designed as a means to measure and control for the individual latency differences that are irrelevant to alphabetic or linguistic processing, it was always entered into the regression equation in the first block and kept in the equation irrespective of its level of statistical significance. The rest of the predictor variables were analysed in a stepwise regression procedure in which predictor variables leading to changes in F-values with a probability of .05 or smaller were to be retained in the equation. Using this method, statistically significant models emerged, and the largest amount of passage reading variance was accounted for when the two significant predictors of syntactic knowledge and vocabulary breadth were retained in addition to the non-significant number matching latency ($F_{(3,65)}=26.68$, $p<.001$. Adjusted $R^2=.53$). The regression coefficients from this model are shown in Table 5.13.

Table 5.13 Summary of multiple regression analysis for variables predicting passage reading comprehension (with number latency)

Variable	B	SE (B)	β	p
Block 1				
Number latency	0.000	0.001	−.071	.406
Block 2				
Syntactic knowledge	0.564	0.150	.447	.000
Vocabulary breadth	0.300	0.100	.362	.004

Since no latency variable is a significant predictor in our best regression model, the number matching latency becomes unnecessary. Removing it from the regression equation and retaining only the two significant predictors of syntactic knowledge and vocabulary breadth slightly improved the predictive power ($F_{(2,66)}=39.85$, $p<.001$. Adjusted $R^2=.53$). The regression coefficients from this revised model are shown in Table 5.14.

Table 5.14 Summary of multiple regression analysis for variables predicting passage reading comprehension (without number latency)

Variable	B	SE (B)	β	p
Syntactic knowledge	0.572	0.150	.453	.000
Vocabulary breadth	0.286	0.098	.344	.005

Of the two significant predictor variables, syntactic knowledge obtained higher standardised beta coefficients than vocabulary breadth in either of the regression models above.

Discussion and conclusion

The regression results obtained above suggest that the knowledge variables can account for the individual differences in passage reading comprehension better than the latency variables do. Both syntactic knowledge and vocabulary breadth contributed uniquely to the prediction of the passage reading performance while none of the word recognition latencies did. Between the two knowledge variables of syntax and vocabulary breadth, the higher beta coefficient obtained by the syntax measure indicates that it is the best predictor of passage reading performance of these Japanese EFL learners.

The partial correlation data (Table 5.12) also coincide with the non-significant regression results found for the word recognition latencies. The only exception is the synonym/antonym decision latency, which had a significant negative partial correlation of −.41 with passage reading ($p<.001$). Since greater performance is represented in higher paper-test scores and lower latency values and vice versa, the negativity of the correlation here should not be surprising. It is also notable that the synonym/antonym decision latency exhibited a somewhat different pattern of correlations from the other word recognition latencies. While the three purely visual processing tasks of word, pseudoword, and irregular letter string matching latencies correlated highly among each other (.75 and above), the synonym/antonym decision latency recorded clearly lower, though still statistically significant, correlations with these latencies ($r=.28$ to .44). On the other hand, it is the only type of word recognition latency that correlated significantly with the three PB test variables ($r=-.41$ to $-.55$, $p<.001$). The strongest of these correlations was with syntactic knowledge (−.55). Synonym/antonym decision is unique among our set of speeded word recognition tasks in that it requires lexical-semantic access. It tests the speeds of both visual word recognition and lexical-semantic access, whereas the other three simple visual discrimination tasks can be completed with no semantic processing. It seems most reasonable to attribute the differences in the pattern of correlations to this difference in the nature of processing requirement.

Interestingly, synonym/antonym latency correlated more strongly with syntax than with vocabulary breadth (−.55 vs −.46). Skills in lexical access and breadth of vocabulary have both been theorised as dimensions of lexical competence (Meara 1996, Nation 2001, Richards 1976), but the correlation results obtained above seem to indicate that either of these two 'dimensions of lexical competence' is more closely related to the ability to identify syntactically acceptable structures than to the other dimension of the same competence. This suggests that it would be of theoretical value to look into the latent trait structure of the L2 learners' lexical, semantic, syntactic, and passage comprehension test results. While the sample size of the present preliminary study does not permit factor analysis, the data discussed thus far lend support for further examination of the relationships among passage reading comprehension, syntax, vocabulary, and synonym/antonym recognition speed in our subsequent main study with a view to exploring the possibility of such latent variables and their relationships with these abilities.

The present preliminary study has focused on the syntactic, lexical-semantic, and orthographic abilities as possible sources of individual differences in passage reading comprehension. Syntax and vocabulary breadth accounted for passage reading much better than word recognition efficiency did, and they are clearly worthy of inclusion in the main study to follow. Word recognition variables showed a weaker but significant link to passage reading comprehension when the task required lexical-semantic access. Synonym/antonym recognition speed should deserve further attention as it was also linked to other linguistic dimensions of L2 proficiency such as syntax and vocabulary breadth. While none of the simple visual word recognition latencies has shown any direct link to careful passage reading comprehension in the present study, it seems premature at this point to abandon these variables entirely after a single instance of data collection. Additionally, since inefficient and effortful decoding of orthographic information is expected to hinder fluent reading, there remains a possibility that the orthographic processing efficiency is importantly related to reading speed, which is another aspect of reading investigated in our subsequent main study. Researching whether or not visual word recognition efficiency accounts for reading speed should contribute to our general knowledge of the L2 reading. From the viewpoint of practicality, the visual processing items are among the least time-consuming. They can be completed in as little as 5 minutes. On these accounts, despite their failure to predict passage reading comprehension for the preliminary sample population, the three visual matching speed variables of real words, pseudowords, and irregular letter strings are kept in our set of latency variables for the subsequent main study.

Preliminary Study E: passage reading comprehension and working memory

Introduction/purpose

This fifth preliminary study explored the strength of relationship between performances in the passage reading comprehension test and Osaka and Osaka's ESL version (1992) of the Reading Span Test (RST) in order to assess whether or not the working memory variable as measured via this test is worth further investigation in the subsequent multivariate study. The study also attempts to investigate whether the higher sentence levels of the RST are truly more difficult than the lower sentence levels. It also aims to gain information which would help reduce the size of the RST without compromising its ability to distinguish persons on the basis of their competence in simultaneously performing short-term verbal memory storage and verbal information processing.

Participants, procedures and method of analysis

Participants were 38 students at two universities (n=32) and at an institute of higher learning (n=6) in northern Kyushu. They took the second version of the passage reading comprehension test described in earlier sections and Osaka and Osaka's ESL version of the RST (1992). The passage reading test was group-administered during their class time as part of their EFL class work and 35 minutes were allowed for this test. The RST was administered individually outside their class hours to those who responded to an invitation to the study. The students received payment for their participation in the RST. Each individual RST session took approximately 30 minutes. The researcher himself administered the RST as a proctor whenever possible. However, when scheduling difficulty necessitated concurrent testing, a trained individual also administered the test as a proctor. The administration procedure for the RST closely followed that of its original author, which is reproduced below:

> . . . Each sentence was printed on a single line across the center of a 13 x 18 cm White card. The cards were arranged in five sets, each of which comprised two, three, four, and five sentences. Blank cards were inserted between the sets. Within a set, the sentences were not related to each other. The subject was asked to read each sentence aloud at his/her own pace. As soon as the subject finished reading a sentence orally, the next sentence was presented and the subject was forced to continue reading aloud. After reading all the sentences in a set, the subject was asked to recall the last word of each sentence within the set. The order of reporting these final words was based on the free recall procedure. The subject was prohibited from reporting the last target word first within each set, in order to avoid the recency effect (1992:288).

The scoring procedure also adhered closely to the original. The student's span level was determined by the level of sentence size at which they succeeded in the recall of all of the sentence final words in the set in three or more of the five available trials. After three or more successful trials at a particular sentence level, if in the subsequent level the student was successful in only two out of the five trials, an additional 0.5 point was awarded. When the student failed in four or five trials at the 2-sentence level, the score of zero was assigned, whereas in the case of a student succeeding in two out of five trials at the 2-sentence level, a score of 1.5 was awarded. Therefore, the possible span score range was between 1.5 to 5.0 with the exception of zero score. Table 5.15 shows a hypothetical distribution of successful recall trials by nine dummy students and the span scores that would be awarded to them for their recall performances. Scores from this RST and the students' passage reading test scores were subjected to Pearson product-moment correlation.

Table 5.15 Hypothetical distribution of successful recall trials and their associated span scores

	Number of successful trials (out of 5)				Span score to be awarded
	2-sentence set	3-sentence set	4-sentence set	5-sentence set	
Student A	5 trials	5 trials	4 trials	3 trials	5.0
Student B	5 trials	5 trials	3 trials	2 trials	4.5
Student C	5 trials	4 trials	3 trials	0 trial	4.0
Student D	5 trials	3 trials	2 trials	0 trial	3.5
Student E	4 trials	3 trials	1 trial	0 trial	3.0
Student F	3 trials	2 trials	0 trial	0 trial	2.5
Student G	3 trials	1 trial	0 trial	0 trial	2.0
Student H	2 trials	1 trial	0 trial	0 trial	1.5
Student I	1 trial	0 trial	0 trial	0 trial	0.0

Unlike the other measures analysed in the earlier sections of this chapter, reading span does not adopt a simple sum of discrete items or average from them. For this reason, no item analysis was performed. Instead, the students' performances on the four different sentence levels were compared using a repeated-measures ANOVA, as presumably, the recall difficulty is a function of the sentence size, and its results should indicate whether or not larger sentence size does coincide with higher difficulty and vice versa. Though a somewhat crude method, it should enable us to check the basic functionality of the RST.

In Osaka and Osaka's (1992) experiment, the test session was terminated at the level where their subject failed in three out of the five trials. In the present study, however, the session was continued until the last of the 5-sentence set was tried in order to analyse the effect of the sentence set size on the recall performance and compare their performances across different sentence set sizes.

Results

Table 5.16 summarises the descriptive statistics on the two measures of passage reading and reading span working memory. The passage reading score mean was 9.26, which is slightly lower than that of the sample for Preliminary Study D, whose mean was 10.07. As expected from the smaller sample size, the present group had a much narrower SD of 2.70, as opposed to 4.38 of the aforementioned sample. The reading span mean was a low 1.89 with an SD of 0.99. Pearson product-moment correlation between these two variables was also a low .23 (n.s.).

Table 5.16 Results of passage reading and reading span test – version 1 (n=38)

Variable	Min	Max	Mean	SD
Passage reading	5	14	9.26	2.70
Reading span	0.0	4.0	1.89	0.99

Success rates for the different sentence levels of the RST are listed in Table 5.17. The smallest sentence sets had the highest success rate and vice versa as expected. The data were subjected to repeated-measures ANOVA, which indicated that there was a significant difference among the levels ($F_{(3, 111)}$=84.05, p<.001; see Table 5.18). Post-hoc pairwise comparisons via Bonferroni's adjustment suggested that, except between 4-sentence and 5-sentence levels, the inter-level differences were statistically significant (p<.001).

To summarise, the difficulty order of the four different levels of the Reading Span Task was:

2-sentence level < 3-sentence level < 4-sentence level and 5-sentence level

Table 5.17 Recall performance by sentence level (n=38)

	Mean	SD
2-sentence level	2.79	1.21
3-sentence level	1.60	0.82
4-sentence level	0.58	0.76
5-sentence level	0.34	0.53

Table 5.18 ANOVA table: within subject effect of sentence level

Source	SS	df	MS	F	Sig.
Sentence level	142.342	3	47.447	84.054	.000
Error	62.658	111	0.564		

Discussion and conclusion

The non-significant correlation between reading span and passage reading might suggest there was little link between these two variables at least as far as the present sample group is concerned. Since low correlation can result from factors other than pure lack of relationship, a careful examination of several aspects of the available data is recommended before we decide whether or not to continue to include a variable as a predictor.

First, as noted earlier, the variance of the criterion passage reading measure is rather small for the current sample, and it might have contributed somewhat to the deflated correlation. Also of significance is the low overall mean of 1.89 on the RST. Osaka and Osaka's (1992) sample had a mean span of 3.23 with an SD of 1.10 on the same version of the RST. It is thus possible that this test is more suitable for higher proficiency groups than the students in the present sample. Since Osaka and Osaka's group had no one who failed the 2-sentence level while the current sample had six such students, the data were reanalysed in a post-hoc fashion to explore the effects of these extreme low performances on the RST. Table 5.19 describes the basic statistics from the sample when the six persons with zero span were removed from the analysis.

Table 5.19 Results of passage reading and reading span test – version 1 Removing extreme low span cases (n=32)

Variable	Min	Max	Mean	SD
Passage reading	5	14	9.25	2.77
Reading span	1.5	4.0	2.17	0.63

Whilst almost no change is observed in the passage reading score mean (9.26 to 9.25), there was change in the reading span mean in the expected direction (from 1.89 to 2.17). Although this reduction in the sample size also accompanied shrinkage in span variance (SD=0.63 compared to previous 0.99), the two variables now correlated significantly ($r=.44$, $p<.05$). Compared to the much higher correlations observed between passage reading and such variables as syntactic knowledge and vocabulary breadth in the previous preliminary study, this is merely a moderate correlation. Nevertheless, it is possible that working memory influence may be larger among higher proficiency subgroups, which our subsequent main study will include, or that it might account for a unique portion of the passage reading variance. Therefore, a slightly conservative judgment was made not to terminate this variable at this stage of the project but to accommodate it in the final test battery. The ANOVA results from this group confirmed that the 2-sentence level was the easiest and the 4-sentence and 5-sentence levels were the most difficult. The differences among the lower three levels suggest that the test was not grossly misbehaving and that these level divisions were meaningful. From the perspective of research economy, the lack of any difference between 4-sentence and 5-sentence levels is taken to lend support for removal of the 5-sentence level. Although the maximum reading span was as high as 4.0, this was recorded by only one out of the 38 students. And the next highest span of 3.5 was also obtained by a single student. These two students constituted the top 6% of the group. The remaining 94% failed to reach this level and, in fact, 85% received a span score of 2.5 or lower, which means even the 4-sentence level items were unnecessary for the great majority of the students. Therefore, a decision was made to remove the 5-sentence level from the next stage of the project.

Preliminary Study F: materials for reading speed measure

Introduction/purpose

The present study requires an adequate set of English sentences to measure the Japanese university student's sentence reading speed during comprehension. Since the object of measurement is the individual differences in reading speed during comprehension and not in the ability to comprehend the sentences quickly, the target sentences should be understandable for most of the students. From Shizuka's (2000) 40 sentences, which he used for his CB sentence reading test, the present preliminary study attempts to identify the easiest 10 sentences to be adopted in the subsequent main study.

Participants, procedures and method of analysis

Participants were 32 students at an institute of higher learning in northern Kyushu. They responded on paper to the 40-sentence reading comprehension items originally developed for CB testing by Shizuka (2000). Each item required the student to read a single sentence followed by four choices in Japanese, only one of which included ideas presented in the target English sentence. The test was group-administered during their class time as part of their EFL class work and 30 minutes were allowed for this test. The students' responses were dichotomously scored and item facility values were compared to select the easiest 10 items.

Results and conclusion

The group mean of 20.28 with an SD of 5.63 was obtained for this 40-item test. The item facility values ranged between .03 and .91 with a mean of .51. Table 5.20 lists the 10 items that had the highest item facility values. These 10 items will thus be employed in the subsequent main study to measure the individual differences in the speed of reading sentences on the computer screen.

Table 5.20 Top 10 sentence reading items by facility value

Item no.	Facility	SD
16	.906	.296
30	.875	.336
3	.844	.369
13	.813	.397
17	.750	.440
27	.750	.440
6	.719	.457
12	.719	.457
32	.688	.471
9	.656	.483
Mean	.772	.415

Chapter review

This chapter has described in detail three additional preliminary studies conducted subsequently to the initial three.

Study D has firstly provided evidence that the three revised PB measures of passage reading and syntactic and vocabulary breadth have acceptable levels of reliability. It has also shown that the CB word recognition latencies can be measured with a reasonable level of reliability. The study also identified problematic synonym/antonym pairs based on response accuracy and helped refine the item set. The regression results from the same dataset have additionally indicated that syntactic and vocabulary breadth account for the passage reading test performance much better than does the word recognition efficiency, and the partial correlation results suggested that word recognition efficiency can have a significant overlap with passage reading performance when the task involves lexical-semantic access.

The chapter then reported on Study E, which examined whether Osaka and Osaka's (1992) ESL version of the RST is suitable as a measure of the working memory span of the prospective sample population and whether the ability assessed via this instrument correlates to a meaningful extent with the passage reading results of our preliminary sample. The comparisons of the recall success rates from the four span levels indicated that the span level divisions are meaningful among the lowest three levels but not between the highest two, which has lent support for the removal of the most time-consuming 5-sentence level set from the instrument. The results of the analysis after removing the extreme low span individuals suggested that the RST and the passage reading do correlate significantly with each other, and a decision has been made to retain the RST in the test battery for the final main study.

The last of the preliminary studies identified the 10 sentence reading items with the highest item facility values from a set of 40 which are to be used in the main study as the target sentences in the assessment of CB reading speed.

6 The main study: background data

Background

As has been specified earlier in the Methodology chapter, the main study adopts multiple regression analysis as its primary statistical tool for investigating observed skill variables that may account for individual differences in Japanese EFL learners' reading comprehension performance, although the data are also subjected to factor analysis to explore the latent trait structure of the measured variables. Confirmatory factor analysis (CFA) follows this exploratory factor analysis (EFA) to evaluate the adequacy of the factor structure obtained from the EFA and to statistically test the significance of specific links among the research variables.

Research variables

The actual research variables and their measurement instruments have been determined as a result of theoretical and practical considerations and preliminary empirical studies reported in Chapters 4 and 5. Specifically, this study involves the following research variables: (1) Passage Reading Comprehension, (2) Computer-Based Sentence Reading Speed, (3) Syntactic Knowledge, (4) Vocabulary Breadth, (5) Number Matching Latency, (6) Real Word Matching Latency, (7) Pseudoword Matching Latency, (8) Irregular Letter String Matching Latency, (9) Synonym/Antonym Decision Latency, (10) Lexicality Effect, (11) Orthography Effect, (12) Reading Span Working Memory, and (13) Language Independent Metacognitive Knowledge about the Text and Reading. Among these variables, some are measured via PB tasks and others via CB tasks.

Computer-based test apparatus

The computer equipment used for the CB instruments is Apple Macintosh PowerBook 5300c, a laptop model with a 10.4 inch active matrix colour LCD display. The stimulus items have been programmed to appear in black letters/numbers in the centre of the display screen over the white background.

Final modifications to the MCQ-based power tests

Since the time of the preliminary piloting, one important addition has been made to make the results of MCQ-based measures more reliable following recent doctoral research conducted in the Japanese EFL context.

Of the three PBT instruments in the present study, the Passage Reading (k=20) and Syntax instruments (k=32) have relatively narrower score ranges compared to the Vocabulary Breadth instrument (k=60). Since test length strongly influences test reliability, it is not surprising that the preliminary studies yielded lower reliability coefficients for Passage Reading (alpha=.80) and Syntax (alpha=.82) compared to Vocabulary Breadth (alpha=.94). These measures are MCQ based and, like any test method, they have certain shortcomings. MCQ-based measures are structurally more prone to the effects of guessing, and such influences are more likely to be a threat to the reliability of tests with a relatively smaller number of items. While alpha values of .80 and .82 are already reasonably high, any valid and efficient method of improving them further and minimising the effects of guessing should be welcome.

A recent language testing study has produced an innovative method of improving the reliability of tests with dichotomously scored items by incorporating information on the candidates' level of confidence on the individual items. Shizuka (2000, 2004) tested, among other things, his hypothesis that 'response confidence contains information about the reader's ability that is not revealed by response correctness in such a way that when response correctness is kept constant, more confident test-takers are more proficient' (2000:274), and he developed what he calls Clustered Objective Probability Scoring (COPS). In COPS, the candidates are asked to rate their level of confidence on each of their responses to the test items along with their answer choices, and the score to be awarded to a response is based on the percentage correct of the responses clustered by the same confidence levels. Shizuka (2004:179) illustrates the system through a hypothetical 10-item test given to five types of candidates. Table 6.1 captures a part of Shizuka's own illustration. It depicts five candidates whose patterns of correctness are identical but whose ratings of their own confidence level on individual items vary significantly. Correct responses are represented by 1s and incorrect responses by 0s in the top half of the table. In the bottom half, the letters H, M, and L denote high, medium, and low confidence levels respectively.

The scoring begins by computing the objective ratio of correct responses for each group or cluster of items receiving a common confidence level within each person. Again, Shizuka's (2004:180) own example should serve best. Table 6.2 is a result of rearranging the items for each person in order that the items are clustered on the basis of their subjective confidence ratings within each person. The third row for each person shows the probability correct for

Table 6.1 Response correctness and reported confidence levels of Shizuka's five hypothetical cases

		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Total
Person 1	Correctness	0	1	0	1	1	1	0	1	0	1	6
Person 2		0	1	0	1	1	1	0	1	0	1	6
Person 3		0	1	0	1	1	1	0	1	0	1	6
Person 4		0	1	0	1	1	1	0	1	0	1	6
Person 5		0	1	0	1	1	1	0	1	0	1	6
Person 1	Confidence	L	H	M	H	H	M	L	L	L	M	
Person 2		H	H	H	H	H	H	H	H	H	H	
Person 3		L	L	L	L	L	L	L	L	L	L	
Person 4		L	M	H	L	M	H	L	M	H	L	
Person 5		L	H	L	H	H	H	L	H	L	H	

Source: Shizuka (2004:179)

Table 6.2 Shizuka's COPS applied to five hypothetical cases

Person 1 'The realist'											**Total**
Correctness	1	0	0	0	1	1	0	1	1	1	6
Confidence rating	L	L	L	L	M	M	M	H	H	H	
Cluster probability	.25				.67			1.00			
COPS item score	.25	0	0	0	.67	.67	0	1.00	1.00	1.00	4.59
Person 2 'The optimist'											**Total**
Correctness	1	1	1	1	1	1	0	0	0	0	6
Confidence level	H	H	H	H	H	H	H	H	H	H	
Cluster probability	.60										
COPS item score	.60	.60	.60	.60	.60	.60	.00	.00	.00	.00	3.60
Person 3 'The pessimist'											**Total**
Correctness	1	1	1	1	1	1	0	0	0	0	6
Confidence level	L	L	L	L	L	L	L	L	L	L	
Cluster probability	.60										
COPS item score	.60	.60	.60	.60	.60	.60	.00	.00	.00	.00	3.60
Person 4 'The randomizer'											**Total**
Correctness	1	1	0	0	1	1	1	1	0	0	6
Confidence level	L	L	L	L	M	M	M	H	H	H	
Cluster probability	.50				1.00			.33			
COPS item score	.50	.50	.00	.00	1.00	1.00	1.00	.33	.00	.00	4.33

Table 6.2 (continued)

Person 5 'The clairvoyant'											Total
Correctness	0	0	0	0	1	1	1	1	1	1	6
Confidence level	L	L	L	L	H	H	H	H	H	H	
Cluster probability	.00				1.00						
COPS item score	.00	.00	.00	.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00

Source: Shizuka (2004:180)

each of H, M, and L clusters (Cluster Probability). Each correct response is then assigned the Cluster Probability value to yield a COPS item score, which is found in the fourth row. Person 1, for instance, was correct on one out of four items on which she rated her confidence to be low, therefore the Cluster Probability for her low-confidence items is .25, which becomes the COPS item score for her correct low-confident response. Her medium-confidence Cluster Probability was .67, which is assigned to each of the two correct medium-confidence responses. Her high-confidence items were all correct yielding a probability of 1.00, which is assigned to each of her three correct responses. It is by summing these COPS item scores that her overall individual score of 4.59 is obtained.

Partial knowledge is a common phenomenon in L2 development, but simple dichotomous scoring fails to distinguish between those who respond correctly and confidently to an item with very good knowledge of the point tested and those who have only partial but enough knowledge to guess the answer correctly. Shizuka's examples illustrate how his system can address this point. Persons 1 and 6 gave exactly the same responses to the 10 items, but Person 1 declared 'middle' confidence on two of her correct responses, presumably due to lesser knowledge of the content tested on these items. Person 6, on the other hand, declared 'high' confidence on all of the six correct responses, presumably because of very clear idea of the tested content, and received a higher total score of 6.00.

Some candidates may have a tendency to give higher confidence ratings consistently while others may constantly underrate themselves. Incorporating the observed response correctness for each confidence cluster rather than blindly assigning pre-set values to candidates' self-ratings, COPS successfully avoids rewarding the constantly overconfident (cf. Person 2) more than those who indicate unreasonably low confidence (cf. Person 3).

In reality, few candidates are likely to respond as extremely as Shizuka's optimist, pessimist, etc., but they are expected to be somewhat more realistic, for example like his 'realist', with varying ratios of declared confidence levels. Demonstrated in Shizuka's hypothetical cases, however, is the potential for

COPS to account for the individual differences in the general level of optimism/pessimism in predicting their own success on test items and for the system to reward accurate self-ratings.

Results of Shizuka's pilot testing with his Japanese EFL students supported a system with four confidence levels expressed by H ('absolutely sure'), M ('sure'), L ('not very sure'), and Zero ('no idea at all'), which rewards only the correct responses with a confidence level of L or above. Shizuka also confirmed that COPS does not complicate the test taking nor lengthen the required test time. It is therefore decided that the Passage Reading and Syntax instruments in this main study adopt COPS with the four confidence levels of High, Middle, Low, and Zero, with the Zero confidence items receiving no score regardless of their response correctness.

There is another useful consequence of adopting COPS for the passage reading test. The test has four passages with five MCQ items each. It would be informative if each of the four passage scores could be analysed in a separate multiple regression so that the stability of the results could be checked. Only unreliable results could have been anticipated from the original dichotomous scoring, which only gives a narrow range from zero to five for each passage. With COPS, however, since each item is assigned a COPS item score which is expressed as a percentage, the passage score can have a range between 0 and 500. Such finer discrimination is expected to improve the reliability of these sub-section measures significantly.

Updated measurement instruments

As detailed in Chapters 4 and 5, all variables except Language Independent Metacognitive Knowledge underwent preliminary testing with tertiary level Japanese students. Since the measure of Language Independent Metacognitive Knowledge became available at a later stage in this research project, concurrent analysis of this measure is required based on the responses of the students in this main study, unlike the other 12 variables previously piloted. Since most aspects of the measurement instruments have already been detailed in Chapters 4 and 5, the accounts to follow aim to review each instrument briefly, and focus is given to some final revisions made since the time of the preliminary testing.

Passage reading comprehension

The PBT type measure of the students' Passage Reading Comprehension consists of four expository reading passages and a total of 20 4-choice MCQ items divided into four subsets of five items that are attached to each of the four passages. The measure thus has a possible raw score range between 0 and 20. The COPS procedure (Shizuka 2000, 2004) is incorporated to improve the reliability of this measure having a relatively small number of items. Details of this instrument have been presented in Chapters 3 and 4.

Computer-based sentence reading speed/time

The CB measurement of Sentence Reading Speed adopts 10 sentences, each of which is followed by a 4-choice MCQ. Figure 6.1 illustrates the steps involved in each item of this test. The rectangles in the Figure represent the actual contents of the screen frames which appear at appropriate key presses. Except for the target English sentence in the second frame, all English translations in the parentheses are inserted solely for the purpose of illustration here and were unavailable during the actual test.

Facing a laptop computer, the student presses a designated key to start an item, which makes the target sentence appear in Times size 18 font. The student then reads it as quickly as possible for comprehension and, having reached the end of the sentence, presses the same key again. This second key press simultaneously prompts the computer to record the time elapsed since the initial key press and to display the post-reading answer choices, lined up vertically in Osaka size 14 font. From the four Japanese answer choices displayed, the student is to choose the one which expresses the idea contained in the target English sentence and presses the corresponding answer key. The computer records the student's decision and the time spent for that decision. The students are told that each of their sentence reading and answer selection is timed and that it is important that they comprehend the sentences and answer correctly but that they do so as quickly as they can. There were detailed individual instructions given orally as well as two practice items preceding the actual test items.

The 10 items are randomly sequenced by the computer for each student. For each student, time spent on each sentence is first divided by the number of syllables contained in the sentence to obtain the sentence reading time in the unit of milliseconds per syllable, and the mean reading time on the items is calculated based only on the items whose comprehension questions are answered correctly.

Syntactic knowledge

The PBT type Syntactic Knowledge measure employs the item format found in the Structure section of TOEFL. Each item requires the student to read an English sentence, a part of which is left blank, and four semantically similar but syntactically distinct choices are provided so that the student is asked to choose the one which would fit in the blank in terms of the syntactic acceptability of the completed sentence. There are a total of 32 Syntax items, thus the possible raw score range for this instrument is between 0 and 32, but the

Figure 6.1 Progression of the computer-based sentence reading task

1. Each item begins with the following idle screen.

赤いキーで提示します
(Press the Red key to start.)

2. The first key press here will prompt the target sentence such as below to appear while setting off the timer.

My brother likes to watch sports on television when he has free time.

3. As soon as the test taker has read the target sentence like the above, s/he is to press the same key again, which will simultaneously stop showing the sentence, record the time elapsed since the initial key press, and display the answer choices such as below.

1. いつも暇だ
   (is/are always free)
2. テレビ観戦が好き
   (likes to watch sports on TV)
3. スポーツ選手の空き時間
   (athlete's free time)
4. 兄弟が欲しい
   (want/s a brother/brothers)

4. S/he then presses one of the four number keys corresponding to her choice from among the four on the screen above, which will simultaneously stop showing the choices, record which key was pressed as well as the time elapsed since the previous key press, and start the subsequent item with the idle screen on top.

COPS procedure (Shizuka 2000) is incorporated which is expected to lead to much finer discrimination within that range. This instrument is detailed in Chapters 3 and 4.

Vocabulary breadth

The PBT measure of Vocabulary Breadth adopts the format found in the Vocabulary Levels Test and asks the student in each task to choose from a list of six words the one which matches a given definition. Each task contains three question items such that three out of six words must be chosen that match the three definitions respectively. The measure has a total of 20 tasks thus 60 items. The Vocabulary Breadth instrument is detailed in Chapters 3 and 4.

Word recognition latency

As detailed in Chapter 5, all of the CB word recognition efficiency instruments require the student to decide as quickly as possible whether each of the stimulus pairs shown on the computer display consists of two matched ('same' or 'synonymous') or unmatched ('different' or 'antonymous') items. As in the preliminary study, although in reduced item sets, the current latency test is administered in the three separate blocks of Number (k=18), Alphabetic String (k=24), and Synonym/Antonym (k=18), with the Alphabetic String block further consisting of the Real Word (k=8), Pseudoword (k=8), and Irregular Letter String pairs (k=8). The basic format and procedure are unchanged from the preliminary study (see Chapter 5), although there have been a few revisions in the details. Namely, all of the number matching stimuli and a few of the alphabetic string stimuli have been renewed to improve balance across items and stimulus types, and additionally, the computer is re-programmed such that, within each block, the test items are randomly sequenced for each session with no item reused with the same individual. Unlike the preliminary study, therefore, latency values on all of the stimulus items are obtained for each individual while the randomness of the stimulus presentation is now achieved only in the sequencing of the items. Additionally, the display font is enlarged to size 48, and the target stimulus items are aligned vertically rather than horizontally when displayed on screen.

As in the preliminary study (cf. Chapter 5), the degree to which each person's word recognition is slowed down by the lack of orthographic regularity is assessed by subtracting the decision latency under Pseudoword condition from that under Irregular Letter String condition (Orthography Effect), whereas the extent to which someone's word recognition is slowed down by the lack of lexical familiarity is assessed by subtracting the decision latency under Pseudoword condition from that under Real Word condition (Lexicality Effect).

Working memory

Working memory is measured through an ESL version of the Reading Span Test (RST) described in Chapter 5. The basic structures of the test and the target items are kept the same, but there are several revisions that have been made for the present main study. Specifically, the 5-sentence set level has been removed, and the sentences are delivered on the computer screen instead of paper cards. This enables random sequencing of the trial sets within the given set size level, as well as simultaneous and accurate measurement and recording of the student's reading and answering time, on the basis of which outliers may be identified. The student proceeds by pressing the designated keys to start the trial and display the target sentences one by one. At the end of each trial, question marks appear on the screen to prompt recall of the target words. As in the preliminary study, the target word recall is hand-recorded by a test proctor present for each individual RST session. To summarise, five sets each of two, three, and four target sentences are presented to each student and all the sentence-final words are to be recalled after each set. The RST scoring procedure is the same as described in Chapter 5.

Language-independent metacognitive knowledge of text and reading

As described in Chapter 3, a 38-item questionnaire is administered to elicit responses from the Japanese students on their language-independent metacognitive awareness of the text in general (k=14) and reading strategies (k=24) in the 'yes'/'no' answer format. Simple sums of responses that are judged to be associated with proficient reading by the original questionnaire developers are adopted, making the scale range from 0 to 38.

Participants

The sample population in the present main study consists of L1-Japanese EFL students from five different universities located in four cities in western Japan. The composition of the present sample group is by no means a result of planned stratified sampling. It may be recalled from the Methodology chapter that the participants are recruited in EFL courses whose instructors have agreed to administer the PBTs in their classes and invite their students to participate in the paid CB tasks, and that it is the availability of such instructors which determined the composition of the sample population.

As summarised in Table 6.3, two of the universities are private institutions and the other three are public, of which two are nationally owned and one prefecturally owned. Although one needs to be aware of the dangers of simplistic ranking of institutions, the EFL proficiencies of the university entrants in Japan are likely to differ depending on how competitive the admission is to the institution. Despite recent introduction of diverse admission processes, PB entrance examination is still one of the most popular methods of applicant evaluation at Japanese universities, and EFL is very often a required subject in those examinations. Public universities are generally more competitive in their own localities unless there are exceptionally popular private institutions. The participants at the three public institutions may show advantage over those at the two private institutions, thus their group means will be examined as part of a background analysis later in this chapter.

In terms of the enrolment, Universities B, C, and E are quite large with over 10,000 matriculated undergraduates, while university A can be considered middle-size with an approximate enrolment of 7,000. University D may be regarded as belonging to a small-size category with about 1,300 students. Universities B and C are located in the same city while the others are from different cities. Universities B, C, and E are located in fairly large cities, whereas University D is in a small city.

Table 6.3 Participating institutions

	Type	Approximate undergraduate enrolment	Location (city population)
University A	Private	7,000	City 1 (230,000)
University B	Private	20,000	City 2 (1,300,000)
University C	National	10,000	City 2 (1,300,000)
University D	Prefectural	1,300	City 3 (50,000)
University E	National	10,000	City 4 (620,000)

The participants in this study belong to one of 10 different academic disciplines listed in Table 6.4. Nine of the 11 participants from University E are postgraduate students, but the remaining 210 are all undergraduates. English majors constitute the largest subgroup (n=60), a large percentage of which is female (n=50). Since English majors normally have higher motivation for improving their EFL proficiency and are receiving more hours of EFL training, they tend to outperform the non-English majors if at the same university. Especially, all of the postgraduate students and the remaining upper-level undergraduates at University E are English education majors at a national university, and they can be expected to display clear advantage over most of the non-English majors. Group means of the 60 English majors and the 159 non-English majors will be examined later in this chapter. Among the five universities, University B has the largest number of participants (n=83) and University E the smallest (n=11). Overall, the male−female ratio is 116 to 103.

Table 6.4 Breakdown of participants by institution, major, and sex

		University A		University B		University C		University D		University E		Major Total	
Business	(M)	38	(22)	15	(8)							53	(30)
	(F)		(16)		(7)								(23)
Economics	(M)					9	(8)					9	(8)
	(F)						(1)						(1)
Engineering	(M)			4	(4)	33	(33)					37	(37)
	(F)				(0)		(0)						(0)
English	(M)			49	(7)					11	(3)	60	(10)
	(F)				(42)						(8)		(50)
Law	(M)	19	(16)	1	(0)							20	(16)
	(F)		(3)		(1)								(4)
Nursing	(M)							4	(0)			4	(0)
	(F)								(4)				(4)
Nutritional science	(M)							8	(1)			8	(1)
	(F)								(7)				(7)
Welfare and health science	(M)							12	(2)			12	(2)
	(F)								(10)				(10)
Sports science	(M)			14	(11)							14	(11)
	(F)				(3)								(3)
Visual design	(M)							2	(1)			2	(1)
	(F)								(1)				(1)
Institution total	(M)	57	(38)	83	(30)	42	(41)	26	(4)	11	(3)	219	(116)
	(F)		(19)		(53)		(1)		(22)		(8)		(103)

The great majority of the participants (96.8%) belong to the typical age range for the Japanese university students: from 18 to 23. Six are in their higher twenties (2.8%) and one in her thirties (0.5%). The arithmetic mean of the participant age is 19.96, and the median is 20. Most participants also have received a similarly ordinary length of EFL instruction with a mean of 8.5 years and a median of eight years. Japanese students typically receive six years of English at secondary schools and, at university, two more years in the case of non-English majors and longer in the case of English majors. Some also start English training at primary schools. About 31.1% of the current sample are in their seventh year of English learning (68 cases), another 30.1% in their eighth year (66 cases), another 30.2% in their ninth or tenth year (64 cases), and the remaining 8.7% mostly in their eleventh and twelfth year with only a few exceptions reporting longer English learning history. Three students have lived overseas for a combined total length of between 10 months to two years each, and 23 others have stayed in an English-speaking country for about a month or shorter. None of the older students and those with longer history of English learning or living experiences overseas exhibited any unusual patterns of test performance. Therefore, no one has been excluded from the analysis on the basis of their biodata.

Descriptive statistics

The basic descriptive statistics are presented in detail in separate sections for each of the observed variables below.

Passage reading comprehension

The basic descriptive statistics on the Passage Reading results are summarised in Tables 6.5 and 6.6. As seen in Table 6.5, the raw score mean was 11.16 out of the maximum possible score of 20, which is 1.09 points higher than the 10.07 obtained for the preliminary sample studied in Chapter 5. The reliability estimated via Cronbach's alpha was .75, a small decrease from the .80 obtained for the preliminary sample. The passage raw score mean ranged from the highest 3.40 for Passage 2 to the lowest 2.26 for Passage 4. The pattern of differences in alpha values paralleled that of the passage score means with a range between .66 for Passage 2 and .37 for Passage 4. The details of the classical item analysis are presented in Appendix 1.

Table 6.5 Descriptive statistics: passage reading raw score (n=219)

	k	Mean	SD	Min	Max	Skew	Kurt	Alpha
Passage reading total (raw)	20	11.16	4.00	2.00	19.00	−0.11	−0.91	.75
Passage 1 (raw)	5	3.21	1.39	0.00	5.00	−0.27	−0.96	.53
Passage 2 (raw)	5	3.40	1.49	0.00	5.00	−0.72	−0.47	.66
Passage 3 (raw)	5	2.29	1.41	0.00	5.00	0.12	−0.78	.48
Passage 4 (raw)	5	2.26	1.32	0.00	5.00	0.08	−0.70	.37

For a Rasch reliability and item analysis, the data from the 219 candidates were merged with those from the 182 candidates in the preliminary study (see Chapter 5). The overall item reliability estimate of .97 and case reliability estimate of .77 were obtained, and the Infit Mean Square ranged from 0.84 to 1.19 with a mean of 1.00 (SD=0.11) for the items, thus all items were within the recommended range. The range for the 219 cases was from 0.66 to 1.56 with a mean of 1.03 (SD=0.17). Fourteen persons were placed above the recommended range and seven below it. Person ability estimates in logits had a mean of 0.32 (SD=1.04) with a range between −2.4 and 3.14. The item difficulty, in logits, ranged between −1.41 and +1.15 with a mean of 0.00 (SD=0.73) and person ability ranged between −2.4 and +3.14 with a mean of 0.32 (SD=1.04). The details of the Rasch analysis on the Passage Reading items are available in Appendix 2.

Table 6.6 presents the results of the same passage reading test scored differently using the COPS procedure discussed earlier in this chapter. The reduced overall and passage means are to be expected since the procedure works by adjusting the item score downwards from 1.00 but never upwards. The score for each passage roughly decreased by one, which is consistent with the 3.97 gap between the two overall means. The use of the COPS procedure seems well rewarded with the obvious improvement in reliability coefficients. The overall reliability is now remarkably high with an alpha value of .91, but of more significance is the uniformly large increase in the reliabilities for the four passage subsections. Passage 2 had the smallest increase of .18, but all the remaining three recorded an increase of .24 or more.

Table 6.6 Descriptive statistics: passage reading COPS score (n=219)

	k	Mean	SD	Min	Max	Skew	Kurt	Alpha
Passage reading total (COPS)	20	7.19	4.11	0.35	17.20	0.38	−0.81	.91
Passage 1 (COPS)	5	2.20	1.33	0.00	5.00	0.20	−1.06	.78
Passage 2 (COPS)	5	2.37	1.42	0.00	5.00	−0.04	−1.21	.84
Passage 3 (COPS)	5	1.36	1.12	0.00	4.80	0.85	−0.02	.72
Passage 4 (COPS)	5	1.26	0.98	0.00	4.65	0.74	−0.08	.63

Skewness and kurtosis values are included to detect significant deviations from normal distribution. Purpura (1998) warns that skewness and kurtosis coefficients exceeding 3.00 in absolute values are problematic. The values for the Passage Reading are well within that range. Detailed item data based on COPS are found in Appendix 3.

Table 6.7 shows the correlations between the two methods of scoring for each of the passage score pairs as well as for the total score pair. The correlations are generally high and indicate that the finer discrimination achieved via COPS does not accompany drastic changes in the positions of the candidates within the sample.

Table 6.7 Correlation between two methods of scoring (n=219)

	r.
Passage reading total	.95
Passage 1	.93
Passage 2	.92
Passage 3	.91
Passage 4	.90

all significant at p=.000

Computer-based sentence reading speed/time

Table 6.8 summarises the results of the CB sentence reading test, which aimed to measure the participants' sentence reading time.

Table 6.8 Descriptive statistics: CB sentence reading (n=219)

	k	Mean	SD	Min	Max	Skew	Kurt	Alpha
Reading time (all 10 items mean)	10	534.87	200.62	146.15	1291.84	1.16	1.46	.92
Reading time (correct responses mean)	variable	527.98	194.11	128.19	1146.98	0.88	0.27	N/A
Comprehension	10	5.42	2.20	1	10	−0.01	−0.90	.57

The first two rows present the reading time data. Individual means of the reading time for the 10 target sentences are further averaged to obtain the sample mean, which was 534.87 milliseconds per syllable. The second row shows the mean reading time for the sample when only the reading time for sentences with correct comprehension responses is averaged. Because of the missing values due to some incorrect responses, no alpha value is available for this method of analysis.

The third row shows the data on the 10 comprehension questions attached at the end of each target sentence. It has to be emphasised that individual differences in comprehension ability are not the primary concern of this measure and that the post-reading comprehension questions served to encourage reading with at least some level of comprehension and discourage extreme scanning or mere key pressing without much visual and linguistic processing. In that sense, low within-sample variance and reliability were to be expected for the comprehension measure compared to the time measures. While the SD values do not show such expected differences between the comprehension and time measures, the alpha values do (see Appendix 4 for detailed data from the classical item analysis).

It is recalled that the 10 items have been selected on the basis of their facility values obtained in a preliminary study (see Chapter 5). In the preliminary study, the facility mean for the 10 items was 77.2%. The CB sentence reading items yielded a much lower mean of 54.2%. One likely cause of this decrease is the fact that in the CB task each target sentence is removed from view when the answer choices are shown, whereas in the preliminary PB task the target sentences could always be referred back to. Also, the time pressure must have contributed to the depressed mean. Underlying differences in general English proficiency between the preliminary and the main study samples may be suspected as well. However, the preliminary sample consisted of students of a non-English major at a tertiary-level institution which is not a university. It is therefore unlikely that the preliminary sample had a substantive proficiency advantage over the main study sample.

Table 6.9 presents the correlation coefficients among the comprehension score and the two methods of obtaining the sentence reading time. There is a noticeable lack of correlation between the comprehension and time (r=−.04 and −.01, n.s.) and a clearly significant correlation between the two different methods of calculating the reading time variable (r=.95, p<.001). However, reading time means based on correctly responded items are adopted for further multivariate analysis, as originally planned, rather than that based on all 10 items, because the low comprehension mean of 5.42, out of the maximum possible 10, (cf. Table 6.8 above) is insufficient to justify the latter.

Table 6.9 Correlations among CB sentence reading variables (n=219)

	Reading time (all items mean)	**Reading time (correct items mean)**
Comprehension	−.04	−.01
Reading time (all items mean)		.95***

****p<.001*

Syntactic knowledge

Descriptive statistics on the Syntax measure are presented in Table 6.10. The raw Syntax mean for the present sample was also higher than that for the preliminary sample, by 1.54 (16.95 versus 15.41). The reliability as estimated via Cronbach's alpha was .81, which is roughly equivalent to that obtained in the preliminary study (.82). The item data based on the dichotomous scoring method are detailed in Appendix 5.

Table 6.10 Descriptive statistics: syntactic knowledge measure (n=219)

	k	**Mean**	**SD**	**Min**	**Max**	**Skew**	**Kurt**	**Alpha**
Syntax (raw)	32	16.95	5.59	4.00	29.00	0.03	−0.63	.81
Syntax (COPS)	32	11.06	6.07	1.15	27.11	0.44	−0.62	.93

The raw response data were merged with the data from the preliminary study (Chapter 5) and submitted to a Rasch reliability and item analysis, which yielded an item estimate of .99 and a case estimate of .82. Infit Mean Square range was between 0.81 and 1.19 for the items and between 0.55 and 1.66 for the persons, with 18 persons placed below the recommended range and 10 above it. The item difficulty, in logit, ranged between −1.61 and +2.54 with a mean of 0.00 (SD=1.10) and person ability ranged between −2.32 and +2.75 with a mean of 0.15 (SD=0.98). The details of this final Rasch analysis on the Syntax instrument are presented in Appendix 6.

The expected decrease in the mean and increase in the reliability due to the use of COPS are observed for this variable as well (see Appendix 7 for item data based on COPS). The two methods of scoring correlated at r=.97 (p<.001) reconfirming that the COPS alters the positions of the candidates within the sample very little.

Vocabulary breadth

Table 6.11 presents the basic descriptive statistics on the Vocabulary Breadth measure. The mean from this sample was 17.26, which was again higher, by 3.12 points, than the group in the preliminary study (see Chapter 5). The longest of all the measures with 60 items, the Vocabulary Breadth measure obtained the highest alpha value of .95 (see Appendix 8 for details).

Table 6.11 Descriptive statistics: vocabulary breadth measure (n=219)

k	Mean	SD	Min	Max	Skew	Kurt	Alpha
60	27.26	13.31	1.00	60.00	0.13	−0.80	.95

Rasch analysis also indicated that the measure has a high reliability. The item estimate was .99 and the case estimate .95. One out of 60 items had an Infit Mean Square of 1.57 exceeding the recommended range, but all others were within it. The mean of this item fit index was 1.00 with an SD of 0.14. One person had a perfect score and a total of 44 persons had Infit Mean Square values below or above the recommended range. The item difficulty estimates ranged from −4.09 to +3.06 with a mean of 0.00 (SD=1.76), whereas the person ability estimates were between −5.52 and +3.34 with a mean of −0.33 (SD=1.66). The Rasch estimate details for the Vocabulary Breadth measure are found in Appendix 9. Since the raw score and the logit score had a correlation of .99 and the use of the logit score would entail losing one candidate who had a perfect score on this measure, the raw score is used for the Vocabulary Breadth variable in the subsequent analysis.

Word recognition latency

Though it is the latency and not accuracy that is being examined through the speeded same-or-different decision tasks, it has to be established that the latency values are based on visual and cognitive processes accompanying a reasonable level of decision accuracy. Table 6.12 thus presents accuracy statistics from the same-or-different speeded decision for the five stimulus types. All types had a mean accuracy exceeding 90% with fairly small within-group variation. With the revised set of stimulus pairs, the accuracy means are higher for all of the five types in this main study ranging from .91 to .95 in comparison to their earlier prototypes in the preliminary study (Chapter 5), which had an accuracy range from .85 to .94.

Table 6.12 Accuracy of speeded matching tasks (n=219)

	Mean	SD	Min	Max
Number	0.95	0.05	0.78	1.00
Word	0.96	0.07	0.63	1.00
Pseudoword	0.95	0.08	0.63	1.00
Irregular string	0.91	0.10	0.50	1.00
Synonym/antonym	0.91	0.09	0.56	1.00

The actual latency means for the five types of visual recognition stimuli are summarised in Table 6.13. With these revised sets of items, the main study group obtained shorter latency means than the pilot study sample in all of the stimulus types except the Synonym/Antonym decision. Though direct comparison requires caution due to the differences in the methods of estimation, the present version, with fewer items, obtained reliability coefficients which are at least at the same level or higher compared to the pilot version (see Appendix 10 for detailed item statistics). The general order of latency means from the five types of stimulus replicated the pilot data, with the Number Latency recording the smallest mean and the Synonym/Antonym Latency the largest.

Table 6.13 Descriptive statistics: matching latencies (n=219)

	k	Mean	SD	Min	Max	Skew	Kurt	Alpha
Number	18	731.66	146.66	498.94	1362.50	1.22	2.12	.90
Word	8	813.90	175.82	456.88	1475.00	1.12	1.80	.79
Pseudoword	8	831.37	182.63	457.50	1548.00	0.98	1.38	.80
Irregular string	8	928.43	208.37	524.88	1552.75	0.68	0.21	.78
Synonym/ antonym	18	1753.28	474.90	939.28	4120.67	1.37	3.03	.90

Kurtosis of 3.03 for Synonym/Antonym type is marginally above the recommended range of 3.00 and indicates that the distribution of this variable is not fully normal. Therefore, further analysis involving this variable requires caution.

From the high decision accuracy means for all of the stimulus types, a large overlap is expected between the average latency from all of the items and that from the correctly judged items only. This indeed was the case, which can be confirmed in the very high correlations between the two methods of averaging the latencies presented in Table 6.14. These high correlations and the high accuracy means provide support for the use of the latency average from all available items in the subsequent multivariate analysis.

Table 6.14 Correlation between two methods of averaging the latencies (n=219)

	r.
Number	.99
Word	.99
Pseudoword	.99
Irregular string	.99
Synonym/antonym	.98

all significant at p=.000

The arithmetic difference between the Real Word and Pseudoword Latency means is calculated to estimate the effects of the absence of the Lexicality, which was 17.47 msecs on average in this main study. Similarly, the gap between the Pseudoword and Irregular Letter String Latency means constitutes the effect of the lack of regular Orthography, which had a sample mean of 97.06 msecs. As seen in Table 6.15, the kurtosis of 3.67 for Lexicality is above the recommended range of 3.00, indicating that the distribution is not normal. Thus, care must be taken in the use of this variable in any subsequent analyses.

Table 6.15 Descriptive statistics: effects of lexicality and orthography (n=219)

	Mean	**SD**	**Min**	**Max**	**Skew**	**Kurt**
Lexicality	17.47	120.25	−464.75	516.88	0.18	3.67
Orthography	97.06	131.65	−293.25	692.38	0.85	2.91

Working memory

Table 6.16 shows the basic descriptive results of the Reading Span Test (RST) for estimating the participants' L2 Working Memory. The sample mean on this measure was a low 1.68, which is noticeably lower compared to the mean span of 1.89 obtained from the preliminary study (see Chapter 5). Since the same target sentences were used in the preliminary and the main studies and the present sample group outperformed the preliminary sample group in Passage Reading (mean=11.16 vs. 9.26), the reduced working memory span was rather unexpected. A few testing conditions differed across the two administrations. Namely, the target sentences were presented on paper cards to the pilot group and on the computer screen to the main study group. Also, the order of sentence presentation was fixed in the PB preliminary administration but randomised in the CB main study across the candidates. It is unlikely that the ordering of the target sentences resulted in the depression of 'overall' span performance, as opposed to the performances on certain items such as those appearing first, last, etc. However, it is possible that the new CB presentation condition affected the performance of the main study group negatively. The effects of method factors on the measurement of working memory seem to constitute an interesting research question of its own, which is beyond the scope of this current research programme.

Table 6.16 Descriptive statistics: RST working memory (n=219)

Mean	SD	Min	Max	Skew	Kurt
1.68	0.86	0.00	4.00	−0.68	0.32

Language-independent metacognitive knowledge of text and reading

The descriptive statistics on the questionnaire to elicit the participants' Metacognitive Knowledge of the text and the reading strategies are found in Table 6.17.

Table 6.17 Descriptive statistics: metacognitive knowledge (n=219)

	k	Mean	SD	Min	Max	Skew	Kurt	Alpha
Metacognitive knowledge	38	26.79	7.58	9	37	−0.68	−0.98	.91

For this variable, a relatively high mean of 26.79 (71%) was obtained, and the internal consistency reliability estimated via Cronbach's alpha was also quite high (alpha=.91). The results of the classical item analysis are detailed in Appendix 11.

A Rasch reliability analysis followed, which obtained an item estimate of .98 and a case estimate of .87. Infit Mean Square range was between 0.43 and 1.79 for the items and between 0.43 and 1.79 for the persons. Fourteen items were below the ideal range and nine above it. Also, a total of 75 persons were either below or above the ideal range. The estimate of item difficulty ranged between −3.29 to +2.36 around a mean of 0.00 (SD=1.48), while person ability estimate ranged between −1.61 and +4.34 with a mean of 1.44 (SD=1.34).

Although the internal consistency of this measure was high, 23 of its 38 items were outside of the ideal range of model fit and, as shown in Appendix 12, the score distribution was bimodal. Thus, caution would be required in evaluating the contribution of Metacognitive Knowledge as measured through this instrument even if it is found to be a good predictor of reading test performances. The details of this final Rasch analysis on the Metacognitive Knowledge questionnaire are presented in Appendix 12.

Identifying subgroups within sample

The final background analysis to be reported has explored whether subgroups of higher and lower achievers can be identified within the sample. The overall score means on Passage Reading are first listed by institution and major for any obvious differences. The group means are listed in Table 6.18 and also shown on a graph in Figure 6.2. It is easy to see that the group means from the three public universities and the English majors from University B are clustered in the area around the score of 10, whereas those from University A and the non-English majors from University B are placed considerably lower, around the score of 4. It thus seems most logical to form two subgroups: the higher-achiever subgroup consisting of all of the public university students and the English majors at University B, and the lower-achiever subgroup consisting of all of the students at University A and the non-English majors at University B.

Table 6.18 Overall passage reading score by institution and major

	English	Non-English	Institution mean
University A		3.58 (1.94)	3.58 (1.94)
University B	8.93 (3.48)	4.32 (2.17)	7.05 (3.77)
University C		10.07 (3.31)	10.07 (3.31)
University D		9.06 (3.14)	9.06 (3.14)
University E	11.55 (4.60)		11.55 (4.60)
Major Mean	9.41 (3.81)	6.35 (3.91)	7.19 (4.11)

These two subgroups are compared on the two criterion measures of Passage Reading overall score and the CB Sentence Reading Speed. The subgroup means are summarised in Tables 6.19 and 6.20. Not assuming equal variances, the means have been submitted to separate independent t-tests, and the differences have been found to be statistically significant (t=15.07, df=209.86, p<.001, 2-tailed on Passage Reading; t=5.56, df=146.49, p<.001, 2-tailed on CB Sentence Reading Speed). For future investigations into the applicability of findings from the overall sample to subgroups, this division will be adopted.

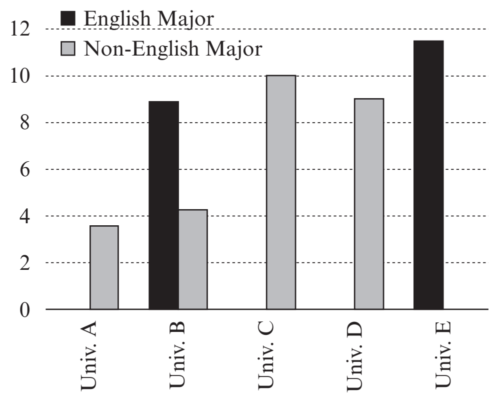
Table 6.19 Passage reading score means for the two subgroups

	n	Mean	SD	SE
Higher achievers	128	9.56	3.52	0.31
Lower achievers	91	3.86	2.05	0.22

Table 6.20 CB sentence reading latency means for the two subgroups (msecs per syllable)

	n	Mean	SD	SE
Higher achievers	128	467.04	147.30	13.02
Lower achievers	91	613.69	218.94	22.95

Figure 6.2 Comparison of passage reading comprehension scores by institution and major



Chapter review

This chapter has detailed the basic descriptive statistics on all of the measured variables after a discussion of a scoring method that is incorporated in some of the measures to improve their reliability. The chapter also discussed the sample population in detail and reported the basic descriptive statistics on all of the measured variables and their reliability coefficients where applicable. Reliability improvement was evident on the instruments which took account of the participants' confidence levels, and the reliabilities of the research instruments have been shown to be generally high. The present chapter also identified two subgroups of differing reading ability, on which separate regression analyses are to be performed subsequently.

7 Results and discussion 1: measured variables accounting for reading performance

Chapter overview

This chapter presents the main results of the analyses of interrelationships among the measured variables and what subskill and knowledge areas best account for the individual differences in passage reading ability and sentence reading speed of the EFL readers at the Japanese universities. Reports on the correlational data precede those on the regression data. The regression data on the passage reading comprehension are detailed first, followed by those on the sentence reading speed.

Interrelationships among the variables

As a background analysis, all of the variables to be treated as the dependent and the independent variables in the main multiple regression analyses are examined for intercorrelations. Correlations involving measures of reading performance are prioritised in our discussion as the main focus of the present study has been on the individual differences in reading skills and how they can be accounted for by factors suggested as significant in the literature by different researchers. The correlational analyses provide an opportunity for initial comparisons of such variables in terms of the degree to which they demonstrate covariation with the reading measures. It should be recalled that the Number Matching Latency has been included in our list of variables in order to use it as a covariate and control for the effect of individual differences in the response latency that is unrelated to alphabetic or linguistic processing speed. In the following two sections, the full bivariate correlations and partial correlations controlling for Number Matching Latency are reported in sequence.

Bivariate correlations

Table 7.1 presents the bivariate correlations among the 17 measured variables. The four passage subsection scores based on COPS are presented

Table 7.1 Simple bivariate correlations among the measured variables (n=219)

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.
1. Passage 1	--	.69***	.65***	.60***	.88***	−.27***	.81***	.71***	−.12	−.04	−.14 *	−.03	−.48***	−.16 *	.15*	.15 *	−.10
2. Passage 2		--	.66***	.48***	.87***	−.18**	.67***	.55***	−.10	−.01	−.11	.00	−.38***	−.16 *	.16*	.13	−.12
3. Passage 3			--	.58***	.85***	−.19**	.70***	.54***	.03	.10	−.02	.10	−.29***	−.17 **	.17**	.10	−.08
4. Passage 4				--	.76***	−.23***	.65***	.57***	−.03	.02	−.03	.02	−.35***	−.08	.08	.07	−.01
5. Passage total					--	−.26***	.84***	.70***	−.07	.02	−.10	.02	−.45***	−.17 *	.17*	.14 *	−.10
6. CB read time						--	−.32***	−.39***	.16 *	.23***	.25***	.22**	.55***	.04	.00	.11	.02
7. Syntax							--	.77***	−.12	−.07	−.13	−.01	−.50***	−.09	.16*	.12	−.17 *
8. Vocabulary								--	−.08	−.06	−.13	−.05	−.60***	−.10	.09	.14 *	−.07
9. Number									--	.66***	.67***	.68***	.44***	.05	.14*	.09	.18 **
10. Real Word										--	.78***	.70***	.35***	−.28 ***	.03	.15 *	.13
11. Pseudoword											--	.78***	.44***	.38 ***	−.15*	.08	.10
12. Irregular string												--	.44***	.17 *	.50***	.14 *	.14 *
13. Syno/antonym													--	.15	.09	.04	.08
14. Lexicality														--	−.27***	−.10	−.04
15. Orthography															--	.11	.08
16. RST–WM																--	.09
17. Metacognitive																	--

Note: *** *p<.001,* ** *p<.01,* * *p<.05* first, followed by the overall Passage Reading comprehension results, also based on COPS. CB Sentence Reading Speed is a time variable whose correlations with PBT power measures are expected to be in the negative direction. The Syntax results are also COPS-based, while the 60-item Vocabulary Breadth is in raw scores. For the five measured latency variables (Variables 9 through 13), smaller values again signify greater performance, thus negative relationships are to be expected with power measures.

All of the intercorrelations among the first six variables representing reading outcomes are statistically significant. The Passage Reading subsections and the overall passage score obtained fairly high intercorrelations ($r=.48$ to $.88$, $p<.001$), although CB Sentence Reading Speed had clearly weaker links to the rest of reading variables ($r=-.18$, $p<.01$ to $-.27$, $p<.001$). The first three of the four passage scores obtained correlations in the mid to high .60s although Passage 4 has somewhat weaker links to them. Their part to total correlations, as observed in Column 5, also display such disparities between Passage 4 ($r=.76$, $p<.001$) and the other three passages ($r=.85$ to $.87$, $p<.001$). One of the possible factors related to this phenomenon may have been the lower reliability achieved by Passage 4 (cf. Chapter 6).

CB Sentence Reading Speed shows somewhat higher correlations with Syntax ($r=-.32$, $p<.001$) and Vocabulary Breadth ($r=-39$, $p<.001$) than with Passage Reading. It is observed in Columns 7 and 8 that the passage subscores and the overall Passage Reading correlated highly and consistently with Syntax ($r=.65$ to $.84$, $p<.001$) and Vocabulary Breadth ($r=.54$ to $.77$, $p<.001$). In the upper half of the remaining columns, low coefficients abound indicating little relationship between the Passage Reading measures and the first four of the latency measures ($r=-.01$, n.s. to $-.14$, $p<.05$). CB Sentence Reading Speed does have some links with them though they are not at all strong ($r=.16$, $p<.05$ to $.25$, $p<.001$). Synonym/Antonym Latency clearly shows a different pattern of relationship with the reading outcomes compared to the other latency variables. It is shown in Column 13 that it has a moderately strong correlation with CB Sentence Reading Speed ($r=.55$, $p<.001$). Though not as strong as with CB Sentence Reading Speed, its correlations with Passage Reading outcomes are also clearly significant ($r=-.29$ to $-.48$, $p<.001$). This disjunction between Synonym/Antonym Latency and the other latencies was also observed in a preliminary study (cf. Chapter 5). The effects of Lexicality and Orthography had only weak relationships with Passage Reading, with correlation coefficients merely as high as .17 ($p=.01$), and they have virtually no relationship with CB Sentence Reading Speed ($r=.04$ and $.00$, n.s.). L2 Working Memory as estimated via RST also failed to correlate with reading variables. Though it recorded statistically significant correlations with the Passage 1 section score and overall Passage Reading, the coefficients are quite small ($r=.15$, $p<.05$ and $r=.14$, $p<.05$). Finally, an index of Metacognitive Knowledge about the text and reading strategies did not correlate significantly with any of the reading variables (r=−.01 to −.12, n.s.).

There are also some correlations that do not involve reading variables but are still noteworthy, such as the ones between Syntax and Vocabulary Breadth (r=.77, p<.001), between Syntax and Synonym/Antonym Latency (r=−.50, p<.001), and between Vocabulary Breadth and Synonym/Antonym Latency (r=−.60, p<.001). All of these variables have shown strong relationships with the reading variables and now among themselves. They are thus likely to be important in a subsequent multiple regression, which explores whether they make unique contributions in accounting for the reading variance.

One more area of interest might be the moderate to high intercorrelations among the latency variables (r=.35 to r=.78, p<.001). Here again, though, Synonym/Antonym shows much weaker relationships with the rest of the latency variables, suggesting fundamental differences in the nature of processing involved.

Extremely high intercorrelations among predictor variables are said to cause an undesirable state in multiple-regression analyses called multicolinearity, and Licht (1995) mentions correlation coefficients higher than .80 as very problematic. Highest among the intercorrelations in the present study was .77 between Syntax and Vocabulary Breadth. Although this was quite close to the suggested figure, examination of Tolerance statistics in subsequent multiple regression analyses identified no predictor variables to be immediately problematic.

Partial correlations controlling for number matching latency

Table 7.2 lists the partial correlation coefficients controlling for Number Matching Latency. From the lack of bivariate correlation between Number Latency and Passage Reading measures, removing from the reading variance that portion shared by Number Latency is expected to cause little change in the coefficients involving reading measures. That indeed is the case. Few notable changes are present in the matrix where the Passage Reading meets the other variables. As regards CB Sentence Reading Speed, the three latency measures of Word, Pseudoword, and Irregular Strings lost some of their original covariation with it so that their correlations are now only as high as .19 (p<.01).

Outside of the sections on reading variables, the largest changes are observed in the intercorrelations among the latency measures. Removing their covariation with Number Latency has caused noticeable reductions in the coefficients. The most remarkable is the plunge in the correlation between Synonym/Antonym and Word Latencies from the original .35 (p<.001) to .08 (n.s.), but the large decrease has been consistent among all of the latency variables.

Table 7.2 Partial correlations controlling for number matching latency (n=219)

	1.	2.	3.	4.	5.	6.	7.	8.	10.	11.	12.	13.	14.	15.	16.	17.
1. Passage 1	--	.69***	.66***	.60***	.88***	−.25***	.81***	.71***	.05	−.08	.07	−.48***	−.15*	.17*	.17*	−.08
2. Passage 2		--	.67***	.48***	.86***	−.17*	.66***	.54***	.08	−.06	.09	−.38***	−.16*	.18**	.14*	−.10
3. Passage 3			--	.58***	.86***	−.20**	.71***	.55***	.11	−.05	.10	−.34***	−.18**	.17*	.10	−.08
4. Passage 4				--	.76***	−.23**	.65***	.57***	.05	−.01	.06	−.38***	−.07	.09	.08	.00
5. Passage total					--	−.25***	.84***	.70***	.09	−.06	.10	−.47***	−.17*	.18**	.15*	−.08
6. CB read time						--	−.30***	−.38***	.16*	.19**	.15*	.54***	.03	−.02	.09	−.01
7. Syntax							--	.77***	.01	−.07	.09	−.50***	−.09	.18**	.13	−.16 *
8. Vocabulary								--	−.01	−.10	.00	−.63***	−.10	.11	.15*	−.05
10. Real word									--	.60***	.45***	.08	−.43***	−.09	.12	.01
11. Pseudoword										--	.60***	.21**	.47***	−.33***	.03	−.03
12. Irregular string											--	.22**	.18**	.56***	.11	.02
13. Syno/antonym												--	.14*	.04	.00	.00
14. Lexicality													--	−.28***	−.10	−.05
15. Orthography														--	.10	.06
16. RST–WM															--	.08
17. Metacognitive																--

Note: *** *p*<.001, ** *p*<.01, * *p*<.05

Summary of the intercorrelations

The data detailed in the present section have indicated that Syntax and Vocabulary Breadth each correlated very strongly with Passage Reading Comprehension, while Synonym/Antonym Decision Latency also had a moderate correlation. The Effect of Orthographic Regularity, the Effect of Lexicality and Reading Span Working Memory had a significant but only weak correlation with Passage Reading, and the remaining variables simply failed to have any meaningful overlap with Passage Reading.

CB Sentence Reading Speed correlated most strongly with Synonym/Antonym Decision Latency, and it also showed significant overlap with Vocabulary Breadth and Syntactic Knowledge. Visual processing latencies demonstrated significant but weak correlations with this speed variable.

The results thus far suggest that it is mainly the breadth of L2 knowledge which predicts Passage Reading Comprehension scores the best, and the only latency variable to have any meaningful overlap with Passage Reading was Synonym/Antonym Decision, which was also the best predictor of CB Sentence Reading Speed. Since these best predictors are highly correlated among themselves, they have to be further evaluated in terms of the unique contributions they make in predicting reading abilities. And, for that we must turn to multiple regression analysis, the results of which are presented in the following sections.

Skill and knowledge areas accounting for individual differences in passage reading comprehension performance: data from multiple regression analyses

The primary purpose of the research project reported in this volume has been to identify skill and knowledge variables which account for the L2 reading comprehension and speed. As has been specified in the Methodology chapter (see Chapter 3), multiple regression is the statistical procedure employed to achieve this goal, since it allows the researcher to identify, on the basis of statistical significance, which of the measured explanatory variables are useful in accounting for the variance in the criterion reading variable. The Number Matching Latency is entered in the equation first regardless of its statistical significance in order to ensure that the variance due to the individual differences in key-pressing speed unrelated to alphabetic or linguistic processing would be accounted for in advance, and thus our comparison of the explanatory strengths of the target latency variables would be free from the effects of such irrelevant individual differences. After the Number Matching Latency, each of the remaining variables is entered into the equation in the stepwise method. That is, each skill variable is assessed in terms of whether adding that variable would improve the regression model significantly (F at $p<.05$), and each time a significant variable is added to the model, all other variables including the ones already found significant are reassessed so that the regression model ends up with the most parsimonious set of explanatory variables. Standardised partial regression coefficients (or 'β' coefficient) of the significant explanatory variables would indicate their relative strengths in accounting for the variance in the criterion reading variables. In other words, the larger the β coefficient for a given skill variable in the regression model, the larger the independent contribution it makes in accounting for the reading performance. The present section focuses on passage comprehension as the criterion variable. The analyses of the whole sample precede those of the better and poorer subgroups within the sample.

Passage comprehension performance of the whole sample

The analysis in the present section first focuses on the overall Passage Reading performance of the whole sample and subsequently examines the performances on each of the four passage subsections.

Overall passage comprehension performance of the whole sample

From a multiple regression analysis with the overall Passage Reading as the criterion variable, a statistically significant model has emerged ($F_{(4,\ 214)}=138.23$, $p<.001$. Adjusted $R^2=.72$), in which three significant predictors of Syntactic Knowledge, Lexicality Effect, and Vocabulary Breadth remained in the equation as well as the non-significant Number Matching Latency which had been entered to control for any construct-irrelevant individual differences in response time (see Table 7.3).

Table 7.3 Summary of multiple regression analysis for variables predicting overall passage reading performance of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.001	0.001	.026	.472
Step 2				
Syntax	0.497	0.039	.734	.000
Lexicality	−0.003	0.001	−.091	.013
Vocabulary	0.040	0.018	.129	.024

As the standardised beta regression coefficients in Table 7.3 show clearly, Syntactic Knowledge ($\beta=.73$, $p<.001$) is the strongest predictor of the overall

Passage Reading Comprehension performance while Vocabulary Breadth (β=.13, $p<.05$) and Lexicality Effect (β=−.09, p=.01) made additional but much smaller contributions to the prediction.

Syntax and vocabulary have often been treated as synonymous to L2 proficiency and researchers have often adopted measures based on them in studies comparing the amount of explained reading variance. The results here support such notions with firmer support for the syntactic factor than lexical.

The Lexicality Effect, as it should be recalled, refers to the gap between the mean response times for those stimulus matching items which constitute existing English words (Real Word Latency) and for those which do not but conform to some common English orthographic regularities (Pseudoword Latency). Haynes (1989) explains it as the 'amount of benefit the individual derived from a stimuli's (sic) familiarity as a meaningful string, as opposed to pseudowords which had no meaning and were unfamiliar' (1989:130). Since the significant regression coefficient for the Lexicality is negative, a logical interpretation is that there is a tendency among the better passage comprehenders to be less affected in terms of processing speed by the lack of such familiarities in a given letter sequence and that the poorer passage comprehenders tended to be delayed by it.

Passage subsection performances of the whole sample

Below are the results of a series of multiple regression analyses taking each of the passage subscores as the criterion variable.

In accounting for the performance on Passage 1, a statistically significant model has emerged ($F_{(3, 215)}$=148.78, $p<.001$. Adjusted R^2=.67), in which Syntactic Knowledge and Vocabulary Breadth formed the core of the equation (see Table 7.4).

Table 7.4 Summary of multiple regression analysis for variables predicting Passage 1 performance of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.000	−.032	.419
Step 2				
Syntax	0.141	0.013	.646	.000
Vocabulary	0.021	0.006	.208	.000

Syntax (β=.65, $p<.001$) is most useful in predicting the reading performance on Passage 1, while Vocabulary Breadth is clearly shown to be a significant predictor as well, though to a much lesser degree (β=.21, $p<.001$).

The same procedure has followed for Passage 2, and again a statistically significant model has been obtained ($F_{(3,\ 215)}$=59.75, p<.001. Adjusted R^2=.45). For this passage, Lexicality followed Syntax in the strength of prediction, while Vocabulary Breadth has disappeared from the equation (Table 7.5).

Table 7.5 Summary of multiple regression analysis for variables predicting Passage 2 performance of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.000	−.018	.727
Step 2				
Syntax	0.153	0.012	.655	.000
Lexicality	−0.001	0.001	−.101	.048

Syntax best accounts for the variance on the candidates' performance on Passage 2 (β=.66, p<.001), and Lexicality is a much weaker predictor though its significance is statistically supported (β=−.10, p<.05).

The multiple regression results for Passage 3 are summarised in Table 7.6. The statistically significant equation ($F_{(3,\ 215)}$=76.87, p<.001. Adjusted R^2=.51) again includes a weighty Syntax component (β=.70, p<.001).

Table 7.6 Summary of multiple regression analysis for variables predicting Passage 3 performance of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.001	0.000	.117	.015
Step 2				
Syntax	0.130	0.009	.704	.000
Lexicality	−0.001	0.000	−.115	.017

Along with Lexicality Effect (β=−.12, p=.017), the analysis also identified Number Matching Latency as an additional significant predictor (β=−.12, p<.05). This was rather unexpected considering the lack of correlation between this variable and the Passage 3 section score (r=.03, n.s.). Nevertheless, the effects of these two variables are marginal compared to that of Syntax.

The last of the four passage subsections yielded a somewhat anticipated result. A statistically significant regression equation has been produced ($F_{(3,\ 215)}$=55.29, p<.001. Adjusted R^2=.43), and, as shown in Table 7.7, Syntax has once again appeared as the strongest predictor of Passage 4 subscore (β=.52, p<.001).

The three predictors of Syntax, Lexicality, and Vocabulary Breadth have been shown to account for the variance on the overall Passage Reading performance, and the separate subsection analyses for the passage subsections

Table 7.7 Summary of multiple regression analysis for variables predicting Passage 4 performance of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.000	.041	.432
Step 2				
Syntax	0.084	0.013	.518	.000
Vocabulary	0.013	0.006	.176	.030

consistently supported Syntax. The statistical significance for the Vocabulary Breadth in this final analysis allows the two secondary predictors of Lexicality Effect and Vocabulary Breadth to remain in the regression equations twice each, which does not contradict the overall passage results.

The data above have revealed that the significance of Syntax is persistent and the overall passage results supporting this variable are not caused by performances on certain subsections. The two other significant predictors of Vocabulary Breadth and the Effect of Lexicality are contrastively more dependent on specific subsection results and are relatively less consistent.

Passage comprehension performance of the better readers

A subgroup of 128 students who have been shown to be higher achievers are the target of analysis here. The same steps of subjecting the overall Passage Reading data and then passage subsection results to multiple regression are taken as for the whole sample analysed above.

Overall passage comprehension performance of the better readers

For the prediction of the overall Passage Reading score of the better subgroup, a significant regression equation has been identified ($F_{(3, 124)}$=39.49, p<.001. Adjusted R^2=.48). In the equation, Syntax and Real Word Latency are included as significant independent variables (Table 7.8). The prominence of Syntax is again obvious from the standardised regression coefficient (β=.68, p<.001), but the analysis of this subsample has also indicated that the efficiency in the visual recognition of high-frequency words can also contribute to the prediction of the better readers' test results.

Passage subsection performances of the better readers

It is expected from the results of the undivided sample and the overall Passage Reading score above that Syntax plays a central role in predicting the Passage

Table 7.8 Summary of multiple regression analysis for variables predicting overall passage reading performance of the better subgroup (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	−0.002	0.002	−.095	.286
Step 2				
Syntax	0.503	0.048	.681	.000
Word latency	0.004	0.002	.183	.042

subsection performance. Our interest now is how consistently it does so and what other variables, including Real Word Latency, would add to the prediction of subsection performances. Results for each subsection are presented sequentially below.

A regression model for Passage 1 has been found statistically significant ($F_{(2, 125)}$=36.26, p<.001. Adjusted R^2=.36). No independent variable other than Syntax has been identified as a significant predictor of this subsection performance (see Table 7.9).

Table 7.9 Summary of multiple regression analysis for variables predicting Passage 1 performance of the better subgroup (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.001	−.059	.411
Step 2				
Syntax	0.132	0.016	.601	.000

For Passage 2, a significant regression model has also emerged ($F_{(2, 125)}$=15.35, p<.001. Adjusted R^2=.18). As seen in Table 7.10, Syntax is again the only significant predictor of this subsection score (β=.44, p<.001).

Table 7.10 Summary of multiple regression analysis for variables predicting Passage 2 performance of the better subgroup (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.001	−.031	.697
Step 2				
Syntax	0.120	0.022	.442	.000

Table 7.11 presents the regression results for Passage 3, for which another significant regression model has emerged ($F_{(2, 125)}$= 38.73, p<.001. Adjusted

R^2=.37). The stability of Syntax is confirmed as the best predictor of Passage Reading test performance from these three subscore analyses so far.

Table 7.11 Summary of multiple regression analysis for variables predicting Passage 3 performance of the better subgroup (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.001	0.001	.128	.070
Step 2				
Syntax	0.149	0.017	.611	.000

Finally, an analysis of the data from Passage 4 has once again led to a significant regression equation ($F_{(3, 124)}$=16.57, p<.001. Adjusted R^2=.27), and Table 7.12 summarises the regression data for this subsection. For this subsection, Syntax (β=.40, p<.01) and Vocabulary Breadth (β=.21, p<.05) have been identified as significant among the independent variables.

Table 7.12 Summary of multiple regression analysis for variables predicting Passage 4 performance of the better subgroup (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.001	.059	.442
Step 2				
Syntax	0.086	0.018	.404	.000
Vocabulary	0.023	0.009	.209	.016

The separate analyses for the subsample of higher achievers have been described thus far, and the consistency of the significance of Syntax has been supported since it was found significant in each of the four subsections analysed. Vocabulary breadth has been identified as contributing additionally to the prediction of the Passage 4 subscore. Nevertheless, it was only in this subsection in which this variable was found significant. Therefore, the evidence for it is much less consistent although it deserves more attention than the other non-significant variables.

Passage comprehension performance of the poorer readers

In this section, the Passage Reading results of the remaining 91 persons who belong to the lower-achiever subgroup are analysed. The organisation of the information will be the same as that for the better subgroup presented above. Analysis of the overall Passage Reading results precedes the subsection analyses.

Overall passage comprehension performance of the poorer readers

Passage Reading data from the subgroup of poorer readers have been analysed and a significant regression model has emerged ($F_{(4, 86)}$=21.24, p<.001. Adjusted R^2=.47). The data on the variables kept in the equation are detailed in Table 7.13. Once again, Syntax is found to play a central role in the model while Lexicality has resurfaced as a secondary but significant predictor. Since Lexicality was not a part of the model for the higher ability subgroup, the results here provide tentative support for the notion that the tendency to be affected by unfamiliar word forms is stronger among the lower-achieving passage readers. Another variable that has resurfaced is Vocabulary Breadth. Its role is also secondary here, but it has to be examined for its consistency to compare its significance across the two subsamples of students. The section to follow will report on the four separate subsection results in order to examine how consistent the Syntax, Lexicality and Vocabulary variables are in accounting for the individual differences in the Passage Reading results of the relatively poorer readers.

Table 7.13 Summary of multiple regression analysis for variables predicting overall passage reading performance of the poorer subgroup (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.001	0.001	.049	.526
Step 2				
Syntax	0.417	0.066	.559	.000
Lexicality	−0.004	0.001	−.224	.005
Vocabulary	0.053	0.025	.184	.040

Passage subsection performances of the poorer readers

The analysis for Passage 1 has produced a statistically significant model ($F_{(3, 87)}$=19.33, p<.001. Adjusted R^2=.38), in which Syntax and Vocabulary Breadth served as the predictor variables (see Table 7.14). This coincides with the data obtained for the whole sample on Passage 1. For the undivided sample, the standardised regression coefficients of .65 for Syntax and .21 for Vocabulary Breadth had been obtained. The figures for the poorer readers here may indicate that the parallel between Syntax and Passage Reading is somewhat weakened and Vocabulary Breadth slightly improves its strength as a predictor of Passage Reading among a relatively weaker sample of EFL readers. That Vocabulary Breadth failed to be a part of the regression model for the subgroup of better readers is also consistent with such an interpretation.

Table 7.14 Summary of multiple regression analysis for variables predicting Passage 1 performance of the poorer subgroup (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.000	−.007	.934
Step 2				
Syntax	0.122	0.026	.451	.000
Vocabulary	0.029	0.010	.277	.005

For Passage 2, the best regression equation obtained was statistically significant ($F_{(3, 87)}$=12.11, p<.001, Adjusted R^2=.27). As shown in Table 7.15, the best predictor among the variables was Syntactic Knowledge (β=.37, p<.001), followed by Vocabulary Breadth (β=.25, p<.05). The gap between the two predictors in the beta values has become even smaller. The Passage 2 data for the better readers only found the significance of Syntax and not Vocabulary. The results here encourage us to theorise that, for some texts, while the breadth of syntactic knowledge is most closely related to their comprehension performance, the breadth of vocabulary does play an additional role, particularly for lower-achieving readers.

Table 7.15 Summary of multiple regression analysis for variables predicting Passage 2 performance of the poorer subgroup (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.001	−.025	.786
Step 2				
Syntax	0.137	0.038	.374	.000
Vocabulary	0.036	0.015	.252	.017

The data from Passage 3 have indicated for the first time that a variable other than Syntax can be the best predictor of a passage subscore. The regression coefficients from a significant model ($F_{(3, 87)}$=5.49, p<.01. Adjusted R^2=.13) are detailed in Table 7.16, and instead of Syntax, it is the Lexicality Effect that has emerged as the most significant independent variable for the prediction (β=−.30, p<.01). Syntax has diminished its role (β=.24, p<.05) and has been found the second best predictor of this subscore for the poorer readers.

Finally, the data on Passage 4 were analysed and a significant regression equation has been obtained ($F_{(2, 88)}$=11.33, p<.001. Adjusted R^2=.24). As found in Table 7.17, Syntax was the only significant predictor of this subsection performance for the lower-achieving subgroup (β=.45, p<.001).

In the set of analyses on the four subsection results from the poorer

Table 7.16 Summary of multiple regression analysis for variables predicting Passage 3 performance of the poorer subgroup (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.001	0.000	.173	.083
Step 2				
Lexicality	−0.001	0.000	−.295	.004
Syntax	0.046	0.019	.238	.018

Table 7.17 Summary of multiple regression analysis for variables predicting Passage 4 performance of the poorer subgroup (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	0.000	0.000	.027	.778
Step 2				
Syntax	0.094	0.020	.454	.000

subsample, the consistency of the Syntax variable has been verified. It was found the strongest predictor of Passage subscore in three of the four subsections and a significant predictor even when it is not the best predictor. Neither of the two other variables which predicted the overall score significantly came close to Syntax in consistency across the passage subsections. As for Lexicality, even though it was identified as a significant predictor of the overall score, the separate subscore analyses have found it significant only for the prediction of performance on Passage 3. It is thus most likely that its significance on the overall Passage Reading was chiefly due to its link with the candidates' performance on Passage 3. One more skill variable requiring a mention is Vocabulary Breadth. It was the third best predictor of the overall score after Lexicality, but it contributed to the explanation of the variances on Passage 1 and 2 as the second best predictor after Syntax. While the Lexicality Effect once exceeded Syntax in the strength of prediction, it is Vocabulary Breadth that seems relatively more stable from its multiple appearances as the second best predictor.

Skill and knowledge areas accounting for individual differences in computer-based sentence reading speed: data from multiple regression analyses

The focus now will shift from the individual differences in the performance on a paper-based passage reading comprehension test to those in the speed of reading sentences as they appear on the computer screen. The procedure of the statistical analysis will be the same as that employed for Passage Reading data, except, of course, the criterion variable will now be computer-based sentence reading speed. As with Passage Reading, the results of the whole sample are presented first, followed by those of the two subgroups of participants.

Sentence reading speed of the whole sample

The analysis of the entire sample of 219 participants has produced a statistically significant regression model ($F_{(3, 215)}$=34.64, p<.001. Adjusted R^2=.32). After the non-significant Number Matching Latency, Synonym/Antonym Decision (β=.59, p<.001) and Real Word Matching Latencies (β=.16, p=.01) have been found significant (Table 7.18).

Table 7.18 Summary of multiple regression analysis for variables predicting CB sentence reading speed of the whole sample (n=219)

Variable	B	SE (B)	β	p
Step 1				
Number latency	−0.268	0.103	−.203	.470
Step 2				
Synonym/antonym latency	0.240	0.026	.586	.000
Real word latency	0.172	0.083	.156	.010

The whole sample data thus suggest that the speed dimension of the reading ability is better accounted for by speed variables rather than power variables. The two speed variables in the model are both based on efficiency in visual alphabetic processing. Therefore, this efficiency must be importantly related to the ability to read sentences quickly on the computer screen. The beta coefficients show that, between the two significant predictors, Synonym/Antonym Decision Latency is more strongly related to the sentence reading speed than is Real Word Matching Latency. Since the Synonym/Antonym task requires speeded semantic access to one's mental lexicon in addition to the orthographic knowledge, it can be inferred that it is the efficiency in this semantic access that is more importantly related to the skill of speeded sentence reading. The sample is now divided into better and poorer readers so that separate analyses can be performed.

Sentence reading speed of the better readers

The data from the subgroup of 128 higher-achievers were submitted to a multiple regression analysis, and a statistically significant model has been obtained ($F_{(5, 122)}$= 10.23, p<.001. Adjusted R^2=.27). As shown in Table 7.19, a number of variables have been identified as collectively contributing to the regression equation. Synonym/Antonym Decision Latency has obtained the largest beta coefficient of .39 ($p=.001$), but the coefficients for the remaining predictors are relatively similar in the strength of prediction. In the order of the coefficient size in absolute values, Real Word Latency, Vocabulary, Number Latency, and RST Working Memory followed Synonym/Antonym Latency.

Table 7.19 Summary of multiple regression analysis for variables predicting CB sentence reading speed of the better readers (n=128)

Variable	B	SE (B)	β	p
Step 1				
Number latency	−0.248	0.114	−.241	.031
Step 2				
Synonym/antonym latency	0.174	0.052	.311	.001
Vocabulary	−4.103	1.275	−.260	.002
RST working memory	35.907	13.630	.207	.010
Word latency	0.23	0.089	.278	.010

The data here suggest that, for better readers, in addition to the efficiency in visual and semantic processing of the text, the breadth of vocabulary and the ability to simultaneously process and maintain information in memory also plays a significant part in accounting for the individual differences in the reading speed among the better readers.

Sentence reading speed of the poorer readers

The analysis of the lower-achieving 91 students has also yielded a significant regression equation ($F_{(3, 87)}=11.99$, $p<.001$. Adjusted $R^2=.27$). The two variables of Synonym/Antonym Latency ($\beta=.58$, $p<.001$) and Syntactic Knowledge ($\beta=.25$, $p<.01$) have been kept in the equation as significant predictors of the sentence reading speed for this subgroup (see Table 7.20).

Table 7.20 Summary of multiple regression analysis for variables predicting CB sentence reading speed of the poorer readers (n=91)

Variable	B	SE (B)	β	p
Step 1				
Number latency	−0.213	0.151	−.146	.161
Step 2				
Synonym/antonym latency	0.247	0.045	.579	.000
Syntax	19.917	7.277	.251	.008

The significance of Synonym/Antonym Latency as the best predictor indicates its invariant strength as a predictor of reading speed across different proficiency levels. The additional predictors that were significant for the better subgroup have disappeared for the lower ability subgroup and Syntax has been identified as the sole secondary predictor. It seems to suggest that, in addition to the speed dimension, the breadth dimension of language proficiency is also importantly related to the speeded performance of the lower ability readers.

Summary and discussion of the regression analyses

The previous sections in this chapter have detailed the results of the background correlational analyses and the main regression analyses. The series of data presented thus far is now considered holistically. Some of the critical information from the regression analyses is now reorganised into Tables 7.21 and 7.22 for easier comparisons of the variables. In each of the tables, the left half refers to the analysis of the whole sample and the right half to the separate analyses of the two subgroups of different reading abilities. The right half is further divided into higher and lower parts, each summarising the results of subgroup analysis. The tables list the adjusted R^2 value and the significant predictor variables for each analysis, and the standardised beta coefficients are attached in absolute values where more than one variable was found significant in the analysis. Table 7.21 first summarises the significant predictors of Passage Reading Comprehension test performance. For each sample, the predictors of the overall score are presented above those of the four separate passage subscores.

The consistency of the Syntactic Knowledge is obvious from its most frequent appearances as the best and/or the sole predictor of Passage Reading performance. It is invariantly the best predictor of the overall score, and only in one subsection analysis did it fail to be the best predictor. Where two or more predictors are found significant, the beta coefficients for Syntax are considerably higher than those of the additional predictors, confirming its consistent usefulness as an important correlate of Passage Reading. A conclusion that can be drawn from the data so far is that the individual differences in passage comprehension scores on certain texts are largely explained by the general breadth of the candidate's syntactic knowledge.

Besides the strongest Syntactic Knowledge, two other additional predictors that clearly deserve our attention are Vocabulary Breadth and the Effect of Lexicality. Clear cut comparisons between these two variables cannot be made simply from the size of the coefficients or the number of their appearances across various regression analyses.

There is one instance in which Lexicality was found a better predictor than

Table 7.21 Summary of the significant predictors of passage reading performances across separate analyses

The undivided sample (n=219)	The two subgroups (n=219)
	The better readers (n=128)
Overall Score (R^2=.72) Syntax (β=.73) Vocabulary (β=.13) Lexicality (β=.09) **Passage 1 (R^2=.67)** Syntax (β=.65) Vocabulary (β=.21) **Passage 2 (R^2=.45)** Syntax (β=.66) Lexicality (β=.10) **Passage 3 (R^2=.51)** Syntax (β=.70) Number latency (β=.12) Lexicality(β=.12) **Passage 4 (R^2=.43)** Syntax (β=.52) Vocabulary (β=.18)	**Overall score (R^2=.48)** Syntax (β=.68) Real word latency (β=.18) **Passage 1 (R^2=.36)** Syntax **Passage 2 (R^2=.18)** Syntax **Passage 3 (R^2=.37)** Syntax **Passage 4 (R^2=.27)** Syntax (β=.40) Vocabulary (β=.21)
	The poorer readers (n=91)
	Overall score (R^2=.47) Syntax (β=.56) Lexicality (β=−.22) Vocabulary (β=.18) **Passage 1 (R^2=.38)** Syntax (β=.45) Vocabulary (β=.28) **Passage 2 (R^2=.27)** Syntax (β=.37) Vocabulary (β=.25) **Passage 3 (R^2=.13)** Lexicality (β=.30) Syntax (β=.24) **Passage 4 (R^2=.24)** Syntax

was Syntax (Poorer Readers − Passage 3), and the results of this particular analysis may have influenced the regression results for the overall score of the lower ability group. It is never a significant predictor in the better subgroup, so its covariation with Passage Reading appears to be limited to the lower ability subgroup. Vocabulary has, on the contrary, appeared as the second best predictor in the analysis of the overall score of the whole sample and also of the lower ability subgroup. It was found significant in two passage subscores as the second best predictor in each of the two analyses just mentioned. Further, it is found significant in at least one of the subsection analyses for the higher ability subgroup. Vocabulary breadth thus seems slightly more consistent than the Effect of Lexicality. Furthermore, since the Syntax-Vocabulary overlap was remarkably large ($r=.77$), it is also true that much of the reading variance which Vocabulary could have predicted had already been accounted for by the Syntax measure. The same does not apply to the Lexicality Effect, which only had a correlation of −.09 (n.s.) with Syntax when the variance associated with Number Matching Latency was partialled out. In fact, a post-hoc regression analysis which only included Vocabulary Breadth and Lexicality Effect as the explanatory variables has demonstrated that Vocabulary is a much stronger predictor of Passage Reading than is the Lexicality Effect ($\beta=.69$, $p<.001$ vs. $\beta=-.10$, $p<.05$). Therefore, although the Effect of Lexicality is almost as useful as Vocabulary Breadth in the prediction of Passage Reading when it is employed along with the Syntax measure, it should be regarded as much less significant when compared with the general breadth of L2 knowledge consisting of Syntax and Vocabulary.

The general usefulness of Vocabulary Breadth as a predictor of passage reading test score seems quite substantive, and its significance as a secondary predictor in the analysis of our undivided sample leads us to an understanding that reading comprehension test variance can be partly explained by vocabulary breadth, but since it was found significant in only one subsection analysis for the higher ability subgroup and more consistently in the analyses of the lower ability subgroup, one may suspect that vocabulary breadth tends to be more strongly related to the passage reading skills of learners below a certain ability level.

One other variable, Real Word Latency, was found a second significant predictor of the overall score of the higher ability subgroup. However, it has failed to contribute uniquely to the prediction of the passage subscores and must be considered much less consistent than Syntax, Vocabulary, and the Lexicality Effect. The significance of Number Matching Latency as a secondary predictor in the subsection analysis of Passage 3 seems accidental from its very single appearance among other significant predictors with more consistent appearances.

It may be recalled that Syntax was the only predictor variable for which the COPS procedure was utilised to take account of the differences in test-taker confidence on their response accuracy. As the criterion Passage Reading measures also incorporated the COPS procedure, one might suspect that the covariation between Syntax and Passage Reading Comprehension is due to this commonality in scoring procedure. However, re-submitting the raw Syntax scores to the same set of regression analyses indicated that it remains the best and most consistent predictor of Passage Reading Comprehension scores, although its predictive power is slightly reduced while that of Vocabulary is slightly increased (see Appendix 13 for more detail).

Table 7.22 lists the significant predictors of the reading speed from the three separate multiple regression analyses. The consistency of Synonym/Antonym Decision Latency is easily observed as it is significant in all three of the analyses. As mentioned earlier, this provides support for the skill of quickly accessing semantic information from the mental lexicon as an important correlate of reading speed.

Table 7.22 Summary of the significant predictors of CB sentence reading speed across separate analyses

<table>
<tr><th>The undivided sample (n=219; R^2=32)</th><th>The two subgroups (n=219)</th></tr>
<tr><td rowspan="4">Synonym/antonym latency (β=.59)
Real word latency (β=.16)</td><td>The better readers (n=128; R^2=27)</td></tr>
<tr><td>Synonym/antonym latency (β=.31)
Real word latency (β=.28)
Vocabulary (β=.26)
Number latency (β=.24)
RST working memory (β=.20)</td></tr>
<tr><td>The poorer readers (n=91; R^2=27)</td></tr>
<tr><td>Synonym/antonym latency (β=.58)
Syntax (β=.25)</td></tr>
</table>

Real Word Latency also contributed to the explanation of reading speed variance in the whole sample and also in the higher ability subgroup. It thus suggests that the efficiency in simple visual identification of high-frequency English words is importantly related to the speed of sentence reading during comprehension.

While it is mostly speed-based variables that are found significant in accounting for the sentence reading speed, Vocabulary and Syntax are also included in the regression equations for the two subgroups. Therefore, the breadth of L2 knowledge seems to constitute a secondary though significant aspect of sentence reading speed.

8 Results and discussion 2: latent variables accounting for reading performance

Chapter overview

Following on from the main findings presented in the previous chapter on the skill and knowledge variables accounting for the individual differences in reading comprehension and speed, this chapter reports on a series of subsequent analyses for examining the data from a slightly different perspective. The data are first reanalysed to identify latent variables that seem to account for the pattern of the intercorrelations among the observed variables and to formulate a model of relationships among the variables. Reports on an evaluation of the fit between such a model and the sample data will follow.

Skill variables for further analysis

A decision is required, prior to embarking on factor analysis, on whether the analysis should include all of the variables measured for the correlation and regression analyses or focus on a selection of variables shown to be more relevant from their results.

As reported in the previous chapter, a number of skill variables had statistically significant correlations with the reading variables. However, some of the correlations found to be statistically significant were still rather low. Also, in a series of multiple regression analyses, only a small number of variables consistently made unique contributions to the explanation of the reading variances.

Since the primary interest of the present study is reading abilities, and since including those variables uncorrelated with reading might unnecessarily complicate the obtained factor structure, a conservative approach is adopted which only includes the skill variables that have exhibited meaningful degrees of overlap with them, together with Passage Reading and CB Sentence Reading Speed variables.

The four separate passage subsection scores are employed rather than the Passage Reading total score to take advantage of the multiple measures of Passage Reading performance, which would be beneficial in interpreting the results of the factor analysis and conducting a structural equation modelling (SEM) study. The regression analyses have highlighted Syntactic Knowledge as the most consistent predictor of Passage Reading Comprehension while CB Sentence Reading Speed was best accounted for by Synonym/Antonym Decision Latency. These two variables are thus included in the analysis. Vocabulary Breadth contributed to a smaller degree and less consistently to the prediction of reading variables but its partial correlations with the reading variables were clearly stronger and more stable than most of the other variables ($r=-.38$ to $.71$). Vocabulary Breadth is therefore added to the analysis. The Lexicality Effect and Real Word Latency have reached statistical significance as secondary predictor variables in some regression analyses, but their partial correlations with the reading variables were only as high as $-.18$. Although this coefficient was statistically significant ($p<.01$), the relationship itself should be considered rather weak. These two variables and the others even less significant are therefore excluded from further statistical analysis. In summary, the following variables are employed in the factor analysis to be reported in this chapter:

- Passage 1 Subscore
- Passage 2 Subscore
- Passage 3 Subscore
- Passage 4 Subscore
- CB Sentence Reading Speed
- Syntactic Knowledge
- Synonym/Antonym Decision Latency
- Vocabulary Breadth

Identifying latent variables: exploratory factor analysis

The eight variables are initially subjected to a principal component analysis (PCA) to explore how many underlying variables seem to be able to account for the variances in all of the observed variables.

Both the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (0.89) and Bartlett's Test of Sphericity (Appropriate Chi-Square=1085.05, $df=28$, $p<.001$) have yielded satisfactory values, indicating that the selection of variables is acceptable for a factor analysis.

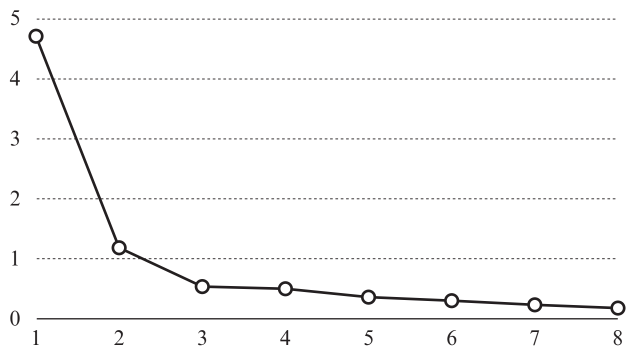
Table 8.1 lists the initial eigenvalues of the eight components and the percentage of total variance explained by each of the components. The first two exceed the conventional criterion of 1.00, providing support for a 2-factor solution. Almost 74% of the total variance is accounted for by these two components.

Table 8.1 Eigenvalues and the percentage of total variance associated with the eight components (n=219)

Component	Eigenvalue	% of Variance	Cumulative %
1	4.72	59.0	59.0
2	1.19	14.8	73.8
3	0.54	6.7	80.5
4	0.50	6.2	86.7
5	0.36	4.5	91.2
6	0.30	3.7	94.9
7	0.24	3.0	97.9
8	0.17	2.1	100

Figure 8.1 graphically presents the eigenvalues for the eight components in a Scree Plot. A plunge in the eigenvalue is observed from the first to the second component followed by what looks to be a beginning of levelling off. The 2-factor solution thus appears to be the most appropriate interpretation of the results from the analysis of the Scree Plot as well. The remaining analyses will hence be based on a 2-factor assumption.

Figure 8.1 Scree Plot: eigenvalues of the eight components



Now that the number of factors has been specified, the data are resubmitted to a factor analysis employing the Maximum Likelihood estimation method with Kaiser-normalised Promax rotation. The choice of an oblique rotation rather than an orthogonal one has been based on the assumption that factors of EFL abilities should be more likely correlated than uncorrelated. The resulting pattern matrix is presented in Table 8.2. The two factors correlated with each other at .61.

It is noticeable that all of the variables except Vocabulary Breadth load clearly on only one of the two factors. Focusing on Factor 1, one can firstly observe that all of the Passage Reading subscores are well accounted for

Table 8.2 Pattern matrix extracted after Promax rotation with Kaiser normalisation* (n=219)

	Factor 1	Factor 2
Passage 1	.845	.052
Passage 2	.798	−.072
Passage 3	.896	−.195
Passage 4	.696	.008
CB sentence reading speed	.125	−.708
Syntax	.874	.076
Vocabulary	.576	.363
Syno/antonym latency	.005	−.850

**Extraction method: Maximum Likelihood. Rotation converged in three iterations.*

by this factor. Compared to the other passage subsections, Passage 4 has a somewhat lower loading of .70, but it is not obvious from within the pattern matrix what is responsible for this discrepancy, and this point will have to be revisited in a more comprehensive analysis of the available data.

Factor 1 is not exclusively associated with the four passage subscores. Syntax shows one of the highest factor loadings of .87, while Vocabulary Breadth also displays a moderate loading of .58. Therefore, the obtained pattern matrix does not support a separate careful passage reading factor *per se*, and neither does it indicate any distinct linguistic, lexical, nor syntactic factor.

The six variables that load on Factor 1 are all PB variables representing ‘power’ as opposed to ‘speed’, which is measured through the two CB tasks that load heavily on Factor 2. More specifically, Factor 1 is further refined as mostly consisting of the ability to carefully process printed expository discourse for successfully answering questions about its contents (the passage subscores) and the ability to recognise syntactically acceptable structures through careful processing of printed sentences (Syntax). It also includes the ability to demonstrate one’s breadth of lexical meanings through PB testing (Vocabulary Breadth). In sum, central to this factor seems to be the ability to process printed text carefully using one’s breadth of linguistic knowledge. For brevity, this latent variable is hereafter referred to as Careful Text Processing Power.

As for Factor 2, two variables clearly load on it. CB Sentence Reading Speed and Synonym/Antonym Latency have fairly high factor loadings (−.71 and −.85, respectively). Though not certainly central to this factor, the factor loading of .36 for Vocabulary Breadth does not seem completely negligible in light of the only moderate loading of .57 it had achieved on Factor 1. Synonym/Antonym Decision requires expeditious access to one’s mental lexicon for retrieval of information on high-frequency word semantics, while CB Sentence Reading must involve speeded construction of sentence meaning drawing on the lexical access skill. Factor 2 can therefore be conceptualised as the speed in lexical-semantic access for rapidly comprehending context-reduced sentences presented on a computer screen. In further discussions, this factor is referred to as Lexical-Semantic Processing Efficiency.

To summarise, the results of the EFA suggest that the pattern of intercorrelations among the eight selected variables can be explained in terms of two latent factors that are mutually related. One of the factors appears to account for the individual differences in Passage Reading subscores and Syntactic Knowledge test performance whereas the other does CB tasks requiring fast lexical-semantic processing. A model to capture the suggested relationships between the two factors and the measured variables is prepared and its adequacy examined with the use of SEM methodology in the following section.

Testing the proposed ability structure: confirmatory factor analysis

About structural equation modelling

Structural equation modelling (SEM), of which confirmatory factor analysis (CFA) is a specific type (Bachman 2004), is a relatively less widely utilised statistical approach among L2 researchers compared to such procedures as multiple regression or exploratory factor analysis, thus a brief explanation on SEM is provided here.

Hoyle (1995) outlines the basic elements of SEM. According to him, it begins with the specification of a model to be estimated. A model in SEM basically refers to a statistical statement about the relations among variables and specification to the exercise of formally stating a model. This model specification involves formulating a statement about a set of parameters, which are constants indicating the nature of the relation between two variables. Parameters are either fixed (usually at zero) or freely estimated from the available data (nonzero).

It is necessary that there is a sufficient number of known parameters (i.e., known variances of the variables and covariances among them) in relation to the number of parameters to be estimated (e.g., path coefficients) based on them. When the total number of known parameters equals that of parameters to be estimated, the model is said to be 'just identified' and has zero degrees of freedom. When the number of known parameters exceeds that of free parameters, the model is said to be 'overidentified' and has degrees of freedom equal to the number of observed variances and covariances minus that of free parameters. When there is a smaller number of known parameters than of free parameters, the model is 'underidentified' and the estimation is not possible.

Therefore, SEM takes an observed covariance matrix, and the model specified by the researcher. If the model is either overidentified or just identified, the analysis can proceed. The obtained covariance matrix and the specifications are consulted to form the implied covariance matrix, which is the covariance matrix that would result if values of fixed parameters and the estimates of free parameters were placed in the structural equations. The implied matrix is refined iteratively so that the distance between the implied and the observed covariance matrices becomes minimised. There are cases in which this minimisation fails even if the model is overidentified. The degree to which the two matrices fit each other can be subjected to a test of statistical significance, and the researcher is often more interested in the result that the two matrices are so similar that the difference is statistically non-significant. There are a number of alternative indices of overall model fit as well. Along with the overall fit indices, the statistical significance of each of the estimated parameters will be reported. Depending on the software which actually performs the required computations, modification indices are available which help the researcher identify specific revisions in the equations that would improve the overall fit of the model. If the recommended modifications are theoretically justifiable, the model may be modified and resubmitted. It is common to compare two or more models of theoretical significance and test their differences statistically. Finally, the parameter estimates are available in both unstandardised and standardised values, analogous to the unstandardised and standardised partial regression coefficients in multiple regression. The former retains the scaling information of the variables, thus comparisons of their magnitudes need to be made on the basis of the latter.

Preliminary analysis

Two of the conditions in the dataset that need to be met when subjecting it to normal-theory SEM are that the data be on a continuous scale and have a multivariate normal distribution (Byrne 2001). All of the eight variables are on a continuous scale, but the data distribution needs empirical verification. The present section is based on an analysis using AMOS Version 4.0 (Arbuckle and Wothke 1999), which includes a test of multivariate normality and the function of outlier analysis. The dataset from the original sample of 219 students has yielded a value of 8.40 with its associated critical ratio of 4.92 on Mardia's (1970) normalised estimate of multivariate kurtosis. This indicates that multivariate normality is not achieved. AMOS produces a list of potential outliers in the order of the distance farthest from the centroid of observed values for the entire sample expressed in Mahalanobis d-squared values. Table 8.3 lists the 20 cases that are farthest from the centroid of the distribution in terms of the Mahalanobis d-squared. When a wide gap in this value is observed relative to all the other gaps, it can be a sign that the cases above such a gap are outliers (Byrne 2001). Between the top two cases is a very wide gap, but also between the second and the third. Since the third and fourth cases are quite close to each other but far from the fifth case, removing the first four cases seems a reasonable solution. Of the two probability columns on the right half of the table, the first is said to show the probability of the given case receiving the displayed Mahalanobis d-squared value under the hypothesis of multivariate normality. The far-right column shows a slightly different kind of probability, of the given case being exceeded by any other case in Mahalanobis d-squared value, again under the assumption of normality. Byrne quotes Arbuckle as stating that small values are to be expected in the p1 column but small figures in the p2 column indicate cases that are improbably far from the centroid under the hypothesis of normality. A sudden increase in the p2 value from the fourth to the fifth case also seems to provide support for drawing the cut-off line between the two. The entire dataset is thus resubmitted to AMOS after removing the four cases with the highest Mahalanobis d-squared values.

Table 8.3 20 cases carrying the largest Mahalanobis d-squared values

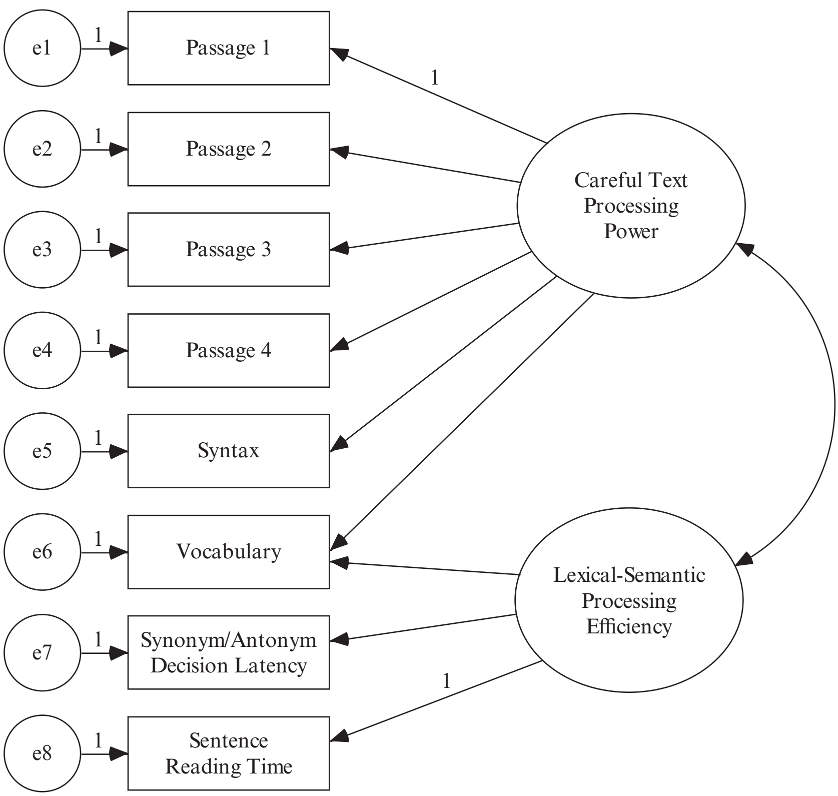
Case number	Mahalanobis d-squared	p1	p2
118	37.267	0.000	0.002
206	30.947	0.000	0.000
204	23.660	0.003	0.020
167	23.630	0.003	0.003
178	20.346	0.009	0.051
143	18.738	0.016	0.151
113	18.238	0.020	0.139
114	17.062	0.029	0.319
101	16.900	0.031	0.246
190	16.801	0.032	0.173
108	16.290	0.038	0.224
161	16.089	0.041	0.193
88	15.783	0.046	0.203
144	15.638	0.048	0.168
15	15.222	0.055	0.226
191	15.214	0.055	0.154
170	15.014	0.059	0.150
24	14.676	0.066	0.196
79	14.654	0.066	0.140
212	14.583	0.068	0.108

Mardia's coefficient for the test of multivariate normality after removing the four cases has improved to 1.77 with its associated critical ratio of 1.02. Since this is within the acceptable range, a decision has been made to proceed with the analysis on this slightly reduced dataset of 215 cases.

Initial model specification

Based on an interpretation of the EFA reported earlier in this chapter, an initial model is now specified. The 2-factor model with each of the factors accounting for different test performances is depicted in the SEM diagram in Figure 8.2. The conventional SEM notations are used in which observed variables appear as rectangles and unobserved latent variables as ovals and circles. The two large ovals are the extracted factors while the small circles labelled e1 to e8 represent error terms for their associated observed variables. Single-headed arrows represent regression paths and the double-headed ones covariances. The 'Careful Text Processing Power' factor is shown to account for the four separate passage subscores as well as the Syntax and Vocabulary test results, whereas the 'Lexical-Semantic Processing Efficiency' is shown to account for the individual differences in Vocabulary Breadth, Synonym/Antonym Decision Latency, and CB Sentence Reading Speed/Time.

Figure 8.2 Initial model of relationships among the variables



The model meets the requirement of being overidentified (df=18) as there are a total of 36 known parameters for 18 parameters to be estimated. As required for the purpose of fixing the scale of the latent variables, one each of the path coefficients is constrained to be 1.0.

On the Maximum Likelihood estimation method, the minimisation of the discrepancies between the model and the data has been achieved, yielding a significant chi-square statistic of 37.70 ($p<.01$), which indicates that the model-to-data gap is still not small enough to be statistically non-significant. However, alternative fit indices provide evidence that the specified model does in fact fit the sample data fairly well. The goodness-of-fit (GFI) value of .96, Bentler-Bonnet normed fit index (NFI) value of .97, and comparative fit index (CFI) value of .98 have been obtained, all indicating acceptable level of fit (see Table 8.4). The root mean square error of approximation (RMSEA) is not lower than the most desired .05 level but still lower than the .08 level, which is said to represent reasonable errors of approximation in the population (Byrne 2001). Nevertheless, given the significant chi-square statistic, adjustments in the model are attempted on the basis of the modification indices which AMOS provides to suggest possible modifications in the model that might improve its overall fit to the sample data.

Table 8.4 Model fit statistics on the initial model (n=215)

Chi-square	df	p	GFI	AGFI	NFI	CFI	RMSEA
37.696	18	.004	.959	.918	.967	.982	.072

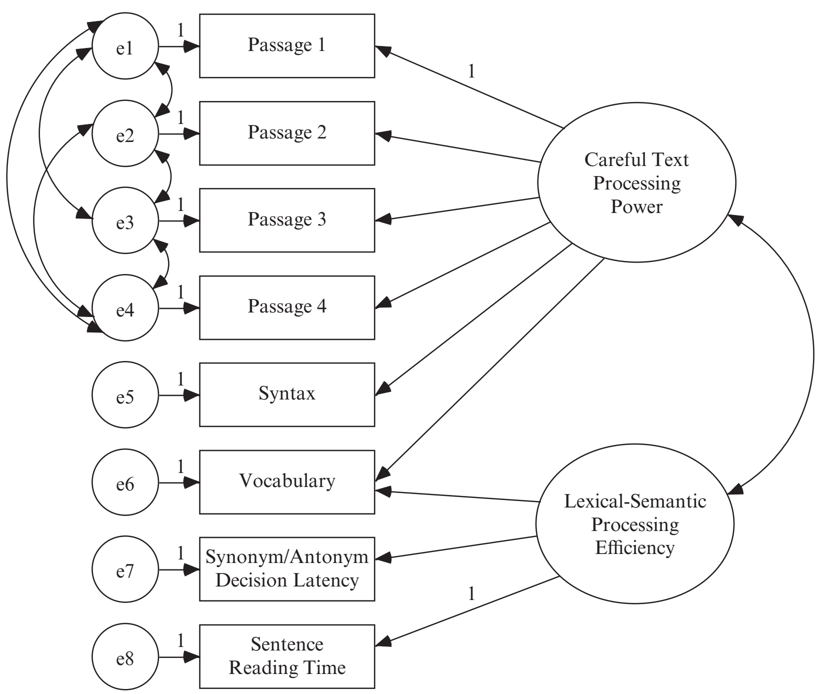
Model modification

Table 8.5 lists the modification indices for the initial model and the expected parameter changes. The modification index for a given path represents the expected decrease in the overall chi-square value should the parameter for that path be freely estimated rather than fixed at zero. The Expected Parameter Change value indicates the amount of expected change for the given path. Among the six proposed paths, the only ones that would seem theoretically reasonable are the covariances between the error terms associated with the passage subscores (i.e., between e3 and e4, between e2 and e3). Since the four passage subscores have the common and unique feature of involving PB passage reading and answering comprehension questions on the passage contents, it seems reasonable to assume that their error terms would yield covariances substantially larger than zero. While adding links only between e3 and e4 and between e2 and e3 might most efficiently lower the overall chi-square, proposing only these two of the six possible links between any two error terms for the passage subscores is theoretically less justifiable. Therefore, regardless of the lack of empirical support, the modification should also include covariances among all of the error terms for the four passage subscores. The model is thus respecified with added links among the four error terms, and this revised model is graphically presented in Figure 8.3. Although six more parameters will have to be estimated with the unchanged number of data points, the model is still overidentified with 12 degrees of freedom.

Table 8.5 Modification indices

Path	Modification index	Expected parameter change
Covariances		
e3 <---> Factor 2	11.007	15.71
e3 <---> e4	7.051	0.094
e7 <---> e3	7.102	41.588
e2 <---> e3	6.259	0.116
Regression weights		
Passage 3 <----- Factor 2	6.959	0.001
Passage 3 <----- Synonym/Antonym	7.552	0.000

Figure 8.3 Revised model of the relationships among the variables



Model comparison

Overall model fit

The revised model has been submitted to AMOS, and the model-to-data discrepancy has been minimised. The model fit statistics of the initial and the revised models are presented in Table 8.6 for comparisons.

Table 8.6 Model fit statistics on the initial and the revised models (n=215)

Model	Chi-square	df	p	GFI	AGFI	NFI	CFI	RMSEA
Initial	37.696	18	.004	.959	.918	.967	.982	.072
Revised	16.917	12	.153	.981	.944	.985	.996	.044

The new model, which allowed the covariances among the Passage Reading error terms to be freely estimated, has resulted in a non-significant chi-square statistic for the given degrees of freedom (chi-square=16.92, df=12, n.s.). It has concurrently emerged that the difference in the chi-square values between the two models is statistically significant (chi-square=20.78, df=6, p<.01). All of the fit indices shown in the table indicate that the revised model fits the sample data better than the original and even the RMSEA has improved to the level considered most desirable.

Parameter estimates

With regard to the individual path coefficients, only small changes are observed as a result of the model revision. In both models, all of the freely estimated paths are significant at p<.05 as shown in Tables 8.7 and 8.8. Table 8.9 summarises the parameter estimates for the covariance between the two factors. The inter-factor link is significant at p<.05 regardless of the model, and the differences between the models are again quite small. The initial model constrained the correlations among the passage section error terms to be zero, while their correlations were freely estimated in the revised model. The parameter estimates of the links among these error terms are shown in Table 8.10. The covariances shown in the top three lines, namely e1−e2, e2−e3, and e3−e4 pairs, are significant, but not the other three. This has been partly expected since only two of the three significant paths had been recommended via the modification index (cf. Table 8.5).

The standardised parameter estimates are also presented on the SEM diagrams in Figures 8.4 and 8.5. The initial model is described in Figure 8.4 and the revised model in Figure 8.5.

While the differences in the parameter estimates across the models are generally small, the discussion should be based on the estimates from the revised model which has been shown to fit our sample data better.

Clearly, the most notable path from the Careful Text Processing Power

Table 8.7 Summary of the path coefficient estimates for the initial model (n=215)

	Unstandardised	SE	CR	Standardised
F1 ----->				
Passage 1	0.208	0.010	20.242	.882
Passage 2	0.193	0.013	14.975	.767
Passage 3	0.157	0.010	15.886	.790
Passage 4	0.121	0.010	12.714	.698
Syntax	1.000			.927
Vocabulary	1.534	0.138	11.115	.659
F2 ----->				
Vocabulary	−0.033	0.007	−4.577	−.276
Synonym/antonym	3.953	0.619	6.387	.955
CB reading speed	1.000			.564

Table 8.8 Summary of the path coefficient estimates for the revised model (n=215)

	Unstandardised	SE	CR	Standardised
F1 ----->				
Passage 1	0.201	0.011	18.691	.867
Passage 2	0.180	0.014	13.293	.730
Passage 3	0.149	0.010	14.380	.764
Passage 4	0.115	0.010	11.813	.679
Syntax	1.000			.945
Vocabulary	1.529	0.136	11.229	.670
F2 ----->				
Vocabulary	−0.032	0.007	−4.436	−.267
Synonym/antonym	3.948	0.617	6.397	.955
CB reading speed	1.000			.565

Table 8.9 Summary of the covariance estimates between Factor 1 and Factor 2 (n=215)

Model	Unstandardised	SE	CR	Standardised
Initial	−344.2	72.694	−4.735	−.568
Revised	−355.1	74.470	−4.768	−.574

Table 8.10 Summary of the covariance estimates between the error terms (n=215)

	Unstandardised	SE	CR	Standardised
e1 <---> e2	0.124	0.059	2.095	.195
e2 <---> e3	0.173	0.059	2.951	.249
e3 <---> e4	0.117	0.042	2.767	.226
e1 <---> e3	0.023	0.044	0.511	.048
e2 <---> e4	0.028	0.054	0.520	.040
e1 <---> e4	0.018	0.041	0.427	.037

Figure 8.4 SEM parameter estimates – initial model

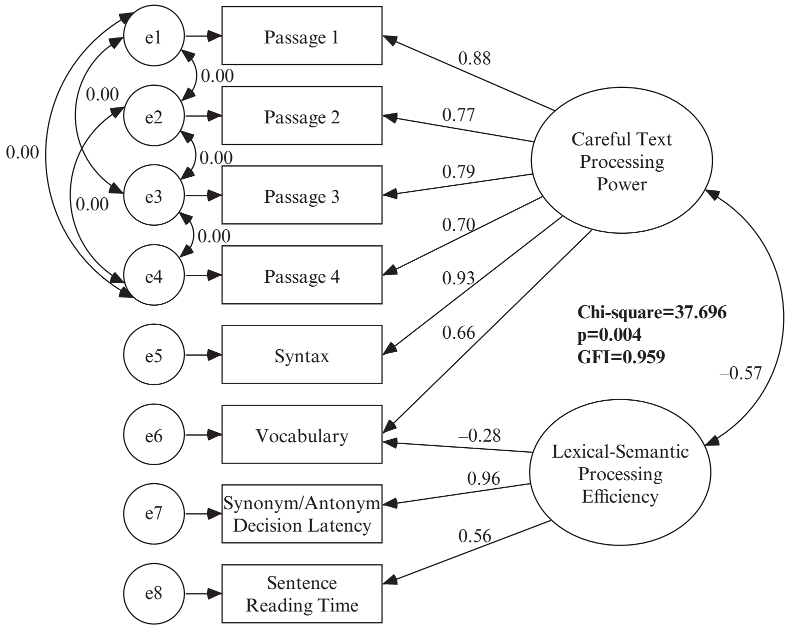
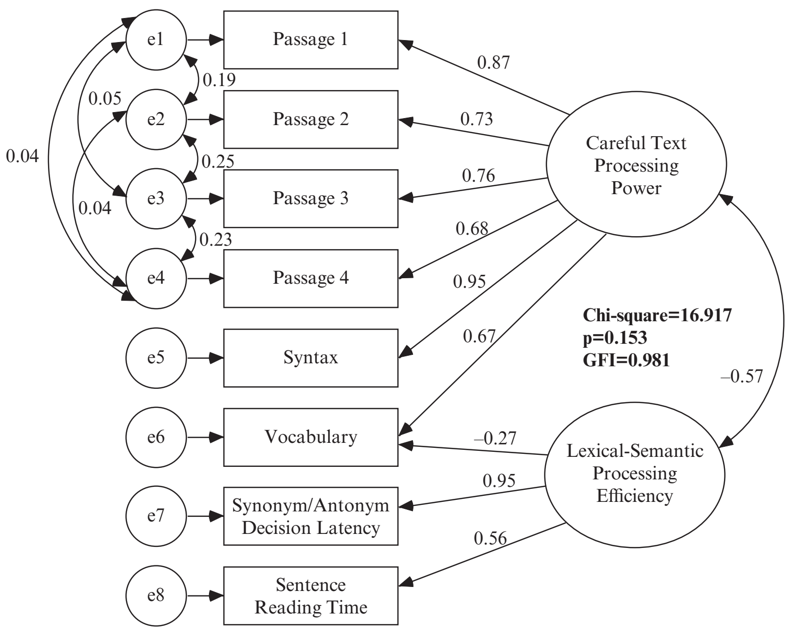


Figure 8.5 SEM parameter estimates – modified model



factor is to the Syntax test performance (.95). This seems to indicate that the ability measured through such a test is most central to this factor. All of the paths to the four passage subscores carry fairly large values as well (.68 to .87), confirming that the Passage Reading subscores are substantially accounted for by this underlying variable. Although the parameter estimate is clearly smaller than for Syntax, the link to Vocabulary Breadth is also fairly strong (.67). This helps confirm that this underlying factor is not solely based on the awareness of sentence structures and discourse markers and the like. The Vocabulary Breadth seems to remain an important dimension of careful text processing.

Of the three paths from the Lexical-Semantic Processing Efficiency factor, Synonym/Antonym Decision Latency carries an outstanding value of .95. CB Sentence Reading Speed also shows a moderate but substantial coefficient (.56) linked to this factor. Since the speed variables are in fact expressed in time, the negative value obtained for the path to Vocabulary Breadth was to be expected. Its absolute value of .27, however, is the lowest among all of the path coefficients. This can be taken to suggest that this second major factor is mostly associated with processing speed though it does have some overlap with the breadth aspect of lexical competence. Also from its strong link to the Synonym/Antonym Decision Latency, it appears that at the core of this factor is the speed of retrieving lexical-semantic information of high-frequency words in the language.

As expected, the two major underlying factors overlap to some extent (−.57). They are both abilities underlying second language text processing and are always likely to show some overlap. The extent of overlap is moderately high, but as the CFA has confirmed, two distinct factors explain our sample data and not one.

An attempt has been made to test a third model positing a higher-order factor accounting for the variances in the two major factors identified but to no avail. Minimisation of the discrepancy between the model and the data has failed. It is possible that the model has become too complicated for the sample size. Multigroup analysis has also failed, possibly from insufficiency of the sample size as well.

To take account of the overlap in the use of the COPS procedure between the criterion Passage Reading variables and the Syntax variable, the data were re-submitted after replacing the COPS Syntax score with the raw Syntax score. This did not change the relative significance of Syntax (cf. Appendix 14).

Summary and discussion of the latent ability analyses

Following on from the main correlation and regression results reported in the previous chapter, the present chapter has focused on eight selected variables to explore whether certain latent skill variables can explain the individual differences in these measured skill and knowledge variables. An EFA has extracted two such variables; one linked to the PB Passage Reading, Syntax, and Vocabulary test performances and the other linked to CB Sentence Reading Speed, Synonym/Antonym Decision Latency, and PB Vocabulary test performance. The two factors have been labelled Careful Text Processing Power and Lexical-Semantic Processing Efficiency based on the pattern of factor loadings obtained and the nature of the measures involved. The obtained pattern of factor loadings and the inter-factor correlation have led to a model of reading abilities, and it has been empirically verified in a subsequent analysis through CFA.

The EFA and CFA results have basically corroborated the regression results, with the only small difference being the clearer significance of Vocabulary Breadth in speeded semantic processing. Nevertheless, EFA and CFA have not simply confirmed what was already found in the regression analyses, but rather they have put the data in a new light to help us understand more about the structures of the abilities underlying the measured performances. The regression data had indicated that Passage Reading comprehension test performance is best accounted for by the breadth of knowledge in the target language syntax and additionally by the breadth of target language vocabulary. Another finding from the regression analysis was that the speed of reading sentences on the computer screen is best explained by the speed of semantic judgment on the contents of high-frequency synonym or antonym word pairs. However, the regression results provided little indication of, for instance, whether the competence underlying passage reading comprehension test performance should be conceptualised as separate from underlying competence responsible for performance on measures of syntactic or vocabulary breadth. The results of EFA and CFA have offered support for a view that there is no clear distinction between such competencies and that a single text processing power factor best explains the pattern of correlations among Passage Reading, Syntax, and Vocabulary. Also, from the two separate regression results, it was not possible to explore how, for instance, Synonym/Antonym Decision Latency may be indirectly related to Syntactic Knowledge. Given the moderate but significant partial correlations between them, it is not surprising that EFA and CFA have shown the two variables to be linked via latent variables.

In summary, the series of analyses reported in this chapter have supported a view that there are distinct power and speed factors underlying the performances on the tasks performed by a certain group of EFL learners and that these distinct factors are somewhat related to each other. The analyses underpin the notion that the ability to recognise syntactic structures is critically related to the ability to carefully process passages for comprehension, and they also add support for the understanding that the speed of lexical-semantic access and the speed of reading sentences on computer rely to a large extent on a common underlying competence.

9 General discussion and conclusion

This final chapter firstly answers the research questions about L2 reading ability posed in Chapter 3 with reference to the obtained data, and the findings will then be discussed in the light of related previous literature. Subsequently, the chapter will discuss the implications of the findings for theories of reading in general, for the theories of L2 development, for language pedagogy, and for language assessment. The chapter will conclude with discussion of some possible limitations of the study and recommendations for future research.

Answers to the research questions about reading performance

Answers to Research Questions 1 and 2: findings on careful passage comprehension

The first two of the four research questions concerned the ability to carefully read printed English passages for comprehension. The first question asked:

> To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills – L2 vocabulary breadth, L2 word recognition efficiency, L2 working memory span, L2 syntactic knowledge, and language-independent metacognitive knowledge about the text and reading – account for the individual differences in L2 careful passage reading comprehension ability of Japanese EFL learners?

In the regression analyses, L2 syntactic knowledge consistently emerged as the best predictor of the passage reading comprehension performance. The breadth of vocabulary and the effect of lexicality followed as additional significant predictors although the amount of their unique contributions was clearly smaller compared to the breadth of syntactic knowledge. Between these two secondary predictors, the breadth of vocabulary was slightly more consistent in the amount of unique contribution than was the effect of lexicality.

Another skill variable showed a considerable overlap with passage reading, though not to the extent its unique contribution would be found statistically significant. After syntactic knowledge and vocabulary breadth, the largest correlation with passage reading was recorded by the synonym/antonym decision latency. It, however, showed sizable overlaps with syntactic and vocabulary breadth as well, lowering the amount of its unique contribution in the explanation of the passage reading variance.

Other skill variables had much weaker relationships with passage reading comprehension. Such weak but statistically significant correlates of passage reading (at the $p<.05$ level) included the effects of orthographic regularity and lexicality and the reading span working memory. The CB sentence reading speed also showed a weak but significant correlation with passage reading comprehension.

The remaining predictors showed very small and statistically non-significant correlations with passage reading comprehension. They are real word, pseudoword and irregular string latencies and the language-independent metacognitive knowledge about the text and reading.

It is therefore mainly the breadth of linguistic knowledge which accounted for the careful passage reading comprehension, while the general efficiency in visual word recognition, working memory span, and metacognitive knowledge about reading were found to have very little relationship with the kind of reading ability assessed. One noteworthy speed variable is synonym/antonym decision latency. The only latency measure to require lexical-semantic access, it had a considerable overlap with passage reading. Together with the failure of the other latency measures to predict reading comprehension test performance, this suggests a clear dissociation between the speeds of superficial visual processing of word-level information and of word recognition accompanying lexical-semantic information retrieval, and it appears that the former has little to do with individual differences in passage comprehension but the latter does to some extent.

The second research question concerned whether there would be any differences between the better and the poorer subgroups of EFL learners in the pattern of relationships between the skill areas and the reading comprehension performance. More specifically, it had asked:

> Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in L2 careful passage reading comprehension ability of Japanese EFL learners as a function of L2 reading ability level, e.g. between 'higher ability' and 'lower ability' subgroups of readers?

A few minor differences aside, the central aspect of the results did not differ across the two subgroups: the breadth of syntactic knowledge accounts for passage comprehension the best and most consistently. While a few other skill types appeared as significant secondary predictors of the subgroup performances, the only one which appeared with any degree of consistency was the breadth of vocabulary, and its significance was more pervasive in the lower-ability subgroup. Two other predictors, real word latency and the effect of lexicality, made unique contributions in explaining passage comprehension in some cases, but the results were much less consistent compared to those on syntax and vocabulary.

It is thus only syntactic knowledge whose relative predictive power was fairly stable across the subgroup boundary, although a trend was also observed for the lower-ability subgroup for which vocabulary breadth emerged as an additional predictor along with the most significant syntax variable.

It should also be recalled at this point that a smaller-scale preliminary study (see Chapter 5) had also produced results demonstrating the strong predictive power of the general breadth of syntactic knowledge (β=.45, p<.001) and a relatively weaker yet clearly significant role of the vocabulary breadth (β=.34, p<.01). Synonym/antonym decision latency had a significant correlation with the reading score (r=−.40, p=.001), but very much like in the main study reported here, it had failed to add significantly to the explained reading variance above and beyond what the breadth of syntax and vocabulary had already accounted for. The general pattern of data from the two studies is therefore strikingly similar, which might be interpreted as a sign of reliability.

Answers to Research Questions 3 and 4: findings on computer-based sentence reading speed

The third and the fourth research questions had addressed the speed aspect of reading proficiency. The third research question had asked:

> To what extent do the linguistic knowledge areas and psycholinguistic and cognitive subskills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners?

And, the fourth research question posed was:

> Are there any differences in the extent to which the linguistic knowledge areas and psycholinguistic and cognitive skills (listed in Q1) account for the individual differences in computer-based L2 sentence reading speed of Japanese EFL learners as a function of L2 reading ability level, e.g. between ‘higher ability’ and ‘lower ability’ subgroups of readers?

As with careful passage comprehension on paper, one strong predictor emerged as clearly more consistent than the others. The synonym/antonym decision latency best accounted for the CB sentence reading speed variance of the full, undivided sample and of the separate subgroups of higher and lower ability readers as well. Other skill and knowledge variables which were significant as additional predictors were real word latency, vocabulary breadth, syntactic knowledge, number matching latency, and reading span working memory. Real word latency helped account for the reading speed variance of the full, undivided sample and the higher ability subgroup but not the lower ability subgroup. Vocabulary, number latency, and working memory added to the amount of explained variance in the reading speed of the higher ability subgroup, while syntax added to that of the lower ability subgroup.

In general, it is thus word recognition efficiency which directly accounted for the speed differences in reading sentences on computer, and the most consistent type of word recognition efficiency variable was the one which required the students access to their mental lexicon for word semantics as opposed to the type which simply required fast visual identification of letter sequences in English words or nonwords. There were some differences in the results for the two subgroups. Among the relatively advanced readers, a complex pattern of relationships emerged with more than a few skill variables accounting for the reading speed variance. Prediction of the reading speed of the less advanced group, on the contrary, could be based on only two skill areas: synonym/antonym latency and syntactic knowledge.

Reading performance and some unobserved variables

The regression analyses, which were employed to answer the research questions above, had relied on the surface-level interrelationships among the measured variables. However, as reported in the previous chapter, a significant overlap among the predictor and the criterion variables in the regression analysis may be better explained from a different perspective by hypothesising some latent variables accounting for such intercorrelations.

Although syntactic knowledge and vocabulary breadth were the most consistent predictors of passage reading performance according to the regression data, the SEM results supported a view that a common unobserved variable explains much of the variance in the test performances on passage reading comprehension, syntax, and vocabulary breadth. Nevertheless, since the link between this unobserved variable and syntax was strongest, it should be reasonable to conjecture that the ability to carefully process syntactic structures in the written input is one of the most important aspects of this unobserved variable.

As regards the sentence reading speed, synonym/antonym latency was found the best and most consistent predictor in the regression analyses. The reading speed and synonym/antonym latency were, however, shown to be explained in terms of yet another unobserved variable based on the results of SEM. It also suggested that this unobserved variable additionally explains the student performance on the vocabulary breadth test. Nonetheless, it seems to be most closely associated with the synonym/antonym latency, with its implication being that the speed of lexical-semantic access is central to this second unobserved variable. This second factor, which can be referred to as Lexical-Semantic Processing Efficiency, must merit attention for better understanding of the EFL proficiency and its development. Since two of the three indicator variables for this latent variable, namely synonym/antonym decision latency and sentence reading time/speed, were the only measures on the model that were obtained through computer-based speed tests, it would remain as a possibility that what is posited here as Lexical-Semantic Processing Efficiency is influenced by the participants' competence in the computer-based speed tests in general and not so much the semantic processing speed *per se*. To clarify this issue, an additional SEM analysis expanding the existing model was conducted, and the results indicated that the Lexical-Semantic Processing Efficiency factor is clearly distinct from general efficiency in visual processing of stimuli on computer screen and key pressing. More information on this post-hoc SEM analysis is presented in Appendix 15.

Answers to some additional questions

The findings reported so far have addressed how well the skill or knowledge variables account for reading comprehension and speed. The data from some of the skill variables can also be analysed to consider questions that are not directly concerned with reading abilities, although it is likely that they are indirectly related to reading.

One such question has been to what extent the learner's breadth of L2 vocabulary parallels their speed at visual word recognition. The correlational data have indicated that there is very little relationship between vocabulary breadth and the word recognition efficiency of high-frequency words and nonwords when the recognition tasks simply required visual matching of letter sequences. However, when the task required access to the semantic information in the participants' mental lexicon, the word recognition efficiency showed a fairly strong link with vocabulary breadth. This is not the first time such results have emerged. In a preliminary study involving a different sample and a slightly different set of word recognition test items, a very similar pattern of relationships was obtained. That is, none of the simple visual matching speed had any significant relationship with vocabulary breadth but synonym/antonym decision did. The converging evidence from the two individual studies reported in this volume may be interpreted as an indication that to EFL learners from certain educational, cultural or L1 backgrounds, the letter sequence or word shape identification subprocess is available separately from the lexical-semantic access subprocess and that those with a larger vocabulary tend to be fast at word recognition because they are fast at the latter of the two subprocesses just mentioned.

Related to such a finding is the question of whether the relationship between vocabulary breadth and lexical-semantic access speed can be attributed to a common lexical competence distinct from other related L2 competencies such as syntactic awareness, reading skill, etc. The main study and a preliminary study both provided evidence that lexical-semantic access speed is related to the breadth of both vocabulary and syntactic knowledge. The relative strength of the relationship differed between the two studies. In the main study, which analysed the performance of 219 individuals, vocabulary breadth was the best predictor of lexical-semantic access speed, and vice versa. This by no means represents conclusive evidence for a distinct lexical competence. In the preliminary study which analysed the data from 69 learners, it was the syntactic knowledge that showed the largest overlap with lexical-semantic access speed. The SEM results from the main study did offer support for a distinct lexical-semantic processing efficiency factor, which partially accounted for breadth of vocabulary. It is, however, also true that vocabulary breadth is best accounted for by the careful text processing power factor postulated in the SEM. The available empirical findings therefore fail to provide full support for a separable lexical competence.

The analysis of the word recognition efficiency variable above has led to a view that, as regards the adult Japanese EFL learners observed in the present study, the word recognition process is dissociable into at least a non-semantic grapheme or word shape identification subprocess and a lexical-semantic access subprocess. Since the word recognition tasks without the requirement of lexical-semantic access exhibited little relationship with reading abilities or vocabulary breadth, one may ask whether there are indeed individual differences in the speed of pre-semantic subprocess among these learners, all of whom have undergone at least six years of exposure to EFL texts as part of their secondary education. Data obtained for two of our studies indicate that there are relatively consistent individual differences in this subprocess. Both Preliminary Study D and the Main Study have produced evidence that the latency measures were internally consistent, and they also showed that even after the variance associated with number matching is statistically removed, the three recognition latencies for real word matching, pseudoword matching, and irregular string matching correlated significantly among each other and fairly highly in most cases ($r=.45$ to $.77$). These suggest that the phenomenon of individual differences in pre-semantic process latency is real. It is simply that such individual differences do not at all parallel those in reading comprehension or speed of reading which accompanies comprehension.

The study has also been an opportunity for investigating the nature of L2 working memory as assessed via the Reading Span Test (RST). If performance on the RST does indeed reflect one's general efficiency in the dual functions of information processing and active maintenance of information within the working memory system, it seems reasonable to hypothesise that those who are more skilled at word recognition or those with broader vocabulary or syntactic knowledge will have to expend fewer cognitive resources while processing the target sentences and thus have more resources available to focus on the maintenance of target words necessary for successful recall and vice versa. Although there was a lack of a consistent link between Reading Span and passage comprehension or reading speed, relationships between working memory and other factors should also be explored. Examination of the correlation coefficients revealed that the roles of such factors as word recognition efficiency and L2 knowledge breadth in the Reading Span performance are at best marginal. The strongest correlate of Reading Span was vocabulary breadth with a correlation coefficient of .14 ($p<.05$), which is far too small to support the hypothesis constructed above.

The findings and the literature

The results on the explanatory knowledge and skill areas employed in the present study are now discussed in turn in light of some of the previous and current research pertaining to such areas.

Syntactic knowledge

Firstly, the general breadth of syntactic knowledge has emerged as making the largest independent contribution to the explanation of the group variance in the kind of passage reading comprehension measured for the present study, and it has been the most consistent among the explanatory variables.

Both Alderson (1993) and Urquhart and Weir (1998) suggested that exploring the relationship between performance on tasks which require syntactic but little semantic processing and those which involve meaningful reading would be interesting. The present study underwent a preliminary content validation study to remove from a test of syntax those items that are actually focusing more on understanding of sentence semantics. The students' performance on this purer measure of syntax demonstrated an extremely strong link with their passage reading comprehension performance. This finding from a group of Japanese participants converges with Alderson's data from a mixture of L1 groups. Alderson found a strong correlation between grammar and reading comprehension test performances of international students in the UK and Australia. Several other findings in the present study also replicated those of Alderson. Specifically, Alderson detected stronger correlation between the grammar test score and the score of any of the several reading comprehension tests than among these reading scores themselves. The present study employed four reading passages, the scores of which recorded substantial variability thanks to the COPS procedure. The measure of syntactic knowledge correlated highly with these passage scores ($r=.65$ to $.81$, $p<.001$) and in fact slightly more than among the four reading scores ($r=.65$ to $.69$, $p<.001$). In that respect, both Alderson's and the present study support claims for the usefulness of grammar tests as an index of the individual's general L2 reading comprehension ability and at the same time point to the danger of relying on a single measure of reading test to estimate someone's reading ability. Another point of convergence between the findings from Alderson's and the present study is observed in the factorial structure of the reading and the syntax test scores. In factor analysis, no separate syntax factor was identified but a common factor best accounted for individual differences in both the syntax and the passage reading test scores. Evidence is thus accumulated from studies conducted in two very different contexts which indicate a very close link between syntactic knowledge and passage reading ability. A subsequent analysis in the present study using SEM has also suggested that four separate passage reading scores and a syntax test score loaded significantly on a common latent factor, and the syntax score had the highest loading (.95), which indicated the centrality of syntactic knowledge in that factor.

Another large-scale study by Purpura (1999) employing SEM with learners from diverse L1 backgrounds discovered a very strong link between what he calls lexico-grammatical knowledge and reading ability. Purpura's grammar measure was based on six items alone and its interpretation requires caution, but the results of his study and the present study seem to converge on the general trend that the three variables of reading, grammar/syntax, and vocabulary skills have close interrelations.

Syntax has been found an important predictor of reading ability elsewhere. A componential reading research by Haynes (1989) including a number of explanatory skill variables found her grammar measure to be the only independent variable to consistently predict her Taiwanese EFL learners' speed of passage reading significantly. While Haynes's results seem to differ from Alderson's (1993) and the present study in that it was the reading speed rather than comprehension that was accounted for by the knowledge of syntax, it is also true that the present study did observe a partial link between syntactic knowledge and the reading speed. Data from our subgroup of lower-ability readers revealed that syntax was one of two significant predictors of computer-based sentence reading speed. In that sense, the breadth of syntactic knowledge may be conceptualised as importantly related to both comprehension and speed. It is also the case that our SEM data have shown an indirect link between syntax and reading speed. Of the two latent variables identified, one seemed to be most logically associated with syntax-based careful text processing power and the other with lexical-semantic processing efficiency. These two latent variables showed a substantial overlap between each other (r=−.57).

Overall, the converging evidence from both the present and some previous studies clearly strengthens the notion that syntactic knowledge is crucially related to both reading comprehension ability and speed and weakens earlier claims from such authors as Ulijn (1981, 1984) that L2 reading requires little syntactic processing but much lexical-conceptual processing. That it requires lexical processing is obvious, but the relative significance of syntax and vocabulary is only beginning to receive serious research treatment. How the results of the present study might fit with the existing research base is the very topic of the following section.

Vocabulary and syntax

Not all of the literature supports the role of syntactic knowledge in L2 reading so much as shown in the previous section. Some studies on L2 reading ability (Brisbois 1995, Yamashita 1999) include the relative significance of syntax and vocabulary in L2 reading as part of their larger scheme of investigation, and reported that reading performance was better predicted by vocabulary than grammar knowledge.

Brisbois (1995) claimed that the reading performance of her beginning level subgroup of French as a foreign language learners in the US was better accounted for by their vocabulary knowledge than by their grammar knowledge. Her vocabulary measure had a stronger correlation with her reading comprehension measure than did her grammar measure (r=.35 vs. r=.26). Brisbois's work has been quoted in more recent publications such as Bernhardt (1999), in which it is praised as 'the most sophisticated of the studies' (1999:5) on the linguistic threshold for L1−L2 reading ability transfer.

On closer inspection, however, there are several elements in Brisbois's work that complicate its interpretation, and even mislead its readers. Firstly, there is the issue of the timing in her data collection. Brisbois reported that her learners' grammar test was given two months prior to the collection of their criterion reading comprehension data, which in itself is problematic. What complicates the matter further is that her vocabulary test was given two weeks, rather than two months, prior to the measurement of their reading comprehension ability. It is difficult to estimate the effects of such a time difference in this type of cross-sectional research, but it certainly discourages us from accepting her claims at face value.

Another source of concern is the difference in the methods of testing grammar and vocabulary knowledge. Brisbois adopted a measure of general breadth of grammar knowledge, similar to what the present study has used. Her vocabulary measure, in contrast, consisted of words appearing in the test passages for the reading comprehension measure but not yet studied in the curriculum. It was therefore not a test of general breadth of vocabulary. Her grammar test measured general breadth and her vocabulary test specific linguistic elements of the target reading passage. Hence, there was a lack of consistency. These are some of the aspects of her methodology which may well have given some advantage to the vocabulary variable as a predictor of reading comprehension performance.

Additionally, in her commentary on the relative contributions of vocabulary and grammar to the prediction of her learners' reading performance, Brisbois fails to note that her grammar measure made a significant independent contribution to the amount of total reading variance explained above and beyond what her vocabulary measure did, and she simply stated that the vocabulary score variance contributed more to the reading score variance than did the grammar score variance. Overall, the role of vocabulary appears somewhat overstated whereas that of grammar understated.

Whether misled by Brisbois's commentary or not, Bernhardt (1999) stated that the L2 reading score variance of Brisbois's learners of French was accounted for by the following explanatory variables: L1 reading (14%), L1 writing (7%), L2 vocabulary (35%), L2 grammar (3%). Nowhere in Brisbois's paper does a percentage so high as 35 appear associated with vocabulary, and it is beyond our comprehension what precisely triggered the misquote. Regardless of its causes, it has to be emphasised here that no evidence was presented to support such a dominant role of vocabulary, and Brisbois's data and her description of them should not be accepted as clear evidence of the relative significance of vocabulary over syntax.

A study of Japanese EFL readers by Yamashita (1999, 2001) is more relevant for comparison with the results of the present study, not only because of the same L1 background of the sample but because of its analysis of the relative significance of syntactic knowledge and vocabulary breadth for L2 reading. Yamashita reported that both grammar and vocabulary made independent contributions in accounting for the L2 reading variance when her entire sample of 241 learners was analysed. That is not particularly inconsistent with the results of the present study. However, there are more points of divergence than convergence. Firstly, for her undivided whole sample, her regression data indicated a stronger role of vocabulary compared to grammar. Secondly, when her sample was split into three subgroups based on reading ability, different explanatory variables emerged as significant depending on the level of the subgroup. Specifically, grammar was the only significant predictor for her lowest-level subgroup, vocabulary was so for mid-level, and between the two significant predictors for her top level group, vocabulary was a stronger predictor than was grammar. Yamashita concluded from her own and Brisbois's (1995) research critiqued above that 'L2 language ability that is necessary for L2 reading achievement is more lexical than grammatical' (2001:195). Yamashita obviously understates the fact that at least among her low level subgroup, grammar was the only explanatory variable which achieved statistical significance. It constitutes a body of evidence that for a certain subgroup of learners, grammar or syntactic knowledge is more importantly related to reading performance than is vocabulary breadth. The data from the present study further weaken Yamashita's (2001) conclusion because they supported syntactic knowledge as the strongest and most consistent predictor of reading comprehension test performance regardless of the level division.

It should also be noted that a correlation coefficient is partly a function of the reliabilities of the measures correlated, and vocabulary measures can often achieve higher reliability than grammar or syntax measures (as in the present research), making direct comparisons of correlation or regression coefficients difficult. To address this potential source of difficulty in interpretation, SEM was employed with an expanded sample of 624 students, including the original 219 for the present study and an additional 405, who only sat the paper-based measures of passage reading comprehension, syntactic knowledge and vocabulary breadth. This new analysis of the data from the expanded sample is detailed in Shiotsu and Weir (2007), but in brief the relative significance of syntactic knowledge over vocabulary breadth was supported even when the influence of differential reliabilities of the predictor variables was accounted for by the use of SEM.

While syntax clearly overshadowed vocabulary as the better predictor both in the present study and in Shiotsu and Weir (2007), we must recognise that vocabulary *is* important. Vocabulary did correlate significantly with reading but not as strongly as did syntactic knowledge, which means that syntax has taken away a large portion of the reading variance, much of which vocabulary would have accounted for if entered into the regression equation before syntax. Therefore, we should continue to develop the learners' lexical competence for improved reading comprehension performance, although our data would recommend rethinking the notion that the level of vocabulary knowledge is a much better determinant of text reading comprehension than the level of syntactic knowledge is. At any rate, the roles of lexical competence in L2 reading clearly require continued research efforts.

Word recognition efficiency

There is now a growing body of evidence from crosslinguistic studies demonstrating qualitative differences in word recognition processes among ESL readers from typologically different L1 orthographic backgrounds with otherwise comparable L2 proficiency (Akamatsu 2003, Muljani, Koda and Moates 1998, Wade-Woolley 1999, Wang and Koda 2005, Wang, Koda and Perfetti 2003). However, research on the significance of word recognition efficiency for adults' text reading performance is still limited (Akamatsu 2005, Haynes 1989, Nassaji and Geva 1999).

The present study has incorporated aspects of word recognition efficiency previously analysed in Haynes (1989) but with significantly increased sample size (from her 60 to the current 219) and improved rigour in the measurement of recognition latency through computerised testing. For Haynes's Taiwanese university students, no significant correlation was found between their reading comprehension score and the latency variables (with the exception of the orthography effect; $r=.23$, $p<.05$). Although a number of latency variables including real word and synonym/antonym latencies showed significant correlations with each of her two measures of reading speed, multiple regression analysis which controlled for the individual differences in the number matching latency indicated much reduced roles for these recognition latencies. Specifically, only synonym/antonym decision was found to make an independent contribution to the amount of explained passage reading speed variance associated with only one of two different texts on which the learners' reading speed was measured. Haynes's learners' reading speed was better accounted for by her measure of L2 linguistic knowledge, namely of grammar. The present study has identified synonym/antonym decision latency as the best predictor of reading speed of the whole sample as well as each of the two subgroups of different reading abilities.

Evidence from the present study and a part of Haynes's together support fast lexical-semantic access as an important aspect of expeditious and meaningful processing of larger and more contextualised linguistic input such as a sentence or a passage. Also, neither Haynes's nor the present study found significant roles for the graphemic and orthographic processing speed in passage reading comprehension or speed. In the present study, real word matching latency was found a significant secondary predictor in the analysis of the reading speed of the whole sample, but its predictive power was much less consistent compared to synonym/antonym decision latency. Combined with Haynes's data, thus, the present study supports a view that recognition latency of tasks requiring no semantic access tends to be unrelated to the individual's performance in more global reading, at least as far as tertiary level students in Taiwanese and Japanese contexts are concerned.

A study with L1-Japanese learners of English by Akamatsu (2005) demonstrated that his higher proficiency learners were better than their lower proficiency counterparts in the naming accuracies and latencies of case-alternated words (e.g., 'cAsE'), with which both groups performed more poorly than with normal-case words (e.g., 'case'). The visual matching tasks, as employed in Haynes's (1989) and the present study, and Akamatsu's word naming tasks resulted in quite different conclusions (cf. Brown and Haynes 1985), thus it is possible that the two tasks require different kinds of processing, but they are both worthy of further investigation. Grapheme-to-phoneme translation and phonological coding are required in word naming but not necessarily in visual matching tasks. Phonological processing involved in word recognition is seldom researched as a source of individual differences in adult L2 text reading performance and deserves more attention.

Nassaji and Geva's (1999) study addressed their L1-Farsi readers' efficiency in both phonological and orthographic processing in L2-English, but after the measures of syntactic and semantic processing skills were entered into the regression equation, only orthographic processing, and not phonological processing, emerged as a statistically significant additional predictor of text reading comprehension score or silent reading rate. Nassaji and Geva's measure of orthographic processing efficiency seems to require deeper knowledge of English orthography (e.g., deciding which of 'gnub' and 'gmub' is orthographically more acceptable in English) compared to the visual matching tasks in the present study (deciding whether the two letter sequences, e.g., 'dace' and 'dace', are identical or not). Nassaji and Geva recorded their individual participants' processing speed in terms of the time required to complete the entire set of items for each measure without isolating the response times on correct or incorrect responses at the item level, and they arrived at efficiency scores by combining the overall response accuracy and the test completion speed. Their approach would benefit from a more refined assessment methodology, perhaps with the help of computerisation of the tasks involved, and also the interpretation of their results requires caution. Nevertheless, their data are indicative of the need to investigate deeper knowledge of orthographic structure and efficiency in processing both frequent and infrequent orthographic sequences among L2 readers with different L1 orthographical backgrounds.

Although not exactly dealing with adult L2 readers, a large-scale study in the Netherlands with secondary school students by van Gelderen, Schoonen, de Glopper, Hulstijn, Simis, Snellings and Stevenson (2004) indicated that word recognition speed, as assessed through a computer-based lexical decision task, did not make a unique contribution to the prediction of reading comprehension variance after metacognitive knowledge and vocabulary knowledge had been entered.

Overall, the results of the present study are consistent with much of the existing literature on the relationship between word recognition efficiency and sentence or text reading performance among adult L2 readers. The contribution of visual word recognition speed, as assessed via stimulus matching decisions, to the prediction of reading comprehension performance is only limited, if at all, especially when compared to other stronger predictors such as the knowledge of syntax or vocabulary. However, the data from the present study with Japanese university students would recommend a closer examination of the relationship between the semantic activation component of word recognition and sentence or text reading.

Working memory

While a number of researchers have been focusing on or including L2 working memory as part of their research (Da Fontoura and Siegel 1995, Geva and Ryan 1993, Osaka and Osaka 1992, Service 1992), it is not so easy to find empirical data on the relationship between L2 text reading comprehension and working memory in published studies. Harrington and Sawyer (1992) is the only published study to have addressed the relationship between L2 reading comprehension and working memory span among adults. While their data demonstrated a significant correlation between performance on their version of the Reading Span Test (RST) and TOEFL's reading section score among their Japanese EFL learners ($r=.54$), more controlled research was sought which separates a passage reading comprehension measure from a composite vocabulary/reading score like TOEFL's. Unlike in Harrington and Sawyer's research, scores on the RST among the participants in the present study showed only a weak correlation with our index of reading comprehension ability ($r=.15$, $p<.05$). Sources of this discrepancy need to be explored.

According to Carpenter and colleagues (e.g., Just and Carpenter 1992), performance on the RST is expected to reflect one's efficient use of limited mental resources through fluent execution of input processing and stable maintenance of target information in memory during such executions. However, different versions of the RST may require differing levels of cognitive processing. Osaka and Osaka's (1992) ESL version of the RST is largely based on Daneman and Carpenter's (1980) original test with the important difference being the use of less demanding target sentences to be read and words to be recalled (see Chapter 5 to review the testing and scoring procedure). Some variants of the RST adopted in other studies have included a means to ensure that the test taker actually engages in meaningful processing of the target sentences rather than in mere key pressing or grapheme-to-phoneme translation without semantic or syntactic processing. As pointed out earlier, Harrington and Sawyer's version is one such case. Their test had a sentence verification component whereby the test taker is asked to judge whether the target sentence consisted of a syntactically and semantically acceptable word sequence. Given that their study involved only advanced ESL learners, their task had seemed too challenging for the group in the present study.

Failing to find an easier counterpart to Harrington and Sawyer's RST that would be suitable for our sample group, the present study adopted Osaka and Osaka's procedure and included no sentence verification component. The obtained mean span of 1.68 suggests that the task was difficult even without the sentence verification component. Nevertheless, the inclusion of sentence verification could be one area where the source of the discrepancy in the results might lie. It is possible that Harrington and Sawyer's version triggered deeper processing of the input whereas the test in the present study provoked much shallower processing, which may have allowed extraneous factors such as recall strategies to have an unexpectedly larger impact on the individuals' reading span performances. It should be recalled that the previous section on word recognition efficiency demonstrated a clear dissociation between the speed of grapheme level input recognition without semantic activation and that of lexical-semantic access. Only those tasks requiring the latter correlated significantly and consistently with sentence reading speed and passage reading comprehension. It gives us reason to speculate that reading the target sentences aloud without semantic activation during the Reading Span task was like the word recognition task without the semantic activation, and it may thus be the case that for the Reading Span performance to register meaningful individual differences among a sample of learners like ours, some sort of mechanism, such as the one included in Harrington and Sawyer, needs to be built into it while, of course, avoiding a floor effect. The present study has prioritised avoiding a floor effect and refrained from adding more demands to the task, but future research may need to ensure processing of sentential structure/meaning during the input processing while keeping the overall task manageable for most of the test takers. Introducing even shorter and easier sentences and target words may be one way to achieve such a balance. Following Waters and Caplan (1996), a study by Walter (2000) adopted a composite index of L2 RST performance based on three performance dimensions of the test: recall accuracy, sentence verification, and sentence reading time. The target sentences she used with her teenage EFL students in France were shorter (six to eight words long) compared to the ones used in the present study (nine to 12 words) and they seem cognitively much less demanding as well (e.g., 'I can't find my new shoes' in her set, as opposed to e.g., 'The old couple left the dinner party without eating anything' in the present study). Walter's Reading Span measure showed a strong correlation with reading comprehension ($r=.58$ to $.70$, depending on the reading task).

The present study has not found any significant role for L2 working memory as measured through an ESL version of RST. Nevertheless, it seems much too premature to conclude that working memory has little effect on L2 reading ability. Closer investigation on the measure of working memory itself may be fruitful before committing ourselves to any claims on working memory's role in L2 reading (Koda 2005).

Language-independent metacognitive knowledge of text and reading

Although the main focus of this study has been on L2 knowledge and processing skills that account for L2 reading abilities, information related to the learners' L1 reading or language-independent reading behaviour was also sought. Instead of testing the participants' L1 reading comprehension ability, the present study made use of a questionnaire on language-independent metacognitive knowledge about the text and reading in general. Previous studies with L1-Dutch and L2-English learners in the Netherlands (Schoonen et al 1998, van Gelderen et al 2003, 2004) have found their measures of metacognitive knowledge to make significant contributions to the prediction of EFL reading comprehension performance. The present study has adopted a questionnaire based on a selection of items used in van Gelderen et al's study (2003) which focus on reading behaviour. The data with our Japanese sample, however, showed no significant correlation between this measure and that of their reading abilities. There are some differences between the Dutch studies and the present study, apart from the participants' L1 and cultural background. One is the participants' age. Schoonen et al's EFL data were from students at age 14 and 16. Van Gelderen et al's sample had an average age of 14. The present study obtained data from a somewhat older group, which had an average age of approximately 20. Additionally, all of the participants in the present study were already university students. Thus, any single or combination of differences in L1, cultural, age, and educational background may have led to the discrepancies in the results. There is no strong clue at this point as to what precisely is the source of such discrepancies. Given the strong connection repeatedly found between metacognitive knowledge and text reading performance in the Dutch context, the variable seems to merit further investigation outside the Dutch context despite the lack of its significant effect found in the present study with university students in Japan.

Implications of the findings

The findings of the present study have implications for the models or theories of general reading, the theories of L2 development, L2 pedagogy, and L2 testing.

Implications for theories of reading

Except in the cases of language deficiencies, by the time we reach our adulthood, we all have mastered the intricacies of the syntactic rules of our mother tongue. The fundamentals of the syntactic rules are largely shared among the native speakers of the language, and relatively few individual differences should be observed. It is thus totally natural that componential analyses of normal adult L1 readers have seldom measured the individual's L1 syntactic knowledge as an explanatory variable. The present study has indicated that the breadth of L2 syntactic knowledge is the best predictor of L2 passage reading comprehension ability in a certain population. When this finding is taken into account, the theory of reading ability must postulate separate sets of components for L1 and L2 reading, with the L2 version alone having a heavy syntax-based component. The breadth of L1 vocabulary has been suggested as a strong predictor of L1 reading comprehension (e.g., Beck and McKeown 1991), which would mean that breadth of adult L1 vocabulary knowledge varies significantly across individuals while knowledge of syntax is largely constant. The present study has found L2 vocabulary breadth to be important for L2 reading as well. However, in the case of L2 readers, vocabulary breadth should be considered to also covary with L2 syntactic knowledge to a considerable extent, and for a certain population of L2 readers, vocabulary breadth is secondary to syntactic knowledge. The current L1 reading theories also place significance on the word recognition component of reading ability (e.g. Stanovich 1991). Fast and accurate word identification is said to characterise skilled reading. The present study has yielded data that suggest, at least as far as adult Japanese EFL readers are concerned, fast and accurate visual identification of alphabetic letter strings does not necessarily characterise skilled L2 reading, since there was little relationship between the speed of visual letter-string matching and more global reading skills when the matching task required no semantic access. This does not mean that rapid visual processing of letter strings is unimportant. Fast and accurate visual processing is clearly an integral part of lexical-semantic, syntactic, sentence-semantic and discourse level processing. It should rather imply that acquiring speed of surface level visual identification may be insufficient for skilled L2 reading, at least as far as Japanese EFL reading is concerned. This is in sharp contrast with most skilled L1 reading situations in an alphabetic language. Most visually presented words that are identified automatically and effortlessly are the ones whose meanings are accessed equally automatically and effortlessly for alphabetic L1 readers. The findings in the present study imply that, for some L2 readers, however, it is possible for a relatively fast visual identification skill to develop without concomitant development in lexical-semantic access fluency. Thus, different models of word recognition component seem to be required for L1 and L2 readers when the target language is alphabetic. The present study did not specifically address the scanning skill, which is a simple search through the given text for a target sequence of letters, words, etc. without any meaningful processing. The lack of relationship between the speed of pre-semantic word recognition and careful global comprehension also implies that scanning skill and careful passage comprehension skill are also likely to have little relationship between each other and constitute independent skill domains. Divisibility of reading ability is also supported in our analysis of latent variables, which postulated a careful text processing power factor and a lexical-semantic process efficiency factor. This finding can be related to the theoretical basis for such a distinction discussed in Khalifa and Weir (2009), an earlier volume on reading in this series. The notion proposed here is not to be equated with conceptualisations such as those of Hoover and Tunmer (1993), in which reading is viewed as consisting of two basic components of word recognition and linguistic comprehension. The present findings would encourage correction to such a view if it were to be applied in the case of adult Japanese EFL readers. Word recognition should be divided into pre-semantic visual identification and lexical-semantic access efficiency, while linguistic comprehension should be divided into careful text processing ability and the speed of comprehension. Whether it is applicable to L2 readers from other L1 and educational backgrounds or not, it can be argued based on our present findings that L2 reading ability should be conceptualised as having different componential structures compared to the case of typical L1 reading.

Implications for theories of L2 development

The division between the power and the speed factors also brings us to the discussion of L2 ability development. At least in the context of EFL education in Japan, there has been a long tradition of emphasis on careful and precise reading comprehension often mediated by detailed translations. The dimension of language ability emphasised in such a tradition is captured in the largely breadth-based power factor in the findings of the present study, and at the heart of such a factor seems to be the knowledge of English syntax. Since cross-sectional studies such as the present one only establish covariations between variables or explain individual differences in one variable in terms of another, they do not really permit identification of causal relationships (cf. Koda 2005). Having that in mind, if one is allowed to speculate on the directionality of relationships between the correlated variables, a notion that a better grasp of syntactic knowledge helps global comprehension seems more logical than a notion that global comprehension ability helps in a discrete point syntax knowledge test. Vocabulary breadth has also been identified as a significant predictor of passage comprehension, and the same logic should apply with regard to the possible direction of skill dependence. A hypothesis emerging from such speculations is that careful global EFL comprehension skill develops as a function of improved knowledge of English syntax and vocabulary. As it is pointed out above, the direction of causality is yet to be established and subject to future empirical investigation. The findings of the present study have, however, provided a foundation on which to begin such an investigation.

Another finding which is relevant to the theories of L2 development concerns the differences or similarities across the two sub-divisions of the sample population. It is clear from the data obtained that syntactic knowledge is the most significant and consistent predictor of passage reading ability among both the lower and higher ability subgroups, with implication being that the interdependence between syntax and reading is not restricted to a particular proficiency level of the learners but instead it is a phenomenon observed across a fairly wide stretch of proficiency levels. On the other hand, the vocabulary breadth tended to predict the reading performances of the lower ability group better than the higher ability group, indicating that as far as vocabulary breadth and passage reading are concerned the interdependence is more localised in terms of learner proficiency.

The obtained data also indicate that the breadth-based power factor is not all there is for Japanese EFL learners to focus on if they wish to develop their overall proficiency. Considerable individual differences are observed in the speed of accessing the meaning of high-frequency words even among the learners who are fairly accurate at identifying their meanings, and this speed of semantic processing is strongly associated with the speed of reading English sentences for comprehension. Clearly, the developmental aspect of this speed dimension of lexis- and sentence-level semantic processing needs to be brought to the fore of overall conceptualisations of L2 reading proficiency. SLA researchers have been investigating the development of L2 proficiency with an interest in so called implicit and explicit linguistic knowledge (e.g. N Ellis 2005), the former of which is sometimes operationally defined through online or timed processing of the target language (e.g. R Ellis 2005, Erlam 2006, Han and Ellis 1998). Given that fluent, effortless, and accurate processing of the target language, considered to require developed implicit knowledge, is a desirable condition for successful L2 comprehension and production, the process efficiency factor that emerged in the latent trait analysis of the present study may deserve more focused research in terms of its relationship with implicit linguistic knowledge. Also, how the efficiencies in lexis- and sentence-semantic processing improve and how they interact with the development of breadth-based careful comprehension ability or even with other global online processing skills such as aural discourse comprehension constitute some interesting questions for the future research.

Implications/recommendations for language pedagogy

While we wait for empirical evidence on causal links between the development of linguistic knowledge and that of careful passage reading comprehension skill to become available through subsequent research, the field of L2 pedagogy must continue its business of developing learners' reading abilities.

If we are to postulate a dependence of careful global reading on the breadth of linguistic knowledge, as opposed to a dependence in the opposite direction, an implication for pedagogy in the context of tertiary-level EFL training in Japan would be that focused efforts to improve syntactic and lexical skills may result in efficient development of careful passage reading comprehension ability.

Such efforts must not end in work on various syntactic and lexical facts in isolation and without context. The syntax and vocabulary measures for the present study were adopted under the explicit goal of minimising extraneous effects which can complicate the interpretation of the research data, and the test items deliberately lacked contextual clues. Thus, their format must not be mistaken to be appropriate for actual instructions or practices.

The target syntactic rules and the lexical items may be initially made salient to learners in either isolated or contextualised presentation conditions but should be eventually embedded in adequately contextualised reading activities. Ideally, the target rules and the words should be presented to learners in naturalistic texts that they repeatedly encounter in the process of integrating textual information for meaning. Programmed vocabulary development is often proposed and available for both L1 and L2 readers, and they may serve the L2 readers well if utilised effectively. Programmed courses for efficiently mastering syntactic rules may turn out to be even more useful for a group of learners like the one observed in the present study. Commercial grammar reference and practice books and materials for extensive reading may together serve them well too. If technically feasible, extensive reading materials delivered on computer which have hyperlinks embedded within the text for the readers in need of work on specific syntactic or lexical points to jump to appropriate grammar reference and practice pages, and perhaps log their online activities for them, and/or for their instructor for review sessions, may offer individualised help that is promising for their development of global reading ability.

The results of the present study also have implications concerning the development of word recognition skills. Although automatic recognition of high-frequency words must be an important condition to achieve fluent reading comprehension, work on pre-semantic graphemic recognition speed alone will probably have only limited effects.

The findings of the present study would recommend efforts to improve lexical-semantic access speed and activities which engage the learners in the mental act of quickly accessing the meaning of words or phrases that they have visually decoded. For instance, Paran (1996) recommends, along with extensive reading, use of exercises specifically targeting the development of automaticity in visual word recognition. He shows a set of example exercise items, in which each item begins with one target word followed by a list of words including it as well as several distractor words that are spelled similarly. The words are also formatted in various typefaces to make the exercise more demanding. The learners are to go through the items as quickly as possible to find the target word from among the list of distractors. According to Crawford (2005), the majority of word recognition exercises in textbooks and articles related to L2 word reading resemble this type (see also Folse 2007). However, at least as far as the participant population of the present study is concerned, this sort of exercise is unlikely to lead to development in global reading skills or comprehension speed since no semantic access is required for the completion of the exercise. What may be more effective is an elaboration of such an exercise by adding a requirement such as the identification of the target word's synonym, antonym, or the word belonging to the same semantic category (animate–inanimate, etc.) from among competitive distractors.

If feasible, exercises of this nature seem to be most suitable for delivery on computer, which enables precise measurement and recording of learner responses and response time as well as automatic activation of programmed steps. Though based on casual observation, at least many of the participants in the present study appeared to enjoy the speeded word matching tasks on computer the most among the battery of tasks they were asked to perform. The factor of learner motivation cannot be neglected, and computerised word recognition exercises may prove effective in that respect as well.

As Paran (1996) stresses, it is also true that extensive reading should be prioritised for automaticity development. In the case of Japanese EFL learners, the tendency to carefully process the details of the text for full translation into L1 at the expense of performance speed may be so strong that simply encouraging them to engage in extensive reading may not easily lead to automaticity. There have been efforts to develop Japanese readers' reading speed with the use of a CALL (computer-assisted language learning) program. Yoshida, Yoshida and Kobayashi (1992, cited in Yoshida 2001) developed a program in which learners are presented with the reading text on a computer screen with various options on presentation format and speed. The learner may, for instance, choose to make the text appear in the unit of single word, phrase, line, etc., which appear in the 'moving window' paradigm (Just, Carpenter and Woolley 1982). That is, the entire area over the text is covered except for one open 'window' through which a part of the text is shown. The text itself is fixed in the background and this small viewable area moves from the beginning of the text to the end, as if to lead the reader's eye movement. The pace of the window's movement is set beforehand by the reader to make the task adequately challenging, but the learner needs to read at a constant pace and there is no option for moving the window backwards. This prohibits regressive eye movement and forces the learner to read forward, keeping the preset pace. Yoshida et al (1992) reported that their experimental group who were trained under this condition for one academic year showed a significantly larger increase in reading comprehension than their control group. Innovative efforts to develop automaticity such as this seem to show promise as a complement to traditional extensive reading courses.

Implications/recommendations for language assessment

The findings of the present study have implications for language assessment as well. The high inter-correlations between the MCQ-based reading comprehension score and the scores on linguistic knowledge measures can be a justification for treating them as different aspects of a common construct.

It should be recalled that performance on any subsection of the passage reading test was better predicted by performance on the syntax knowledge test than by that on any other passage subsection, which largely overlaps with what was found by Alderson (1993). This higher consistency associated with the syntax measure gives it an advantage as a valid test of the ability that the reading and syntax tests are commonly measuring. As a surrogate of the reading test, however, discrete-point items focusing on grammar rules will suffer the problem of face and content validity. It is, in fact, for this reason that the Grammar paper was removed from the IELTS (International English Language Testing System) battery despite its higher consistency than any of the individual reading tests (Alderson 1993).

It should be pointed out that TOEFL, in its administrations prior to 1995, included a vocabulary subsection within its Reading paper, while it also had a separate Structure and Written Expression paper, a form of which became one of the main sources of the Syntactic Knowledge test items for the present study (as detailed in Chapters 3 and 4). The inclusion of the vocabulary subsection in the TOEFL Reading paper reflects a view of the reading construct in which vocabulary plays a significant part. Such a view, which seems to draw support from high correlations between vocabulary and reading measures, is by no means antiquated (e.g., Qian and Schedl 2004). The correlational data from the present study would favour a configuration of proficiency test battery in which both a grammar and a vocabulary component are parts of the reading paper over one in which only a vocabulary component is.

It is of note that when the Reading Comprehension component of the TOEFL Reading paper that Qian and Schedl (2004) used in their investigation was divided into five item groups, each focusing on ‘Factual Detail’ (18 items), ‘Inference’ (5), ‘Main Idea’ (4), ‘Reference’ (6), and ‘Organisation and Logic’ (5), their vocabulary measures highly correlated with the Factual Detail item group, at .70 to .72, whereas they correlated only moderately with the other item types, at .44 to .58. This is indicative of the following. High correlations between the TOEFL Reading and Vocabulary measures may be largely dependent on the variance on the Factual Detail item group, and the relationships between vocabulary and other important facets of reading comprehension skills such as inferencing, identifying main ideas, resolving referential links, and grasping textual organisation are yet to be clearly demonstrated. The Passage Reading items in the present study avoided testing factual details that can be located within single sentences and attempted to require integration of information from across multiple sentences. This reading measure correlated very strongly with both vocabulary and syntax but more highly with the latter. It appears that any practice of attaching a vocabulary component as part of a reading paper should be re-examined on the basis of more detailed research on the constructs of reading and vocabulary.

In general, shorter items seem to be more compatible with CAT (computer-adaptive testing), in which, owing to item banks and accumulated item parameter statistics, the candidate's response to a test item can be referenced by the algorithm so that, for the subsequent item, it selects one that matches the candidate's estimated ability; this process is successively applied to each item so that by the end of the test, the candidate's ability is much better calibrated than if the items had been randomly chosen from a broad spectrum of difficulties to accommodate candidates' individual differences. If performance on the syntax test is as informative as the data in the present study have suggested, efforts to construct an item bank of syntax items should be justified. Such efforts are expected to eventually yield empirical evidence which helps us identify small sets of well-discriminating items. Additionally, if performances on the syntax test and the passage reading test can be conceptualised as indices of a largely unidimensional skill, it also seems to be a viable option to combine both test types for a CAT administration. The syntax items, each of which is relatively shorter thus more economical, can precede the passage reading items so that a fairly accurate estimation of the person's ability is made based on a much less time-consuming but reliable set of syntax items, which will help determine the most effective set of passage reading items for the particular candidate. Calibrating person ability solely through full length passage reading would be time consuming and demanding on the candidate, and yet it can still result in less reliable scores compared to a short test consisting of syntax items alone. If the difficulty of the passage reading items are adequate for the candidate's ability, which is expected if the CAT system is effective, a relatively smaller amount of passage reading should be required. Combining the syntax and the passage reading items on a CAT has the potential of reliably and economically assessing the reading ability of the candidates belonging to the sample population similar to that of the present study.

Traditional reading tests have been more concerned with the assessment of comprehension ability than comprehension speed. The interrelated but independent factors of careful text processing power and lexical-semantic access efficiency have been identified through the present study, and this 2-factor solution, rather than one, suggests that only assessing careful processing dimension of the candidate's reading ability is unlikely to help us gain a balanced profile of their reading proficiency. The speed dimension should be assessed as much as the test format permits. Again, CB testing seems to offer more options. Shizuka (2000, 2004) attempted to determine, among other things, whether the speed of reading the target sentences in a discrete point sentence comprehension test would provide more information about candidate ability than when only the comprehension score is considered. His CB test data indicated that the reading speed information enabled a finer assessment of his candidate abilities than when the assessment was based solely on response accuracy. CB testing gives the test developers the option of what elements of the candidate responses are recorded. The results of Shizuka's and the present study would recommend that both the actual actions performed by the candidate and the timing of those actions be recorded. The actions may be assessed for accuracy, appropriateness, etc., while the timing data may be used to assess fluency or speed of the actions. The result would be a much more detailed profile of candidate ability, which the users of the test data can refer to for various kinds of decision making. It may be considered as a type of a diagnostic test because the data may identify, for instance, a candidate who is a competent careful processor but a slow reader, etc., and appropriate recommendations may be offered based on such a finding.

Limitations of the study

The present study focused on a number of variables identified as important in the literature (e.g. Daneman 1991, Grabe 1991, Harrington and Sawyer 1992, Schoonen et al 1998) and collected data from a reasonably large sample of learners (n=219) having a common L1 background. To the best of the author's knowledge, no previous L2 study had considered syntax, vocabulary breadth, word recognition speed, working memory, and metacognitive knowledge of the participant group of this scale simultaneously. Preparation and delivery of the group- and individually-administered instruments for this fairly large set of independent variables and the dependent reading ability variables did not leave much time and resources for in-depth, qualitative investigations into the individual members of the participant group, which would have been helpful in obtaining further insights into the nature of processes involved in their performance.

For the same reason, the present study was limited to a cross-sectional design as opposed to a longitudinal one, whose advantages would include the testing of causal relationships among the components and more detailed explorations into the developmental changes in the involvement of component skills.

Also, in the present study, there was a strong focus on conducting the main study on a solid foundation of piloted instruments. The purpose of this preliminary testing was also to manage the total size of test battery to one that was not prohibitively large for the available data collection opportunities. Since the most time-consuming of the instruments is the passage reading test, it had to rely on a format which can contain a relatively large number of items for the given length of reading text and most efficiently generate the greatest and most reliable variance among the participants. For that purpose, the study adopted a measure in the MCQ format. The need for multiple measures of reading have been addressed (Alderson 2000), and as much as the research conditions permit, additional methods of testing should also have been applied. For the present study, limited data collection opportunities for an already large test battery did not permit such additions. As argued by Rupp, Ferne and Choi (2006), MCQ may involve a unique set of cognitive operations not involved in reading under non-test conditions, and its limitation as a measure of an individual's ability in authentic reading comprehension has to be acknowledged. Nevertheless, it is also clear that the participants in the present study had to be engaged in the processing of multiple texts and answer questions which required integration of information from across sentence boundaries. Fortunately, the use of the COPS procedure led to finer discrimination of the participants on the test and its subsections, and it helped reduce the effects of guessing. Test taking strategies may have played a role, but that must be the case in most other test formats. At the least, the MCQ data obtained here should constitute an initial set of evidence which follow-up research can refer to and use for making appropriate comparisons after collecting data based on additional methods of reading ability assessment.

Related to the choice of instruments is the inability of the present research programme to obtain information on the learners' level of proficiency on external and well-referenced measures such as TOEFL, IELTS, etc., which would have permitted direct comparisons of the levels of the participants against those observed by other researchers. The study, however, needed to prioritise the conciseness of the measures and the adequacy of the item difficulties for the participant group examined.

Regarding the battery of instruments adopted, there is a fairly large difference in the amount of preliminary investigation each of them underwent. Some underwent repeated, thus, more thorough piloting, and others less. The questionnaire for measuring metacognitive knowledge was administered to the participants in the main study without any pilot administrations with a Japanese sample. Since the measure was originally developed by a group of Dutch researchers for use with students in secondary schools in the Netherlands, there was the possibility of a mismatch with the Japanese sample. The time constraints prohibited preliminary testing for the present study, but future research will need to proceed through these important preliminary steps.

Another instrument which was delivered after a relatively small amount of preliminary refinement is the CB sentence reading speed measure. It was discovered that, contrary to the preliminary data, the comprehension items turned out to be fairly difficult for many of the students in the main study. It is best to estimate the person's reading speed on as many target texts as possible which the person can actually comprehend. Subsequent research needs to refine this measure based on information on the compatibility between item difficulties and person abilities.

As for the sample size, it was adequate for most types of the analyses conducted. However, it seems that for the multi-group analysis in SEM, it was not large enough to obtain a satisfactory solution. For such an analysis, a larger sample needed to be available.

Recommendations for further research

Synthesising relevant L1 and L2 research up to the middle of the first decade in this century, Koda (2005) identified and discussed at length such essential components of reading as word recognition, vocabulary, intra-word awareness, sentence processing, discourse processing, and text comprehension. In her relatively recent volume, Koda acknowledges the significance of component skills analysis but observes that only a handful of L2 reading studies thus far have included investigations into lower-level processes. In fact, no other component skills analyses of adult L2 readers are identified than the ones discussed in Chapter 2 of the present volume (Haynes and Carr 1990, Nassaji and Geva 1999), suggesting a lack of development in this area, despite the promise that she attributes to it. Further component skills analyses of L2 reading are worthwhile, and as we consider future research development, there are several recommendations that the discussions in the previous sections would lead to.

Firstly, it must be desirable to improve some aspects of the present study and replicate it incorporating considerations of its possible limitations mentioned in the previous section. Such a replication might include multiple methods of assessing participant skills and characteristics, linking the measures to standardised instruments for enhanced research comparability, consideration of more qualitative data, and increased sample sizes. Instead of the Japanese sample, it may be applied with different L1 groups or several subgroups simultaneously for crosslinguistic comparisons.

There should be more empirical research on the syntactic knowledge and sentence parsing skills of L2 readers (cf. Juffs 2005), as well as on their processing efficiency at the word level and beyond (cf. Fender 2003).

As far as the syntactic rules are concerned, identifying the kinds of rules or syntax test items that differentiate good from poor readers should be illuminating. This may be pursued within a group sharing a common L1 initially, but later with multiple groups from differing L1 typological backgrounds, as L1−L2 distance may interact with the learner proficiency level and the difficulty of the rules or test items. Accumulating information on the difficulties of various syntactic rules and the likelihood of learners at various L2 abilities having knowledge of such rules should also be meaningful for future CAT and CALL applications.

Related to the identification of significant syntactic rules is the issue of a causal link between syntactic knowledge and reading ability. Experimental research design might be introduced and the effects of explicit instruction or implicit learning of certain syntactic rules might be evaluated in terms of learners' improvement in their reading proficiency. Similarly, the effects of lexical access training on reading speed might be explored.

In addition to the syntactic and lexical access skills, the present study has identified vocabulary breadth as a significant predictor of reading ability. Significance of aspects of vocabulary knowledge other than vocabulary breadth, such as 'vocabulary depth', has been demonstrated as a good predictor of reading test performance among certain learner groups (e.g. Qian 1999, 2002, Qian and Schedl 2004). It may thus be useful to obtain data from a sample of learners on their vocabulary breadth, vocabulary depth, and lexical access speed in order to identify their relationships among each other and with reading performance as well as to determine whether each of the three dimensions of lexical competence makes a unique contribution to the prediction of reading performance. Contributions of what Laufer, Elder, Hill, and Congdon (2004) refer to as vocabulary 'strength' to reading test performance and how it interacts with lexical access speed may help improve our understanding of both L2 reading and L2 vocabulary.

Componential studies involving language skills other than reading should improve our understanding of L2 proficiency. For instance, the kinds of explanatory variables adopted in the present study can be easily applied to investigate listening comprehension skill (e.g. Joyce 2008). The elements of efficiency in such areas as lexical access and working memory should be vital for fluent comprehension of the message transmitted through streams of sounds in listening activities; thus they seem to have a potential for being some of the significant predictors along with breadth of syntactic and vocabulary knowledge.

Conclusion

Researchers have claimed significance for various linguistic and cognitive variables as predictors of L2 reading comprehension by establishing a relationship between the predictors of their choice and learners' reading ability (e.g. Harrington and Sawyer 1992, Laufer 1992b). Such studies tended to focus on a limited number of predictors and did not compare sets of many variables that have been in a kind of 'theoretical competition'. The present study has carefully identified such predictors, selected candidate instruments, refined such instruments, progressively evaluated the predictors for further consideration, and finally subjected them to a fully multivariate analysis. The study has adopted a larger sample and improved rigour compared to earlier componential studies (e.g. Haynes 1989, Nassaji and Geva 1999), and the analysis has successfully compared the relative significance of the predictors. Use of a number of appropriate tests together with careful item analyses and a recent innovation in language assessment helped improve the reliability and validity of the findings as well. The results obtained through the main multiple regression study were cast into a new light by hypothesising a model of relationships and subjecting that model to a powerful statistical procedure, structural equation modelling, for empirical verification. The study focused on a sample of learners with a common L1 and homogeneous cultural and educational background, which is useful for data interpretation and possible crosslinguistic comparison in the future.

The study does have some limitations. Nevertheless, it is certainly possible to interpret the findings in light of such limitations, and they do offer suggestions for further research.

Language ability is multifaceted. The present study has attempted to explore predictors of passage reading comprehension ability and sentence reading speed. Each of these skills can also be conceptualised as multifaceted. In that sense, the study leaves a number of facets of L2 abilities unexplored. It has, however, shed light on some of the significant facets of L2 abilities and indicated some directions for further research which, the author believes, will eventually address more facets of the L2 abilities so far unexplored, thus contributing to a better understanding of a stimulating and significant phenomenon and the subject of our occupation, second language development and assessment.

Appendix 1 Test of passage reading comprehension: Classical item analysis (dichotomous scoring)

Dichotomous Scoring: The whole 20 item set

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R1	.69	.46	10.47	14.82	.27	.74
2	R2	.79	.41	10.38	14.76	.34	.74
3	R3	.58	.49	10.58	14.21	.42	.73
4	R4	.58	.49	10.58	14.68	.29	.74
5	R5	.58	.49	10.58	14.52	.33	.74
6	R6	.75	.43	10.41	14.54	.39	.73
7	R7	.57	.50	10.60	14.72	.27	.74
8	R8	.61	.49	10.55	14.21	.42	.73
9	R9	.70	.46	10.46	14.30	.43	.73
10	R10	.77	.42	10.40	14.47	.42	.73
11	R11	.47	.50	10.69	13.94	.49	.72
12	R12	.58	.50	10.59	15.45	.08	.76
13	R13	.37	.48	10.79	15.16	.16	.75
14	R14	.48	.50	10.68	13.89	.50	.72
15	R15	.39	.49	10.77	14.34	.39	.73
16	R16	.39	.49	10.78	14.83	.25	.74
17	R17	.42	.49	10.75	14.82	.25	.74
18	R18	.45	.50	10.72	14.88	.23	.74
19	R19	.52	.50	10.65	15.30	.12	.75
20	R20	.49	.50	10.67	14.96	.20	.75

Dichotomous Scoring: Passage 1 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R1	.69	.46	2.53	1.43	.25	.49
2	R2	.79	.41	2.43	1.52	.24	.50
3	R3	.58	.49	2.63	1.21	.44	.37
4	R4	.58	.49	2.63	1.32	.32	.45
5	R5	.58	.49	2.63	1.43	.22	.52

Dichotomous Scoring: Passage 2 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R6	.75	.43	2.65	1.59	.42	.60
2	R7	.57	.50	2.84	1.60	.31	.65
3	R8	.61	.49	2.79	1.42	.49	.56
4	R9	.70	.46	2.70	1.57	.40	.61
5	R10	.77	.42	2.63	1.58	.44	.59

Dichotomous Scoring: Passage 3 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R11	.47	.50	1.82	1.34	.34	.37
2	R12	.58	.50	1.71	1.62	.10	.53
3	R13	.37	.48	1.92	1.53	.18	.48
4	R14	.48	.50	1.81	1.28	.41	.32
5	R15	.39	.49	1.90	1.41	.29	.40

Dichotomous Scoring: Passage 4 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R16	.39	.49	1.87	1.30	.19	.31
2	R17	.42	.49	1.84	1.22	.25	.26
3	R18	.45	.50	1.81	1.35	.13	.36
4	R19	.52	.50	1.74	1.37	.11	.37
5	R20	.49	.50	1.77	1.23	.23	.27

Appendix 2 Rasch analyses of the test of passage reading comprehension (data from preliminary study D and the main study combined): Persons-to-items map, model fit, and item difficulty estimates

```
-------------------------------------------------------------------------------
   Item Estimates (Thresholds)
   all on all (N=401 L=20 Probability Level=0.50)
-------------------------------------------------------------------------------
4.0                                     |
                                        |
                                    XXX |
3.0                                     |
                                        |
                                   XXXX |
2.0                        XXXXXXXXXXXX |
                                        |
                       XXXXXXXXXXXXXXX  |
                           XXXXXXXXXXX  |
                                        | 16
1.0                        XXXXXXXXXXXX |
                                        | 13      17
                  XXXXXXXXXXXXXXXXXXX   | 14      15
                           XXXXXXXXXXX  | 18      20
                                        | 11
             XXXXXXXXXXXXXXXXXXXXXX     | 19
0.0                       XXXXXXXXXXXXX |  3       5       12
                                        |  7
                        XXXXXXXXXXXXXX  |  4
                                        |  8
                XXXXXXXXXXXXXXXXXXX     |
                     XXXXXXXXXXXXXXX    |  1       9
-1.0                        XXXXXXXXXX  |
                                        |  6
                        XXXXXXXXXXXXX   | 10
                                        |  2
                               XXXXXXXX |
-2.0                                 XX |
                                        |
                                    XXX |
-3.0                                    |
                                      X |
                                        |
-4.0                                    |
-------------------------------------------------------------------------------
   Each X represents 2 students
===============================================================================
```

```
-----------------------------------------------------------------------
Item Fit
all on all (N=401 L=20 Probability Level=0.50)
-----------------------------------------------------------------------
INFIT
  MNSQ          0.63    0.71    0.83    1.00    1.20    1.40    1.60
------------------+-------+-------+-------+-------+-------+-------+----
  1  item  1                  :           |*          :
  2  item  2                  :           *           :
  3  item  3                  :   *       |           :
  4  item  4                  :           |  *        :
  5  item  5                  :          *|           :
  6  item  6                  :         * |           :
  7  item  7                  :           |      *    :
  8  item  8                  :          *|           :
  9  item  9                  :        *  |           :
 10  item 10                  :       *   |           :
 11  item 11                  :       *   |           :
 12  item 12                  :           |      *    :
 13  item 13                  :           |   *       :
 14  item 14                  :   *       |           :
 15  item 15                  :          *|           :
 16  item 16                  :           |*          :
 17  item 17                  :           |*          :
 18  item 18                  :           |  *        :
 19  item 19                  :           |        *  :
 20  item 20                  :           |  *        :
=======================================================================
```

Item Estimates (Thresholds) in input order
all on all (N=401 L=20 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	268	401	-0.65 .12	1.03	1.00	0.5	0.1
2 item 2	319	401	-1.41 .13	1.01	0.90	0.1	-0.6
3 item 3	207	401	0.10 .11	0.84	0.78	-3.9	-2.8
4 item 4	233	401	-0.21 .11	1.06	1.05	1.4	0.6
5 item 5	211	401	0.05 .11	0.94	0.94	-1.4	-0.7
6 item 6	301	401	-1.12 .13	0.93	0.86	-1.2	-1.1
7 item 7	227	401	-0.14 .11	1.16	1.20	3.3	2.2
8 item 8	245	401	-0.36 .11	0.98	0.99	-0.4	-0.1
9 item 9	267	401	-0.64 .12	0.90	0.79	-2.1	-2.2
10 item 10	312	401	-1.29 .13	0.88	0.73	-1.9	-2.0
11 item 11	188	401	0.33 .11	0.88	0.81	-2.8	-2.4
12 item 12	208	401	0.09 .11	1.19	1.27	4.1	3.0
13 item 13	148	401	0.82 .12	1.12	1.11	2.3	1.2
14 item 14	163	401	0.63 .11	0.84	0.78	-3.6	-2.6
15 item 15	156	401	0.72 .12	0.98	0.95	-0.5	-0.5
16 item 16	123	401	1.15 .12	1.01	0.96	0.2	-0.3
17 item 17	154	401	0.75 .12	1.03	1.04	0.5	0.4
18 item 18	172	401	0.52 .11	1.06	1.13	1.3	1.5
19 item 19	199	401	0.20 .11	1.16	1.21	3.6	2.4
20 item 20	177	401	0.46 .11	1.05	1.04	1.2	0.5
Mean			0.00	1.00	0.98	0.0	-0.2
SD			0.73	0.11	0.16	2.3	1.7

Appendix 3 Test of passage reading comprehension: Classical item analysis (COPS)

COPS: The whole 20 item set

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	RCP1	.47	.37	6.72	14.94	.63	.90
2	RCP2	.59	.36	6.60	15.02	.61	.90
3	RCP3	.41	.38	6.78	14.89	.63	.90
4	RCP4	.36	.36	6.83	15.21	.55	.90
5	RCP5	.37	.35	6.82	15.28	.54	.90
6	RCP6	.56	.37	6.63	14.92	.64	.90
7	RCP7	.36	.35	6.83	15.21	.55	.90
8	RCP8	.41	.37	6.78	15.11	.57	.90
9	RCP9	.50	.37	6.69	14.73	.70	.90
10	RCP10	.55	.36	6.64	14.93	.66	.90
11	RCP11	.31	.36	6.88	14.97	.64	.90
12	RCP12	.32	.33	6.86	15.76	.38	.91
13	RCP13	.19	.27	7.00	15.99	.37	.91
14	RCP14	.30	.35	6.89	14.87	.70	.90
15	RCP15	.24	.33	6.95	15.41	.53	.90
16	RCP16	.23	.31	6.96	15.65	.45	.90
17	RCP17	.24	.31	6.95	15.64	.46	.90
18	RCP18	.26	.32	6.93	15.74	.40	.91
19	RCP19	.29	.32	6.90	15.94	.33	.91
20	RCP20	.25	.29	6.94	15.70	.46	.90

COPS: Passage 1 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	RCP1	.47	.37	1.72	1.14	.62	.72
2	RCP2	.59	.36	1.61	1.21	.53	.75
3	RCP3	.41	.38	1.79	1.11	.64	.72
4	RCP4	.36	.36	1.83	1.21	.54	.75
5	RCP5	.37	.35	1.83	1.27	.47	.77

COPS: Passage 2 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	RCP6	.56	.37	1.82	1.33	.65	.81
2	RCP7	.36	.35	2.01	1.44	.54	.84
3	RCP8	.41	.37	1.96	1.35	.63	.81
4	RCP9	.50	.37	1.87	1.28	.70	.79
5	RCP10	.55	.36	1.82	1.31	.71	.79

COPS: Passage 3 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	RCP11	.31	.36	1.05	.78	.56	.64
2	RCP12	.32	.33	1.03	.94	.34	.72
3	RCP13	.19	.27	1.17	1.00	.35	.71
4	RCP14	.30	.35	1.06	.75	.65	.60
5	RCP15	.24	.33	1.12	.85	.50	.66

COPS: Passage 4 subset

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	RCP16	.23	.31	1.04	.65	.44	.54
2	RCP17	.24	.31	1.02	.65	.45	.54
3	RCP18	.26	.32	1.00	.71	.29	.62
4	RCP19	.29	.32	.97	.72	.28	.62
5	RCP20	.25	.29	1.02	.66	.45	.54

Appendix 4 Test of computer-based sentence reading: Classical item analysis

Computer-Based Sentence Reading (Comprehension)

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_TB01	.50	.50	4.92	3.70	.46	.48
2	A_TB02	.57	.50	4.84	3.90	.35	.52
3	A_TB03	.45	.50	4.97	3.95	.32	.52
4	A_TB04	.38	.49	5.04	4.38	.11	.58
5	A_TB05	.45	.50	4.96	3.85	.38	.51
6	A_TB06	.67	.47	4.75	3.90	.39	.51
7	A_TB07	.79	.41	4.63	4.57	.06	.59
8	A_TB08	.57	.50	4.84	4.44	.07	.59
9	A_TB09	.60	.49	4.82	4.20	.20	.56
10	A_TB10	.44	.50	4.97	4.18	.20	.56

Computer-Based Sentence Reading (Latency: msecs per syllable)

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	CB01	519.02	265.20	4829.72	3287035.94	.69	.91
2	CB02	650.65	367.62	4698.10	3098484.53	.61	.92
3	CB03	575.93	287.77	4772.81	3216432.11	.70	.91
4	CB04	471.60	233.79	4877.14	3330435.43	.75	.91
5	CB05	579.62	288.07	4769.13	3171593.95	.75	.91
6	CB06	505.89	203.08	4842.86	3475739.40	.67	.92
7	CB07	533.48	220.67	4815.27	3356685.26	.77	.91
8	CB08	521.79	239.48	4826.96	3323407.46	.74	.91
9	CB09	523.03	262.15	4825.71	3245131.77	.75	.91
10	CB10	467.74	211.11	4881.01	3381289.62	.77	.91

Appendix 5 Test of syntactic knowledge: Classical item analysis (dichotomous scoring)

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	G1	.09	.29	16.86	30.39	.25	.81
2	G2	.77	.42	16.18	29.21	.41	.80
3	G3	.37	.48	16.58	29.52	.29	.81
4	G4	.80	.40	16.15	30.07	.24	.81
5	G5	.68	.47	16.26	28.87	.43	.80
6	G6	.60	.49	16.35	29.83	.22	.81
7	G7	.57	.50	16.38	28.61	.45	.80
8	G8	.72	.45	16.23	28.81	.47	.80
9	G9	.37	.48	16.58	30.01	.19	.81
10	G10	.82	.38	16.13	29.91	.29	.81
11	G11	.78	.42	16.17	29.86	.27	.81
12	G13	.79	.41	16.16	29.68	.31	.81
13	G14	.25	.43	16.70	29.46	.34	.81
14	G15	.59	.49	16.36	28.92	.40	.80
15	G16	.75	.43	16.20	28.77	.49	.80
16	G17	.75	.43	16.20	30.26	.17	.81
17	G19	.68	.47	16.27	28.78	.45	.80
18	G20	.59	.49	16.36	29.34	.31	.81
19	G22	.42	.49	16.53	28.96	.39	.80
20	G23	.48	.50	16.47	28.58	.45	.80
21	G24	.31	.46	16.64	30.09	.19	.81
22	G25	.47	.50	16.47	28.65	.44	.80
23	G26	.24	.43	16.71	30.24	.18	.81
24	G27	.47	.50	16.48	29.08	.36	.81
25	G28	.14	.35	16.81	31.55	-.11	.82
26	G29	.16	.36	16.79	30.89	.06	.82
27	G30	.50	.50	16.45	28.76	.42	.80
28	G31	.55	.50	16.40	29.19	.34	.81
29	G32	.63	.48	16.32	28.57	.48	.80
30	G33	.68	.47	16.27	29.99	.21	.81
31	G34	.47	.50	16.47	29.49	.28	.81
32	G35	.45	.50	16.50	30.09	.17	.81

Appendix 6 Rasch analyses of the test of syntactic knowledge (data from preliminary study D and the main study combined): Persons-to-items map, model fit, and item difficulty estimates

```
-----------------------------------------------------------------------
Item Estimates (Thresholds)
all on all (N=401 L=32 Probability Level=0.50)
-----------------------------------------------------------------------
 4.0                                       |
                                           |
                                           |
                                          X|
 3.0                                       |
                                         XX|
                                           |   1
                                        XXX|
                                           |
                                        XXX|  25        26
 2.0                                       |
                                      XXXXX|
                                       XXXX|
                                           |
                                   XXXXXXXX|  23
                                      XXXXX|  13
                                           |  21
 1.0                             XXXXXXXXXX|
                                XXXXXXXXXXX|
                                           |   3         9
                                XXXXXXXXXXX|  19        32
                             XXXXXXXXXXXXXX|
                                 XXXXXXXXXX|  20        22          31
                             XXXXXXXXXXXXXX|  24        27
 0.0                                       |  28
                               XXXXXXXXXXXX|
                                XXXXXXXXXXX|   7        14
                                XXXXXXXXXXX|   6        18          29
                                 XXXXXXXXXX|
                                           |  30
                                 XXXXXXXXXX|   5
                          XXXXXXXXXXXXXXXXX|
-1.0                             XXXXXXXXXX|   8        15          17
                                           |  16
                              XXXXXXXXXXXXX|
                                      XXXXX|  11        12
                                           |  2 4
                                         XX|  10
                                          X|
-2.0                                   XXXX|
                                          X|
                                           |
                                          X|
                                           |
-3.0                                       |
-----------------------------------------------------------------------
Each X represents 2 students
=======================================================================
```

```
----------------------------------------------------------------------
Item Fit
all on all (N=401 L=32 Probability Level=0.50)
----------------------------------------------------------------------
INFIT
  MNSQ          0.63    0.71    0.83    1.00    1.20    1.40    1.60
------------------+-------+-------+-------+-------+-------+-------+---
   1  item  1                 :          *|           :
   2  item  2                 :      *    |           :
   3  item  3                 :           |*          :
   4  item  4                 :           |*          :
   5  item  5                 :         * |           :
   6  item  6                 :           |  *        :
   7  item  7                 :         * |           :
   8  item  8                 :        *  |           :
   9  item  9                 :           |  *        :
  10  item 10                 :          *|           :
  11  item 11                 :          *|           :
  12  item 12                 :          *|           :
  13  item 13                 :          *|           :
  14  item 14                 :         * |           :
  15  item 15                 : *         |           :
  16  item 16                 :           |   *       :
  17  item 17                 :          *|           :
  18  item 18                 :           |*          :
  19  item 19                 :         * |           :
  20  item 20                 :  *        |           :
  21  item 21                 :           |     *     :
  22  item 22                 :     *     |           :
  23  item 23                 :           |    *      :
  24  item 24                 :           |*          :
  25  item 25                 :           |      *    :
  26  item 26                 :           |    *      :
  27  item 27                 :     *     |           :
  28  item 28                 :           *           :
  29  item 29                 :  *        |           :
  30  item 30                 :           | *         :
  31  item 31                 :           *           :
  32  item 32                 :           |      *    :
======================================================================
```

Item Estimates (Thresholds) in input order
all on all (N=401 L=32 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	40	401	2.54 .18	0.98	1.34	-0.1	1.4
2 item 2	316	401	-1.50 .13	0.93	0.82	-1.0	-1.3
3 item 3	145	401	0.68 .11	1.05	1.06	1.0	0.7
4 item 4	313	401	-1.45 .13	1.03	1.10	0.5	0.7
5 item 5	265	401	-0.77 .12	0.94	0.93	-1.4	-0.7
6 item 6	224	401	-0.27 .11	1.09	1.13	2.1	1.6
7 item 7	221	401	-0.23 .11	0.93	0.88	-1.7	-1.5
8 item 8	276	401	-0.91 .12	0.92	0.82	-1.7	-1.8
9 item 9	143	401	0.70 .12	1.08	1.11	1.7	1.3
10 item 10	323	401	-1.61 .13	0.96	0.90	-0.5	-0.7
11 item 11	305	401	-1.32 .13	0.96	0.93	-0.5	-0.5
12 item 12	310	401	-1.40 .13	0.98	0.86	-0.4	-1.1
13 item 13	110	401	1.15 .12	0.97	0.94	-0.5	-0.6
14 item 14	220	401	-0.22 .11	0.93	0.90	-1.8	-1.2
15 item 15	280	401	-0.96 .12	0.81	0.68	-4.0	-3.4
16 item 16	291	401	-1.12 .12	1.13	1.27	2.3	2.2
17 item 17	280	401	-0.96 .12	0.97	0.96	-0.6	-0.3
18 item 18	229	401	-0.33 .11	1.03	1.02	0.8	0.3
19 item 19	159	401	0.50 .11	0.92	0.90	-1.8	-1.2
20 item 20	180	401	0.25 .11	0.84	0.79	-4.0	-2.9
21 item 21	116	401	1.07 .12	1.19	1.34	3.0	3.0
22 item 22	178	401	0.27 .11	0.88	0.86	-2.9	-1.9
23 item 23	96	401	1.36 .13	1.14	1.32	2.0	2.5
24 item 24	189	401	0.14 .11	1.04	1.04	1.1	0.5
25 item 25	55	401	2.14 .15	1.19	1.70	1.8	3.3
26 item 26	56	401	2.12 .15	1.13	1.49	1.3	2.4
27 item 27	184	401	0.20 .11	0.92	0.91	-1.9	-1.3

Item Estimates (Thresholds) in input order
all on all (N=401 L=32 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
28 item 28	198	401	0.04 .11	0.99	0.97	-0.2	-0.3
29 item 29	234	401	-0.39 .11	0.84	0.80	-4.2	-2.6
30 item 30	255	401	-0.64 .11	1.06	1.07	1.4	0.8
31 item 31	173	401	0.33 .11	1.01	1.00	0.2	0.1
32 item 32	153	401	0.58 .11	1.17	1.22	3.5	2.4
Mean			0.00	1.00	1.03	-0.2	0.0
SD			1.10	0.10	0.22	2.0	1.8

Appendix 7 Test of syntactic knowledge: Classical item analysis (COPS)

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	GCP1	.07	.22	10.99	35.86	.36	.93
2	GCP2	.61	.37	10.45	34.34	.55	.93
3	GCP3	.24	.35	10.82	34.66	.51	.93
4	GCP4	.53	.32	10.53	34.54	.58	.93
5	GCP5	.49	.37	10.57	33.88	.65	.93
6	GCP6	.38	.36	10.68	34.76	.47	.93
7	GCP7	.37	.36	10.69	33.99	.66	.93
8	GCP8	.50	.36	10.57	33.75	.72	.93
9	GCP9	.23	.33	10.83	35.27	.37	.93
10	GCP10	.59	.34	10.47	34.43	.59	.93
11	GCP11	.50	.34	10.56	34.41	.59	.93
12	GCP13	.48	.31	10.58	34.28	.67	.93
13	GCP14	.16	.30	10.90	35.10	.46	.93
14	GCP15	.42	.39	10.64	33.91	.61	.93
15	GCP16	.55	.37	10.52	33.65	.72	.93
16	GCP17	.48	.33	10.59	34.61	.55	.93
17	GCP19	.47	.36	10.60	33.96	.66	.93
18	GCP20	.35	.33	10.71	34.59	.56	.93
19	GCP22	.27	.35	10.79	34.50	.54	.93
20	GCP23	.30	.34	10.76	34.18	.64	.93
21	GCP24	.18	.28	10.89	35.51	.37	.93
22	GCP25	.32	.37	10.74	34.04	.63	.93
23	GCP26	.15	.30	10.91	35.58	.34	.93
24	GCP27	.30	.34	10.77	34.53	.55	.93
25	GCP28	.06	.18	11.00	36.71	.05	.94
26	GCP29	.10	.23	10.97	36.29	.18	.94
27	GCP30	.32	.35	10.75	34.29	.60	.93
28	GCP31	.33	.34	10.73	34.44	.58	.93
29	GCP32	.42	.36	10.64	33.87	.68	.93
30	GCP33	.40	.33	10.66	34.83	.50	.93
31	GCP34	.27	.32	10.79	34.88	.49	.93
32	GCP35	.23	.28	10.83	35.53	.37	.93

Appendix 8 Test of vocabulary breadth: Classical item analysis

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	V1	.43	.50	26.84	168.61	.64	.95
2	V2	.54	.50	26.72	168.56	.64	.95
3	V3	.83	.38	26.43	172.38	.47	.95
4	V4	.85	.35	26.41	176.05	.10	.95
5	V5	.94	.25	26.33	175.30	.27	.95
6	V6	.96	.20	26.31	175.74	.26	.95
7	V7	.68	.47	26.58	169.72	.59	.95
8	V8	.74	.44	26.53	171.54	.47	.95
9	V9	.54	.50	26.72	169.49	.57	.95
10	V10	.80	.40	26.46	172.22	.46	.95
11	V11	.40	.49	26.86	171.24	.44	.95
12	V12	.70	.46	26.57	170.62	.52	.95
13	V13	.71	.45	26.55	170.43	.55	.95
14	V14	.37	.48	26.90	169.87	.56	.95
15	V15	.61	.49	26.66	169.18	.61	.95
16	V16	.25	.43	27.02	170.25	.59	.95
17	V17	.28	.45	26.98	171.47	.46	.95
18	V18	.29	.45	26.98	171.04	.50	.95
19	V19	.10	.30	27.17	175.09	.25	.95
20	V20	.64	.48	26.63	169.72	.57	.95
21	V21	.81	.39	26.46	172.29	.45	.95
22	V22	.70	.46	26.57	170.08	.57	.95
23	V23	.37	.49	26.89	169.26	.61	.95
24	V24	.63	.49	26.64	168.56	.66	.95
25	V25	.84	.36	26.42	172.60	.46	.95
26	V26	.54	.50	26.72	170.43	.50	.95
27	V27	.62	.49	26.65	168.57	.66	.95
28	V28	.12	.32	27.15	172.87	.49	.95
29	V29	.21	.41	27.06	172.96	.38	.95
30	V30	.16	.37	27.10	172.53	.46	.95
31	V31	.49	.50	26.77	170.53	.49	.95
32	V32	.15	.35	27.12	173.55	.37	.95
33	V33	.08	.27	27.19	174.92	.30	.95
34	V34	.22	.42	27.04	171.67	.48	.95
35	V35	.40	.49	26.86	170.27	.52	.95
36	V36	.21	.41	27.06	171.82	.49	.95
37	V37	.82	.38	26.44	171.69	.53	.95
38	V38	.33	.47	26.93	172.26	.38	.95
39	V39	.62	.49	26.64	170.48	.51	.95
40	V40	.26	.44	27.01	170.99	.52	.95
41	V41	.10	.30	27.16	174.82	.28	.95
42	V42	.15	.36	27.11	173.15	.41	.95
43	V43	.17	.38	27.09	173.52	.35	.95
44	V44	.20	.40	27.07	173.11	.37	.95
45	V45	.37	.48	26.90	168.85	.64	.95
46	V46	.32	.47	26.95	169.29	.63	.95
47	V47	.32	.47	26.95	169.63	.60	.95
48	V48	.27	.44	27.00	169.84	.61	.95

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
49	V49	.85	.36	26.42	173.20	.40	.95
50	V50	.18	.38	27.09	173.25	.37	.95
51	V51	.21	.41	27.06	173.18	.36	.95
52	V52	.16	.37	27.10	172.27	.49	.95
53	V53	.30	.46	26.96	168.96	.67	.95
54	V54	.56	.50	26.71	169.26	.59	.95
55	V55	.74	.44	26.52	171.26	.50	.95
56	V56	.38	.49	26.89	168.32	.68	.95
57	V57	.59	.49	26.67	169.51	.58	.95
58	V58	.17	.38	27.09	173.45	.35	.95
59	V59	.25	.43	27.01	172.45	.39	.95
60	V60	.74	.44	26.53	171.03	.51	.95

Appendix 9 Rasch analyses of the test of vocabulary breadth (data from preliminary study D and the main study combined): Persons-to-items map, model fit, and item difficulty estimates

```
------------------------------------------------------------------------------
   Item Estimates (Thresholds)
   all on all (N=401 L=60 Probability Level=0.50)
------------------------------------------------------------------------------
6.0                                     |
                                        |
                                        |
                                      X |
5.0                                     |
                                        |
                                      X |
                                        |
                                        |
4.0                                     |
                                        |
                                      X |
                                      X |
                                     XX |  33
3.0                                     |
                                  XXXXX |  19
                                  XXXXX |  28
                                   XXXX |  41
                                        |
2.0                           XXXXXXXXX |  30    32
                                   XXXX |  42    43    51    52    58
                                  XXXXX |  29    36    44    50
                             XXXXXXXXXX |  34    40
                         XXXXXXXXXXXXXX |  16    48    59
1.0                    XXXXXXXXXXXXXXXX |  17    18    47
                     XXXXXXXXXXXXXXXXXX |  38    46    53
                       XXXXXXXXXXXXXXXX |  14    23
                                        |
                          XXXXXXXXXXXXX |   1    35    45    56
0.0                      XXXXXXXXXXXXXX |  11
            XXXXXXXXXXXXXXXXXXXXXXXXXXX |  31
               XXXXXXXXXXXXXXXXXXXXXXXX |   9    54
             XXXXXXXXXXXXXXXXXXXXXXXXXX |  26
                   XXXXXXXXXXXXXXXXXXXX |   2    15    27    39    57
-1.0          XXXXXXXXXXXXXXXXXXXXXXXXX |  24
             XXXXXXXXXXXXXXXXXXXXXXXXXX |  20    22
                     XXXXXXXXXXXXXXXXXX |   7    12    13
                         XXXXXXXXXXXXXX |   8    60
                               XXXXXXXX |
-2.0                XXXXXXXXXXXXXXXXXXX |  55
                             XXXXXXXXXX |  10    37
                                  XXXXX |
                                        |
                                   XXXX |   3     4    21    25    49
-3.0                 XXXXXXXXXXXXXXXXXX |
                                        |
                                 XXXXXX |
                                 XXXXXX |
                                     XX |
-4.0                                  5 |   6
                                     XX |
                                        |
                                        |
-5.0                                    |
                                        |
                                        |
                                      X |
                                        |
-6.0                                    |
------------------------------------------------------------------------------
   Each X represents 1 student
==============================================================================
```

```
------------------------------------------------------------------------------------------
Item Fit
all on all (N=401 L=60 Probability Level=0.50)
------------------------------------------------------------------------------------------
INFIT
  MNSQ              0.63      0.71      0.83      1.00      1.20      1.40      1.60
----------------------+---------+---------+---------+---------+---------+---------+------
   1  item  1                       :       *      |              :
   2  item  2                       :         *    |              :
   3  item  3                       :              *              :
   4  item  4                       :              |              :        *
   5  item  5                       :          *   |              :
   6  item  6                       :              *              :
   7  item  7                       :     *        |              :
   8  item  8                       :              |*             :
   9  item  9                       :       *      |              :
  10  item 10                       :       *      |              :
  11  item 11                       :              |          *   :
  12  item 12                       :             *|              :
  13  item 13                       :              *              :
  14  item 14                       :         *    |              :
  15  item 15                       :        *     |              :
  16  item 16                       :       *      |              :
  17  item 17                       :              |     *        :
  18  item 18                       :            * |              :
  19  item 19                       :              |          *   :
  20  item 20                       :             *|              :
  21  item 21                       :        *     |              :
  22  item 22                       :       *      |              :
  23  item 23                       :       *      |              :
  24  item 24                       *              |              :
  25  item 25                       :              |*             :
  26  item 26                       :              |  *           :
  27  item 27                       : *            |              :
  28  item 28                       :      *       |              :
  29  item 29                       :              |    *         :
  30  item 30                       :             *|              :
  31  item 31                       :              |          *   :
  32  item 32                       :              |    *         :
  33  item 33                       :              |  *           :
  34  item 34                       :              |     *        :
  35  item 35                       :              |  *           :
  36  item 36                       :             *|              :
  37  item 37                       :       *      |              :
  38  item 38                       :              |             *:
  39  item 39                       :             *|              :
  40  item 40                       :             *|              :
  41  item 41                       :              |             *:
  42  item 42                       :              |   *          :
  43  item 43                       :              |      *       :
  44  item 44                       :              |          *   :
  45  item 45                       :       *      |              :
  46  item 46                       :      *       |              :
  47  item 47                       :   *          |              :
  48  item 48                       :        *     |              :
  49  item 49                       :              | *            :
  50  item 50                       :              |*             :
  51  item 51                       :              |     *        :
==========================================================================================
```

```
--------------------------------------------------------------------------
Item Fit
all on all (N=401 L=60 Probability Level=0.50)
--------------------------------------------------------------------------
INFIT
 MNSQ             0.63      0.71      0.83      1.00      1.20      1.40      1.60
-------------------+---------+----.----+---------+---------+--.------+---------+----
 52  item 52                        .           *|              .
 53  item 53                        .   *        |              .
 54  item 54                        .          * |              .
 55  item 55                        .            |  *           .
 56  item 56                        *            |              .
 57  item 57                        .     *      |              .
 58  item 58                        .            |     *        .
 59  item 59                        .            |    *         .
 60  item 60                        .            *              .
==========================================================================
```

Item Estimates (Thresholds) In input Order
all on all (N=401 L=60 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
1 item 1	149	401	0.27 .12	0.88	0.87	-2.2	-0.9
2 item 2	221	401	-0.78 .12	0.92	0.81	-1.5	-1.5
3 item 3	337	401	-2.78 .16	0.99	1.02	0.0	0.2
4 item 4	339	401	-2.83 .16	1.46	4.07	4.2	5.5
5 item 5	373	401	-3.95 .22	0.97	1.66	-0.2	1.1
6 item 6	376	401	-4.09 .23	1.00	2.00	0.0	1.4
7 item 7	260	401	-1.35 .13	0.86	0.74	-2.4	-1.7
8 item 8	277	401	-1.62 .13	1.02	1.03	0.3	0.2
9 item 9	194	401	-0.39 .12	0.91	0.81	-1.8	-1.6
10 item 10	317	401	-2.34 .14	0.90	1.03	-1.3	0.2
11 item 11	157	401	0.15 .12	1.23	1.23	3.8	1.6
12 item 12	270	401	-1.51 .13	0.98	1.09	-0.3	0.6
13 item 13	270	401	-1.51 .13	1.01	0.88	0.2	-0.7
14 item 14	136	401	0.47 .13	0.95	0.84	-0.9	-1.1
15 item 15	221	401	-0.78 .12	0.91	0.82	-1.7	-1.4
16 item 16	89	401	1.28 .14	0.89	0.80	-1.5	-0.9
17 item 17	106	401	0.97 .13	1.08	1.70	1.2	3.1
18 item 18	100	401	1.08 .14	0.97	1.26	-0.4	1.3
19 item 19	37	401	2.63 .19	1.19	1.34	1.4	0.9
20 item 20	254	401	-1.26 .12	0.98	0.87	-0.4	-0.8
21 item 21	331	401	-2.64 .15	0.91	1.02	-1.0	0.2
22 item 22	257	401	-1.31 .13	0.89	0.91	-2.0	-0.5
23 item 23	137	401	0.45 .13	0.89	0.81	-1.8	-1.3
24 item 24	232	401	-0.94 .12	0.77	0.63	-4.6	-3.1
25 item 25	334	401	-2.71 .15	1.04	0.91	0.5	-0.2
26 item 26	210	401	-0.62 .12	1.08	1.09	1.5	0.7
27 item 27	227	401	-0.86 .12	0.80	0.68	-4.0	-2.6
28 item 28	43	401	2.42 .18	0.88	1.44	-1.0	1.1

Item Estimates (Thresholds) in input order
all on all (N=401 L=60 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
29 item 29	74	401	1.60 .15	1.12	1.81	1.4	2.6
30 item 30	51	401	2.18 .17	0.97	1.19	-0.2	0.6
31 item 31	180	401	-0.19 .12	1.20	1.25	3.6	1.8
32 item 32	51	401	2.18 .17	1.12	1.05	1.1	0.3
33 item 33	27	401	3.06 .22	1.05	1.35	0.4	0.8
34 item 34	85	401	1.36 .14	1.10	1.28	1.2	1.2
35 item 35	152	401	0.22 .12	1.05	1.05	0.9	0.4
36 item 36	71	401	1.66 .15	0.97	1.49	-0.3	1.7
37 item 37	316	401	-2.32 .14	0.87	0.74	-1.8	-1.0
38 item 38	116	401	0.79 .13	1.24	1.54	3.4	2.7
39 item 39	221	401	-0.78 .12	0.96	0.94	-0.8	-0.4
40 item 40	85	401	1.36 .14	0.96	0.76	-0.4	-1.0
41 item 41	44	401	2.39 .18	1.23	2.06	1.9	2.2
42 item 42	62	401	1.88 .16	1.07	2.23	0.8	3.2
43 item 43	67	401	1.76 .15	1.13	1.45	1.4	1.5
44 item 44	75	401	1.58 .15	1.19	1.59	2.1	2.0
45 item 45	138	401	0.44 .13	0.85	1.18	-2.6	1.2
46 item 46	113	401	0.84 .13	0.83	0.65	-2.7	-2.2
47 item 47	107	401	0.95 .13	0.81	0.69	-2.9	-1.8
48 item 48	98	401	1.11 .14	0.88	0.76	-1.7	-1.2
49 item 49	332	401	-2.66 .15	1.07	0.80	0.8	-0.6
50 item 50	70	401	1.69 .15	1.03	1.00	0.3	0.1
51 item 51	63	401	1.86 .16	1.12	1.05	1.2	0.3
52 item 52	63	401	1.86 .16	0.98	1.24	-0.2	0.9
53 item 53	112	401	0.86 .13	0.81	0.64	-3.0	-2.3
54 item 54	197	401	-0.43 .12	0.97	0.91	-0.6	-0.7
55 item 55	298	401	-1.98 .13	1.06	0.92	0.8	-0.3

Item Estimates (Thresholds) in input order
all on all (N=401 L=60 Probability Level=0.50)

ITEM NAME	SCORE	MAXSCR	THRSH 1	INFT MNSQ	OUTFT MNSQ	INFT t	OUTFT t
56 item 56	138	401	0.44 .13	0.77	0.65	-4.0	-2.7
57 item 57	214	401	-0.68 .12	0.89	0.78	-2.1	-1.8
58 item 58	62	401	1.88 .16	1.17	1.54	1.7	1.7
59 item 59	88	401	1.30 .14	1.11	1.04	1.4	0.3
60 item 60	280	401	-1.67 .13	1.00	3.72	0.1	8.9
Mean			0.00	1.00	1.20	-0.2	0.3
SD			1.76	0.14	0.63	1.9	2.0

Appendix 10 Measures of recognition latency: Classical item analysis

Accuracy Measures

Number Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_N11	.97	.16	15.94	2.88	.42	.70
2	A_N12	.91	.28	16.00	2.83	.25	.71
3	A_N21	.95	.22	15.96	2.83	.36	.70
4	A_N22	.95	.23	15.97	2.85	.32	.71
5	A_N31	.95	.21	15.96	2.98	.17	.72
6	A_N32	.92	.28	16.00	2.97	.11	.73
7	A_N41	.98	.15	15.94	2.87	.50	.70
8	A_N42	.91	.28	16.00	2.74	.34	.70
9	A_N51	.90	.30	16.01	2.81	.25	.71
10	A_N52	.95	.21	15.96	2.82	.40	.70
11	A_N61	.96	.20	15.95	2.86	.36	.70
12	A_N62	.90	.30	16.01	2.76	.30	.71
13	A_N71	.96	.20	15.95	2.88	.33	.71
14	A_N72	.96	.19	15.95	2.84	.43	.70
15	A_N81	.98	.13	15.93	2.85	.60	.69
16	A_N82	.88	.32	16.03	2.78	.24	.72
17	A_N91	.98	.13	15.93	2.85	.60	.69
18	A_N92	.88	.32	16.03	2.91	.11	.73

Real Word Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_W11	.97	.18	6.66	.50	.32	.42
2	A_W12	.97	.18	6.66	.51	.28	.44
3	A_W21	.97	.18	6.66	.54	.16	.47
4	A_W22	.95	.21	6.68	.50	.21	.46
5	A_W31	.97	.18	6.66	.49	.36	.41
6	A_W32	.91	.29	6.72	.48	.12	.51
7	A_W41	.97	.16	6.66	.50	.35	.41
8	A_W42	.92	.27	6.71	.49	.12	.51

Pseudoword Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_P11	.94	.25	6.64	.43	.29	.26
2	A_P12	.92	.27	6.66	.47	.10	.37
3	A_P21	.97	.16	6.61	.51	.18	.33

Pseudoword Matching (continued)

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
4	A_P22	.95	.22	6.63	.50	.12	.35
5	A_P31	.97	.16	6.61	.51	.22	.31
6	A_P32	.93	.25	6.65	.51	.02	.41
7	A_P41	.95	.22	6.63	.48	.18	.32
8	A_P42	.95	.23	6.63	.48	.16	.33

Irregular String Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_I11	.91	.29	6.32	.81	-.06	.30
2	A_I12	.95	.23	6.28	.78	.08	.22
3	A_I21	.86	.34	6.37	.69	.10	.21
4	A_I22	.95	.21	6.27	.76	.17	.18
5	A_I31	.92	.28	6.31	.69	.21	.14
6	A_I32	.92	.27	6.31	.74	.12	.19
7	A_I41	.90	.31	6.33	.73	.08	.22
8	A_I42	.82	.38	6.41	.67	.07	.23

Synonym/Antonym Decision

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	A_SM11	.93	.26	15.26	3.49	.41	.59
2	A_SM12	.91	.29	15.28	3.50	.35	.60
3	A_SM21	.95	.21	15.24	3.71	.24	.61
4	A_SM22	.94	.25	15.26	3.71	.19	.62
5	A_SM31	.92	.28	15.27	3.45	.42	.59
6	A_SM32	.95	.23	15.25	3.73	.20	.62
7	A_SM41	.90	.30	15.29	3.50	.33	.60
8	A_SM42	.88	.33	15.32	3.49	.29	.61
9	A_SM51	.84	.37	15.36	3.64	.12	.63
10	A_SM52	.91	.29	15.28	3.67	.18	.62
11	A_SM61	.83	.38	15.37	3.85	-.03	.66
12	A_SM62	.91	.28	15.28	3.61	.24	.61
13	A_SM71	.90	.30	15.29	3.56	.27	.61
14	A_SM72	.87	.33	15.32	3.45	.32	.60
15	A_SM81	.84	.37	15.35	3.60	.16	.63
16	A_SM82	.87	.33	15.32	3.59	.19	.62
17	A_SM91	.91	.29	15.28	3.60	.25	.61
18	A_SM92	.94	.24	15.25	3.71	.20	.62

Latency Measures

Number Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R_N11	616.53	172.59	12553.32	6497669.17	.50	.90
2	R_N12	734.46	197.23	12435.40	6363754.31	.57	.90
3	R_N21	737.62	216.78	12432.24	6322258.48	.55	.90
4	R_N22	770.37	197.89	12399.49	6393102.79	.54	.90
5	R_N31	707.71	180.17	12462.14	6367709.65	.63	.90
6	R_N32	771.98	278.19	12397.87	6080163.77	.59	.90
7	R_N41	694.79	278.03	12475.06	6179731.35	.52	.90
8	R_N42	820.48	273.38	12349.37	6267858.84	.46	.90
9	R_N51	833.99	326.28	12335.86	5927978.49	.59	.90
10	R_N52	802.35	213.73	12367.51	6248722.54	.63	.90
11	R_N61	666.83	216.70	12503.03	6287770.72	.58	.90
12	R_N62	771.69	252.84	12398.16	6195605.03	.56	.90
13	R_N71	665.74	194.09	12504.12	6306688.38	.64	.90
14	R_N72	743.62	263.21	12426.24	6193681.59	.54	.90
15	R_N81	679.05	270.14	12490.80	6180697.65	.53	.90
16	R_N82	777.54	284.03	12392.32	6097634.46	.56	.90
17	R_N91	628.15	181.03	12541.70	6344342.65	.65	.90
18	R_N92	746.97	219.04	12422.89	6272197.45	.59	.90

Real Word Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R_W11	814.41	267.81	5696.77	1559080.14	.52	.77
2	R_W12	761.75	249.94	5749.42	1611599.02	.48	.77
3	R_W21	856.22	301.31	5654.95	1551521.95	.45	.78
4	R_W22	753.62	285.55	5757.56	1595470.67	.42	.78
5	R_W31	762.09	237.75	5749.08	1606876.22	.52	.77
6	R_W32	863.54	276.84	5647.63	1544284.08	.52	.77
7	R_W41	796.31	270.93	5714.87	1546993.66	.53	.77
8	R_W42	903.23	301.94	5607.94	1458650.72	.59	.76

Pseudoword Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R_P11	823.25	271.26	5827.72	1685445.10	.53	.77
2	R_P12	757.70	263.06	5893.27	1758669.09	.44	.78
3	R_P21	879.05	317.33	5771.92	1613028.97	.52	.77
4	R_P22	753.32	277.69	5897.66	1669986.18	.54	.77
5	R_P31	850.16	301.49	5800.82	1700535.21	.44	.78
6	R_P32	861.69	301.32	5789.28	1629378.42	.54	.77
7	R_P41	864.00	304.24	5786.97	1640355.41	.52	.77
8	R_P42	861.80	232.96	5789.17	1759404.28	.52	.77

Irregular String Matching

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R_I11	1027.86	361.96	6399.57	2123250.00	.50	.75
2	R_I12	733.08	247.66	6694.35	2405807.76	.41	.77
3	R_I21	1113.07	380.59	6314.36	2072388.52	.51	.75
4	R_I22	724.84	245.13	6702.59	2378815.93	.45	.76
5	R_I31	953.86	344.12	6473.57	2257155.70	.39	.77
6	R_I32	906.91	326.90	6520.52	2123763.59	.58	.74
7	R_I41	963.46	344.51	6463.98	2125574.87	.53	.74
8	R_I42	1004.34	383.11	6423.10	2073639.25	.51	.75

Synonym/Antonym Decision

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	R_SM11	1775.56	667.64	29783.42	66293934.61	.58	.90
2	R_SM12	2007.13	942.61	29551.85	64136235.72	.53	.90
3	R_SM21	1573.51	546.65	29985.47	67937953.33	.54	.90
4	R_SM22	1508.25	609.40	30050.73	67363202.37	.53	.90
5	R_SM31	1956.97	894.08	29602.00	62644343.26	.68	.89
6	R_SM32	1275.79	400.74	30283.18	69391416.90	.53	.90
7	R_SM41	1751.11	838.00	29807.86	64081152.69	.62	.89
8	R_SM42	2011.53	882.96	29547.44	64559750.61	.55	.90
9	R_SM51	1752.34	849.61	29806.63	64667965.41	.56	.90
10	R_SM52	1864.90	879.48	29694.07	64968607.18	.52	.90
11	R_SM61	1319.56	384.31	30239.42	70284777.89	.41	.90
12	R_SM62	1717.53	745.71	29841.44	66550160.41	.49	.90
13	R_SM71	1653.76	698.19	29905.21	65373382.47	.64	.89
14	R_SM72	1884.47	727.90	29674.51	65441141.28	.60	.90
15	R_SM81	1564.92	561.96	29994.06	67762569.98	.54	.90
16	R_SM82	2119.25	1083.77	29439.73	62720159.42	.53	.90
17	R_SM91	1983.16	958.67	29575.81	62588623.49	.63	.89
18	R_SM92	1839.21	865.33	29719.76	63217337.24	.66	.89

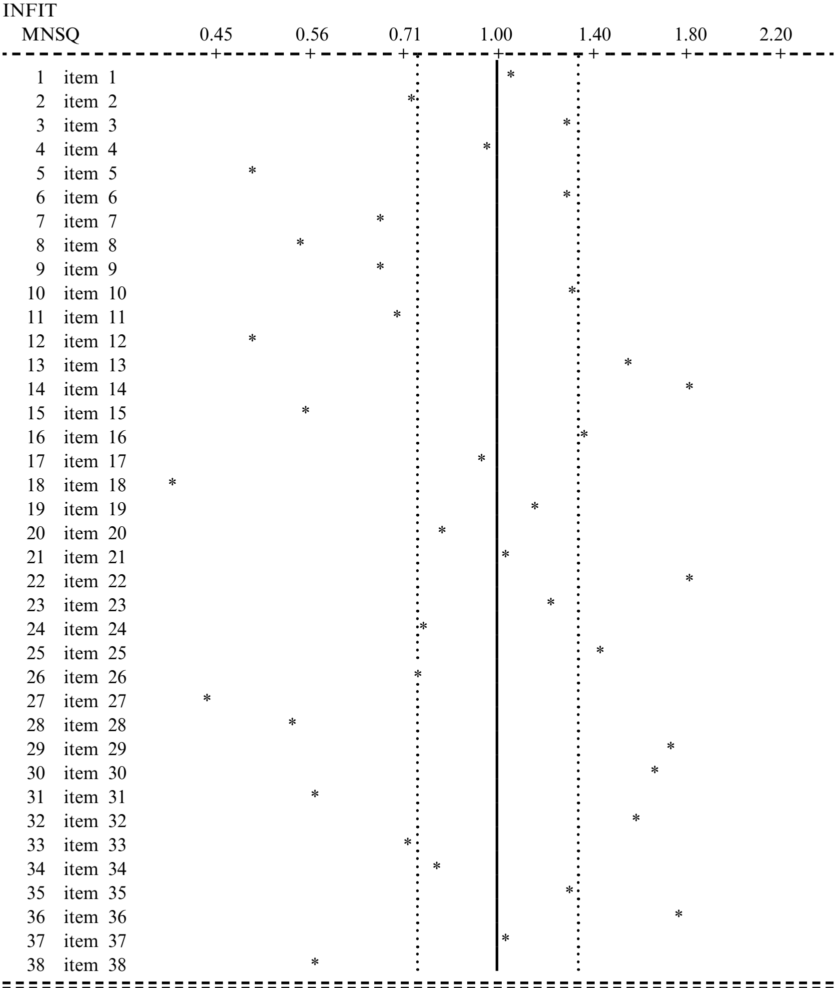
Appendix 11 Measure of language-independent metacognitive knowledge: Classical item analysis

No.	Item Code	Mean	SD	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Alpha if Item Deleted
1	TXA101	.39	.49	26.39	54.24	.42	.90
2	TXA102	.61	.49	26.18	52.40	.69	.90
3	TXA103	.95	.22	25.84	57.77	-.10	.91
4	TXA104	.98	.13	25.80	57.16	.16	.91
5	TXA105	.68	.47	26.11	51.63	.84	.90
6	TXA106	.93	.25	25.85	57.24	.05	.91
7	TXA107	.58	.50	26.21	52.04	.73	.90
8	TXA108	.69	.46	26.10	51.85	.81	.90
9	TXA109	.61	.49	26.18	52.03	.74	.90
10	TXA110	.91	.28	25.87	57.13	.07	.91
11	TXA111	.59	.49	26.20	52.13	.72	.90
12	TXA112	.71	.46	26.08	51.74	.85	.90
13	TXA113	.73	.45	26.06	56.35	.14	.91
14	TXA114	.69	.46	26.09	57.50	-.03	.91
15	RDA201	.62	.49	26.16	51.66	.80	.90
16	RDA202	.80	.40	25.98	56.48	.14	.91
17	RDA203	.98	.15	25.81	57.03	.20	.91
18	RDA204	.68	.47	26.11	51.32	.89	.90
19	RDA205	.94	.24	25.84	57.20	.07	.91
20	RDA206	.49	.50	26.29	52.87	.60	.90
21	RDA207	.94	.25	25.85	56.97	.13	.91
22	RDA208	.68	.47	26.11	57.32	-.01	.91
23	RDA209	.86	.35	25.93	56.47	.17	.91
24	RDA210	.42	.49	26.37	52.98	.59	.90
25	RDA211	.89	.32	25.90	57.35	.01	.91
26	RDA212	.55	.50	26.24	52.49	.66	.90
27	RDA213	.68	.47	26.11	51.45	.87	.90
28	RDA214	.64	.48	26.15	51.66	.81	.90
29	RDA215	.64	.48	26.15	56.70	.08	.91
30	RDA216	.84	.37	25.95	57.58	-.04	.91
31	RDA217	.63	.48	26.16	51.70	.80	.90
32	RDA218	.34	.47	26.45	57.59	-.04	.91
33	RDA219	.49	.50	26.29	52.53	.65	.90
34	RDA220	.42	.49	26.37	53.17	.57	.90
35	RDA221	.83	.38	25.96	56.32	.18	.91
36	RDA222	.80	.40	25.99	57.74	-.07	.91
37	RDA223	.97	.18	25.82	57.21	.10	.91
38	RDA224	.64	.48	26.15	51.76	.80	.90

Appendix 12 Rasch analyses of the measure of language-independent metacognitive knowledge about the text and reading: Persons-to-items map, model fit, and item difficulty estimates

```
--------------------------------------------------------------------------------
    Item Estimates (Thresholds)
    all on all (N=219 L=38 Probability Level=0.50)
--------------------------------------------------------------------------------
 5.0                               |
                                   |
                               XXX |
                                   |
 4.0                               |
                                   |
                            XXXXXX |
                                   |
                                   |
 3.0              XXXXXXXXXXXXXXXX |
                 XXXXXXXXXXXXXXXXX |
                   XXXXXXXXXXXXXXX |
                                   |  32
             XXXXXXXXXXXXXXXXXXXXX |
 2.0             XXXXXXXXXXXXXXXXX |  1
                   XXXXXXXXXXXXXXX |  24      34
                                   |
                   XXXXXXXXXXXXXXX |  20      33
                    XXXXXXXXXXXXXX |
                          XXXXXXXX |  26
 1.0                         XXXXX |  7
                                 X |  2       9       11
                                 X |  15      28      29      31      38
                                   |
                                 X |  5       8       14      18      22
                               XXX |  12
                            XXXXXX |  13
 0.0                          XXXX |
                   XXXXXXXXXXXXXXX |
                    XXXXXXXXXXXXXX |  16      36
                           XXXXXXX |  35
                              XXXX |  30
                             XXXXX |  23
-1.0                           XXX |
                                   |  25
                                XX |
                                 X |  10
                                   |
                                   |  6       21
-2.0                               |  19
                                   |  3
                                   |
                                   |  37
                                   |
-3.0                               |  17
                                   |
                                   |  4
                                   |
                                   |
-4.0                               |
--------------------------------------------------------------------------------
    Each X represents 1 student
    Some thresholds could not be fitted to the display
================================================================================
```

Item Fit
all on all (N=219 L=38 Probability Level=0.50)



Appendix 13 Summary of the significant predictors of reading comprehension performance across separate analyses when syntax raw score was used in place of COPS score

<table>
<tr><th>The undivided sample (n=219)</th><th>The two subgroups (n=219)</th></tr>
<tr><td rowspan="4">Overall Score (R^2=.63)

Syntax (β=.53)
Vocabulary (β=.29)
Lexicality (β=.10)

Passage 1 (R^2=.62)

Syntax (β=.52)
Vocabulary (β=.31)

Passage 2 (R^2=.36)

Syntax (β=.38)
Lexicality (β=.26)

Passage 3 (R^2=.41)

Syntax (β=.50)
Vocabulary β=.31)
Lexicality(β=.12)
Number Latency (β=.11)

Passage 4 (R^2=.39)

Syntax (β=.40)
Vocabulary (β=.27)</td><td>The better readers (n=128)</td></tr>
<tr><td>Overall Score (R^2=.30)

Syntax (β=.54)
Real Word Latency (β=.22)

Passage 1 (R^2=.26)

Syntax

Passage 2 (R^2=.10)

Syntax (β=.26)
Irregular Strings (β=.25)

Passage 3 (R^2=.21)

Syntax

Passage 4 (R^2=.23)

Syntax (β=.34)
Vocabulary (β=.24)</td></tr>
<tr><td>The poorer readers (n=91)</td></tr>
<tr><td>Overall Score (R^2=.34)

Syntax (β=.36)
Lexicality (β=-.20)
Vocabulary (β=.18)

Passage 1 (R^2=.30)

Vocabulary (β=.39)
Syntax (β=.31)

Passage 2 (R^2=.17)

Vocabulary

Passage 3 (R^2=.14)

Lexicality (β=.29)
Syntax (β=.26)

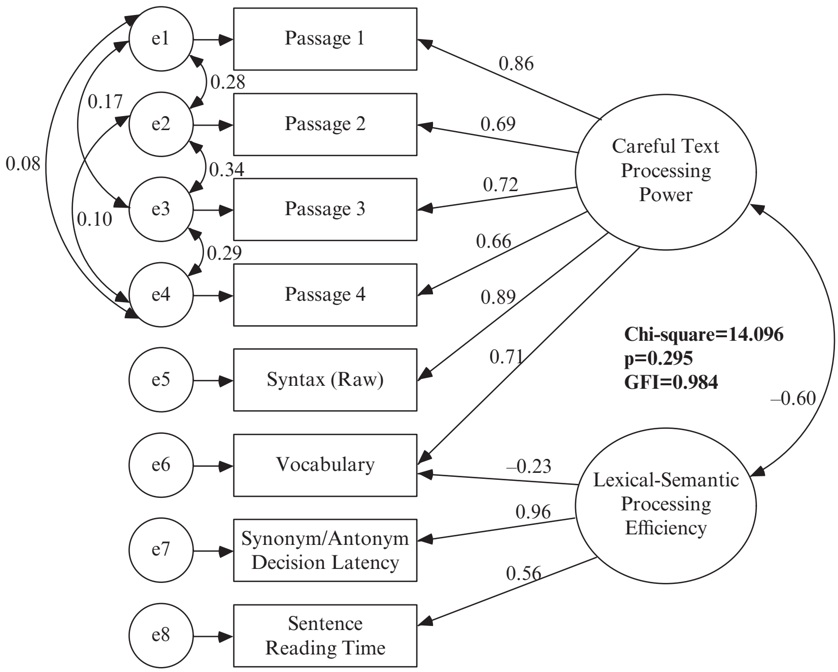
Passage 4 (R^2=.10)

Syntax</td></tr>
</table>

Appendix 14 SEM fit statistics and diagram for the model with the syntax raw score in place of the syntax COPS score

Chi-square	df	p	GFI	AGFI	NFI	CFI	RMSEA
14.096	12	.295	.984	.953	.987	.998	.029

Note: All paths originating from Careful Text Processing Power and Lexical-Semantic Processing Efficiency factors are statistically significant at p<.05.

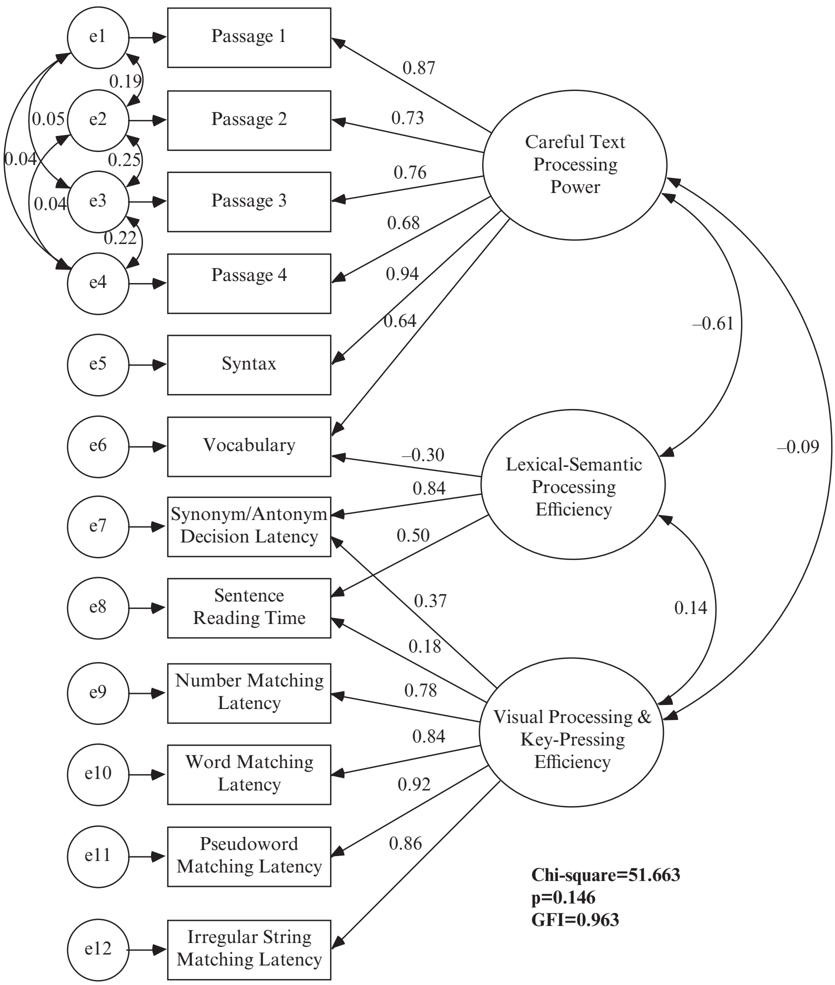


Appendix 15 Results of a post-hoc SEM analysis that explored the separability of lexical-semantic processing efficiency from computer-based visual processing/key-pressing efficiency

To explore the degree to which the Lexical-Semantic Processing Efficiency factor is separable from the efficiency in visual processing and key pressing on computer, another latent variable, Visual Processing/Key-Pressing Efficiency, was added to the model and hypothesised to account for performances in all of the computer-based latency measures (bottom six rectangles). The fit statistics indicate good model fit overall (non-significant Chi-square with a GFI of .96), and the inter-factor correlation between the two latent variables in question (two ovals at bottom right in the diagram) was small (r=.14) and non-significant. Therefore, even though Synonym/Antonym Decision Latency and Sentence Reading Speed/Time were partially accounted for by the Visual Processing and Key-Pressing Efficiency factor, they were much better explained by the Lexical-Semantic Processing Efficiency factor, which is clearly dissociable from the efficiency in non-semantic, visual processing and key-pressing on computer.

Chi-square	df	p	GFI	AGFI	NFI	CFI	RMSEA
51.663	42	.146	.963	.931	.972	.995	.033

Note: All leftward paths originating from Careful Text Processing Power factor, Lexical-Semantic Processing Efficiency factor, and Visual Processing/ Key-Pressing factor are statistically significant at p<.05. The inter-factor relationship was significant only between Careful Text Processing Power factor and Lexical-Semantic Processing Efficiency factor (p<.05).



References

Adams, R J and Khoo, S T (1993) *Quest: The Interactive Test Analysis System (Version 2.1)* [computer program], Camberwell, Australia: Australian Council for Educational Research.

Akamatsu, N (2003) The effects of first language orthographic features on second language reading in text, *Language Learning* 53 (2), 207–231.

Akamatsu, N (2005) Effects of second language reading proficiency and first language orthography on second language word recognition, in Cook, V J and Bassetti, B (Eds) *Second Language Writing Systems*, Clevedon: Multilingual Matters, 238–259.

Alderson, J C (1984) Reading in a foreign language: a reading problem or a language problem? in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language*, Harlow, UK: Longman, 1–24.

Alderson, J C (1993) The relationship between grammar and reading in an English for academic purposes test battery, in Douglas, D and Chapelle, C (Eds) *A New Decade of Language Testing Research: Selected Papers from the 1990 Language Testing Research Colloquium*, Alexandria, VA: TESOL, 203–219.

Alderson, J C (2000) *Assessing Reading,* Cambridge: Cambridge University Press.

Alderson, J C and Urquhart, A H (1984) Introduction: what is reading? in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language,* Harlow, UK: Longman, xv–xxviii.

Anderson, R C and Pearson, P D (1984) A schema-theoretic view of basic processes in reading comprehension, in Pearson, P D and Barr, R (Eds) *Handbook of Reading Research,* New York: Longman, 255–291.

Arbuckle, J and Wothke, W (1999) *AMOS 4.0 User's Guide,* Chicago: Smallwaters.

Bachman, L F (1986) The Test of English as a Foreign Language as a measure of communicative competence, in Stansfield, C W (Ed.) *Toward Communicative Competence Testing: Proceedings of the Second TOEFL Invitational Conference, TOEFL Research Report 21,* Princeton, NJ: Educational Testing Service, 67–88.

Bachman, L F (2004) *Statistical Analyses for Language Assessment*, Cambridge: Cambridge University Press.

Bachman, L, Davidson, F, Ryan, K and Choi, I C (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language,* Cambridge: UCLES/Cambridge University Press.

Baddeley, A, Logie, R, Nimmo-Smith, I and Brereton, N (1985) Components of fluent reading, *Journal of Memory and Language* 24, 119–131.

Baddeley, A D and Hitch, G J (1974) Working memory, in Bower, G (Ed.) *The Psychology of Learning and Motivation, Volume 8,* New York: Academic Press, 47–89.

Beck, I L and McKeown, M (1991) Conditions of vocabulary acquisition, in Barr, R, Kamil, M L, Mosenthal, P and Pearson, P D (Eds) *Handbook of Reading Research, Volume II*, Mahwah, NJ: Lawrence Erlbaum, 789–824.

Beck, I L, Perfetti, C A and McKeown, M G (1982) Effects of long-term vocabulary instruction on lexical access and reading comprehension, *Journal of Educational Psychology* 74, 506–521.

Beglar, D and Hunt, A (1999) Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests, *Language Testing* 16 (2), 131–162.

Berman, R A (1984) Syntactic components of the foreign language reading process, in Alderson, J C and Urquhart, A H (Eds) *Reading in a Foreign Language,* Harlow, UK: Longman, 139–156.

Bernhardt, E (1999) If reading is reader-based, can there be a computer-adaptive test of reading? in Chalhoub-Deville, M (Ed.) *Issues in Computer-Adaptive Testing of Reading Proficiency,* Cambridge: UCLES/Cambridge University Press, 1–10.

Bernhardt, E B (1986) Reading in the foreign language, in Wing, B H (Ed.) *Listening, Reading, Writing: Analysis and Application,* Middlebury, VT: Northeast Conference on the Teaching of Foreign Languages, 93–115.

Bernhardt, E B (1991a) A psycholinguistic perspective on second language literacy, in Hulstijn, J H (Ed.) *Reading in Second Languages: AILA review* 8, 31–44.

Bernhardt, E B (1991b) *Reading Development in a Second Language: Theoretical, Empirical, and Classroom Perspectives,* Norwood, NJ: Ablex.

Bernhardt, E B and Kamil, M L (1995) Interpreting relationships between L1 and L2 reading: consolidating the linguistic threshold and the linguistic interdependence hypotheses, *Applied Linguistics* 16 (1), 15–34.

Bossers, B (1989) Lezen in de tweede taal: een taal- of leesprobleem? [Reading in a second language: a language or a reading problem?] *Toegepaste Taalwetenschap in Artikelen* 38, 176–188.

Bossers, B (1991) On thresholds, ceilings and short-circuits: the relation between L1 reading, L2 reading and L2 knowledge, in Hulstijn, J H (Ed.) *Reading in Second Languages: AILA review* 8, 45–60.

Brisbois, J E (1995) Connections between first- and second-language reading, *Journal of Reading Behavior* 27 (4), 565–584.

Brown, T and Haynes, M (1985) Literacy background and reading development in a second language, in Carr, T H (Ed.) *The Development of Reading Skills*, San Francisco: Jossey-Bass, 19–34.

Brown, T, Carr, T H and Chaderjian, M (1987) Orthography, familiarity, and meaningfulness reconsidered: attention strategies may affect the lexical sensitivity of visual code formation, *Journal of Experimental Psychology: Human Perception and Performance* 13 (1), 127–139.

Bryant, F B and Yarnold, P R (1995) Principal-component analysis and exploratory and confirmatory factor analysis, in Grimm, L G and Yarnold, P R (Eds) *Reading and Understanding Multivariate Statistics*, Washington, DC: American Psychological Association, 99–136.

Byrne, B M (2001) *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*, Mahwah, NJ: Lawrence Erlbaum.

Carr, T H and Levy, B A (1990a) Preface, in Carr, T H and Levy, B A (Eds) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press, xi.

Carr, T H and Levy, B A (Eds) (1990b) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press.

Carrell, P L (1991) Second language reading: reading ability or language proficiency? *Applied Linguistics* 12 (2), 159–179.

Carrell, P L and Eisterhold, J C (1983) Schema theory and ESL reading pedagogy, *TESOL Quarterly* 17 (4), 553–573.
Carson, J E, Carrell, P L, Silberstein, S, Kroll, B and Kuehn, P A (1990) Reading-writing relationships in first and second language, *TESOL Quarterly* 24 (2), 245–266.
Chalhoub-Deville, M (Ed.) (1999) *Issues in Computer-Adaptive Testing of Reading Proficiency*, Cambridge: UCLES/Cambridge University Press.
Chern, C L (1993) Chinese students' word-solving strategies in reading in English, in Huckin, T N, Haynes, M and Coady, J (Eds) *Second Language Reading and Vocabulary Learning*, Norwood, NJ: Ablex, 67–85.
Clapham, C (1996) *The Development of IELTS: A Study of the Effect of Background Knowledge on Reading Comprehension*, Cambridge: UCLES/Cambridge University Press.
Clarke, M A (1978) Reading in Spanish and English: evidence from adult ESL students, *Language Learning* 29 (1), 121–150.
Clarke, M A (1980) The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance, *Modern Language Journal* 64, 203–209.
Coady, J (1979) A psycholinguistic model of the ESL reader, in Mackay, R, Barkman, B and Jordan, R R (Eds) *Reading in a Second Language: Hypotheses, Organization, and Practice*, Rowley, MA: Newbury House, 5–12.
Coady, J and Huckin, T N (Eds) (1997) *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*, Cambridge: Cambridge University Press.
Cohen, A D (1993) *Assessing Language Ability in the Classroom*, Boston: Heinle and Heinle.
Cohen, J D, MacWhinney, B, Flatt, M and Provost, J (1993) PsyScope: a new graphic interactive environment for designing psychology experiments, *Behavioral Research Methods, Instruments, and Computers* 25 (2), 257–271.
Cormier, P and Dea, S (1997) Distinctive patterns of relationship of phonological awareness and working memory with reading comprehension, *Reading and Writing: An Interdisciplinary Journal* 9, 193–206.
Cowan, N (1996) Short-term memory, working memory, and their importance in language processing, *Topics in Language Disorders* 17 (1), 1–18.
Crawford, M J (2005) Adding variety to word recognition exercises, *English Teaching Forum* 43 (2), 36–41.
Cummins, J (1979) Linguistic interdependence and the educational development of bilingual children, *Review of Educational Research* 49, 222–251.
Cunningham, A E, Stanovich, K E and Wilson, M R (1990) Cognitive variation in adult college students differing in reading ability, in Carr, T H and Levy, B A (Eds) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press, 129–159.
Da Fontoura, H A and Siegel, L S (1995) Reading, syntactic, and working memory skills of bilingual Portuguese-English Canadian children, *Reading and Writing: An Interdisciplinary Journal* 7, 139–153.
Daneman, M (1991) Individual differences in reading skills, in Barr, R, Kamil, M L, Mosenthal, P and Pearson, P D (Eds) *Handbook of Reading Research, Volume II*, Mahwah, NJ: Lawrence Erlbaum, 512–538.
Daneman, M and Carpenter, P A (1980) Individual differences in working memory and reading, *Journal of Verbal Learning and Verbal Behaviour* 19, 450–466.

Daneman, M and Carpenter, P A (1983) Individual differences in integrating information between and within sentences, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 9 (4), 561–584.
Davies, A (1990) *Principles of Language Testing*, Oxford: Basil Blackwell.
Dixon, P, LeFevre, J A and Twilley, L C (1988) Word knowledge and working memory as predictors of reading skill, *Journal of Educational Psychology* 80 (4), 465–472.
Duran, R P, Canale, M, Penfield, J, Stansfield, C W and Liskin-Gasparro, J E (1985) *TOEFL from a Communicative Viewpoint on Language Proficiency: A Working Paper*, Princeton, NJ: Educational Testing Service.
Durgunoğlu, A Y, Nagy, W E and Hancin-Bhatt, B J (1993) Cross-language transfer of phonological awareness, *Journal of Educational Psychology* 85 (3), 453–465.
Ellis, N C (2005) At the interface: dynamic interactions of explicit and implicit language knowledge, *Studies in Second Language Acquisition* 27 (2), 305–352.
Ellis, R (2005) Measuring implicit and explicit knowledge of a second language: a psychometric study, *Studies in Second Language Acquisition,* 27 (2), 141–172.
Engle, R W, Nations, J K and Cantor, J (1990) Is "working memory capacity" just another name for word knowledge? *Journal of Educational Psychology* 82 (4), 799–804.
Erlam, R (2006) Elicited imitation as a measure of L2 implicit knowledge: an empirical validation study, *Applied Linguistics* 27, 464–491.
Eskey, D E (1988) Holding in the bottom: an interactive approach to the language problems of second language readers, in Carrell, P L, Devine, J and Eskey, D E (Eds) *Interactive Approaches to Second Language Reading*, Cambridge: Cambridge University Press, 93–100.
Fender, M (2003) English word recognition and word integration skills of native Arabic-speaking and Japanese-speaking learners of English as a second language, *Applied Psycholinguistics* 24, 289–315.
Field, M L (1985) A psycholinguistic model of the Chinese ESL reader, in Larson, P, Judd, E and Messerschmitt, D S (Eds) *On TESOL '84, a Brave New World for TESOL: Selected Papers from the Eighteenth Annual Convention of Teachers of English to Speakers of Other Languages, Houston, Texas, March 6–11, 1984*, Washington, DC: TESOL, 171–183.
Folse, K (2007) *Rapid Reading Practices: Developing Word Automaticity in Reading,* Ann Arbor: University of Michigan Press.
Gathercole, S E and Baddeley, A D (1993) *Working Memory and Language*, Howe, UK: Lawrence Erlbaum.
Geva, E and Ryan, E B (1993) Linguistic and cognitive correlates of academic skills in first and second languages, *Language Learning* 43 (1), 5–42.
Goodman, K S (1967) Reading: a psycholinguistic guessing game, *Journal of the Reading Specialist* 6, 126–135.
Goswami, U and Bryant, P (1990) *Phonological Skills and Learning to Read*, Hove, UK: Lawrence Erlbaum.
Gough, P B (1972) One second of reading, in Kavanagh, J F and Mattingly, I G (Eds) *Language by Ear and by Eye: The Relationships Between Speech and Reading*, Cambridge, MA: MIT Press, 331–358.
Grabe, W (1991) Current developments in second language reading research, *TESOL Quarterly* 25, 375–406.

Grabe, W and Stoller, F L (2002) *Teaching and Researching Reading*, Harlow, UK: Longman.
Green, S B (1991) How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research* 26, 499–510.
Hacquebord, H (1989) *Tekstbegrip van Turkse en Nederlandse leerlingen in het voortgezet onderwijs [Reading Comprehension of Turkish and Dutch Students Attending Secondary Schools]*, Groningen: RUG.
Han, Y and Ellis, R (1998) Implicit knowledge, explicit knowledge, and general language proficiency, *Language Teaching Research* 2 (1), 1–23.
Hansen, J and Bowey, J A (1994) Phonological analysis skills, verbal working memory, and reading ability in second-grade children, *Child Development* 65, 938–950.
Harrington, M and Sawyer, M (1992) L2 working memory capacity and L2 reading skill, *Studies in Second Language Acquisition* 14, 25–38.
Haynes, M and Carr, T H (1990) Writing system background and second language reading: a component skills analysis of English reading by native speaker-readers of Chinese, in Carr, T H and Levy, B A (Eds) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press, 375–421.
Haynes, M E (1989) *Individual Differences in Chinese Readers of English: Orthography and Reading*, unpublished PhD dissertation, Michigan State University, East Lansing.
Heaton, J B (1975) *Writing English Language Tests: A Practical Guide for Teachers of English as a Second or Foreign Language*, London: Longman.
Henning, G (1987) *A Guide to Language Testing: Development, Evaluation, Research*, Cambridge, MA: Newbury House.
Hoover, W A and Tunmer, W E (1993) The components of reading, in Thompson, G B, Tunmer, W E and Nicholson, T (Eds) *Reading Acquisition Processes*, Philadelphia: Multilingual Matters, 1–19.
Hoyle, R H (1995) The structural equation modeling approach: basic concepts and fundamental issues, in Hoyle, R H (Ed.) *Structural Equation Modeling: Concepts, Issues, and Applications*, Thousand Oaks, CA: Sage, 1–15.
Huckin, T N, Haynes, M and Coady, J (Eds) (1993) *Second Language Reading and Vocabulary Learning*, Norwood, NJ: Ablex.
Hughes, A (1989) *Testing for Language Teachers*, Cambridge: Cambridge University Press.
Hulstijn, J H (1991) How is reading in a second language related to reading in a first language? in Hulstijn, J H (Ed.) *Reading in Second Languages: AILA Review* 8, 5–14.
Jackson, M D and McClelland, J L (1979) Processing determinants of reading speed, *Journal of Experimental Psychology: General* 108 (2), 151–181.
Jolly, D (1978) The establishment of a self-access scheme for intensive reading, paper presented at the Goethe-Institut, British Council Colloquium on Reading, Paris.
Joyce, P (2008) *Linguistic Knowledge and Psycholinguistic Processing Skills as Components of L2 Listening Comprehension,* unpublished PhD thesis, Roehampton University, London.
Juffs, A (2005) The influence of first language on the processing of wh-movement in English as a second language, *Second Language Research* 21 (2), 121–151.
Just, M A and Carpenter, P A (1980) A theory of reading: from eye fixations to comprehension, *Psychological Review* 87 (4), 329–354.

Just, M A and Carpenter, P A (1992) A capacity theory of comprehension: individual differences in working memory, *Psychological Review* 99 (1), 122–149.

Just, M A, Carpenter, P A and Woolley, J D (1982) Paradigms and processes in reading comprehension, *Journal of Experimental Psychology: General*, 111(2), 228–238.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Cambridge: UCLES/Cambridge University Press.

King, J and Just, M A (1991) Individual differences in syntactic processing: the role of working memory, *Journal of Memory and Language* 30, 580–602.

Kintsch, W and van Dijk, T A (1978) Toward a model of text comprehension and production, *Psychological Review* 85 (5), 363–394.

Koda, K (1994) Second language reading research: problems and possibilities, *Applied Psycholinguistics* 15, 1–27.

Koda, K (1996) L2 word recognition research: a critical review, *Modern Language Journal* 80 (iv), 450–460.

Koda, K (1998) The role of phonemic awareness in second language reading, *Second Language Research* 14 (2), 194–215.

Koda, K (2005) *Insights into Second Language Reading: A Cross-linguistic Approach*, Cambridge: Cambridge University Press.

Kučera, H and Francis, W N (1967) *Computational Analysis of Present-day American English*, Providence, RI: Brown University Press.

Kunnan, A J (1995) *Test Taker Characteristics and Test Performance: A Structural Modeling Approach*, Cambridge: UCLES/Cambridge University Press.

Kunnan, A J (1998) An introduction to structural equation modelling for language assessment research, *Language Testing* 15 (3), 295–332.

Laufer, B (1992a) Reading in a foreign language: how does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading* 15 (2), 95–103.

Laufer, B (1992b) *The Lexical Threshold of Second Language Reading Comprehension: What it is and How it Relates to L1 Reading Ability*, paper presented at the Second EUROSLA Conference, Jyväskylä.

Laufer, B and Nation, P (1995) Vocabulary size and use: lexical richness in L2 written production, *Applied Linguistics* 16 (3), 305–322.

Laufer, B and Nation, P (1999) Vocabulary-size test of controlled productive ability, *Language Testing* 16 (1), 33–51.

Laufer, B, Elder, C, Hill, K and Congdon, P (2004) Size and strength: do we need both to measure vocabulary knowledge? *Language Testing* 21 (2), 202–226.

Leather, C V and Henry, L A (1994) Working memory span and phonological awareness tasks as predictors of early reading ability, *Journal of Experimental Child Psychology* 58, 88–111.

Lee, J W and Schallert, D L (1997) The relative contribution of L2 language proficiency and L1 reading ability to L2 reading performance: a test of the threshold hypothesis in an EFL context, *TESOL Quarterly* 31 (4), 713–739.

Levy, B A and Carr, T H (1990) Component process analysis: conclusions and challenges, in Carr, T H and Levy, B A (Eds) *Reading and its Development: Component Skills Approaches*, San Diego: Academic Press, 424–438.

Licht, M H (1995) Multiple regression and correlation, in Grimm, L G and Yarnold, P R (Eds) *Reading and Understanding Multivariate Statistics*, Washington, DC: American Psychological Association, 19–64.

MacDonald, M C, Just, M A and Carpenter, P A (1992) Working memory constraints on the processing of syntactic ambiguity, *Cognitive Psychology* 24, 56–98.
Mardia, K V (1970) Measures of multivariate skewness and kurtosis with applications, *Biometrika* 57, 519–530.
Masson, M E J and Miller, J A (1983) Working memory and individual differences in comprehension and memory of text, *Journal of Educational Psychology* 75 (2), 314–318.
McLaughlin, B, Rossman, T and McLeod, B (1983) Second language learning: an information-processing perspective, *Language Learning* 33 (2), 135–158.
McNamara, T F (1996) *Measuring Second Language Performance*, London: Longman.
Meara, P (1992) Vocabulary in a second language: volume 3, *Reading in a Foreign Language* 9 (1), 761–844.
Meara, P (1996) The dimensions of lexical competence, in Brown G, Malmkjaer, K and Williams, J (Eds) *Performance and Competence in Second Language Acquisition*, Cambridge: Cambridge University Press, 35–53.
Meara, P and Jones, G (1990) *The Eurocentres 10K Vocabulary Size Test*, Zurich: Eurocentres.
Meara, P M (1983) *Vocabulary in a Second Language: Volume 1*, London: Centre for Information on Language Teaching and Research.
Meara, P M (Ed.) (1987) *Vocabulary in a Second Language: Volume 2*, London: Centre for Information on Language Teaching and Research.
Miyake, A, Just, M A and Carpenter, P A (1994) Working memory constraints on the resolution of lexical ambiguity: maintaining multiple interpretations in neutral contexts, *Journal of Memory and Language* 33, 175–202.
Möbius, P (1995) Individual differences in 'working memory', verbal efficiency and semantic processing in German-English bilinguals: an exploratory study, *Working Papers in English and Applied Linguistics* 2, 27–41.
Muljani, D, Koda, K and Moates, D R (1998) The development of word recognition in a second language, *Applied Psycholinguistics* 19, 99–113.
Nassaji, H and Geva, E (1999) The contribution of phonological and orthographic processing skills to adult ESL reading: evidence from native speakers of Farsi, *Applied Psycholinguistics* 20, 241–267.
Nation, I S P (1990) *Teaching and Learning Vocabulary*, New York: Newbury House.
Nation, I S P (2001) *Learning Vocabulary in Another Language*, Cambridge: Cambridge University Press.
Negishi, M (1996) *The Latent Trait Structure of L2 Reading Comprehension Tests,* unpublished PhD thesis, University of Reading.
Osaka, M and Osaka, N (1992) Language-independent working memory as measured by Japanese and English reading spans, *Bulletin of Psychonomic Society* 30 (4), 287–289.
Osaka, M and Osaka, N (1994) Working memory capacity related to reading: measurement with the Japanese version of reading span test, *The Japanese Journal of Psychology* 65 (5), 339–345.
Osaka, M, Osaka, N and Groner, R (1993) Language-independent working memory: evidence from German and French reading span tests, *Bulletin of the Psychonomic Society* 31 (2), 117–118.
Palmer, J, MacLeod, C M, Hunt, E and Davidson, J E (1985) Information processing correlates of reading, *Journal of Memory and Language* 24, 59–88.

Paran, A (1994) *Reading Processes in a First and in a Second Language: An On-line Comparison of English and Hebrew,* unpublished PhD thesis, University of Reading.
Paran, A (1996) Reading in EFL: facts and fictions, *ELT Journal* 50 (1), 25–33.
Perkins, K, Brutten, S R and Pohlmann, J T (1989) First and second language reading comprehension, *RELC Journal* 20 (2), 1–9.
Purpura, J E (1998) *LTRC workshop: structural equation modeling on EQS*, unpublished manuscript.
Purpura, J E (1999) *Learner Strategy use and Performance on Language Tests: A Structural Equation Modeling Approach*, Cambridge: UCLES/Cambridge University Press.
Qian, D D (1999) Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension, *The Canadian Modern Language Review* 56 (2), 282–307.
Qian, D D (2002) Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective, *Language Learning* 52 (3), 513–536.
Qian, D D and Schedl, M (2004) Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance, *Language Testing* 21 (1), 28–52.
Rankin, J L (1993) Information-processing differences of college-age readers differing in reading comprehension and speed, *Journal of Reading Behavior* 25 (3), 261–278.
Rayner, K and Pollatsek, A (1989) *The Psychology of Reading*, Englewood Cliffs, NJ: Prentice Hall.
Read, C, Zhang, Y F, Nie, H Y and Ding, B Q (1986) The ability to manipulate speech sounds depends on knowing alphabetic writing, *Cognition* 24, 31–44.
Read, J (1997) Vocabulary and testing, in Schmitt, N and McCarthy, M (Eds) *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press, 303–320.
Reicher, G M (1969) Perceptual recognition as a function of meaningfulness of stimulus material, *Journal of Experimental Psychology* 81, 275–280.
Richards, J (1976) The role of vocabulary teaching, *TESOL Quarterly* 10 (1), 77–89.
Richards, J C, Platt, J and Platt, H (1992) *Longman Dictionary of Language Teaching and Applied Linguistics*, Harlow, UK: Longman.
Rosner, J and Simon, D P (1971) The Auditory Analysis Test: an initial report, *Journal of Learning Disabilities* 4, 384–392.
Rumelhart, D E (1977) Toward an interactive model of reading, in Dornic, S (Ed.) *Attention and Performance VI*, Hillsdale, NJ: Lawrence Erlbaum.
Rupp, A A, Ferne, T and Choi, H (2006) How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective, *Language Testing* 23 (4), 441–474.
Sasaki, M (1993) Relationships among second language proficiency, foreign language aptitude, and intelligence: a structural equation modelling approach, *Language Learning* 43 (3), 313–344.
Schmitt, N and McCarthy, M (Eds) (1997) *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge: Cambridge University Press.
Schmitt, N, Schmitt, D and Clapham, C (2001) Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test, *Language Testing* 18 (1), 55–88.

Schoonen, R, Hulstijn, J and Bossers, B (1998) Metacognitive and language-specific knowledge in native and foreign language reading comprehension: an empirical study among Dutch students in grades 6, 8 and 10, *Language Learning* 48 (1), 71–106.

Schoonen, R, van Gelderen, A, de Glopper, K, Hulstijn, J, Simis, A, Snellings, P and Stevenson, M (2003) First language and second language writing: the role of linguistic knowledge, speed of processing, and metacognitive knowledge, *Language Learning* 53 (1), 165–202.

Segalowitz, N (1986) Skilled reading in the second language, in Vaid, J (Ed.) *Language Processing in Bilinguals: Psycholinguistic and Neuropsychological Perspectives*, Hillsdale, NJ: Lawrence Erlbaum, 3–19.

Segalowitz, N, Poulsen, C and Komoda, M (1991) Lower level components of reading skill in higher level bilinguals: implications for reading instruction, in Hulstijn, J H (Ed.) *Reading in Second Languages: AILA Review* 8, 5–14.

Segalowitz, N, Watson, V and Segalowitz, S (1995) Vocabulary skill: single-case assessment of automaticity of word recognition in a timed lexical decision task, *Second Language Research* 11 (2), 121–136.

Segalowitz, N S and Segalowitz, S J (1993) Skilled performance, practice, and the differentiation of speed-up from automatization effects: evidence from second language word recognition, *Applied Psycholinguistics* 14, 369–385.

Segalowitz, S J, Segalowitz, N S and Wood, A G (1998) Assessing the development of automaticity in second language word recognition, *Applied Psycholinguistics* 19 (1), 53–67.

Service, E (1992) Phonology, working memory, and foreign language learning, *The Quarterly Journal of Experimental Psychology* 45A (1), 21–50.

Shiffrin, R and Schneider, W (1977) Controlled and automatic human information processing: II, Perceptual learning, automatic attending, and a general theory, *Psychological Review* 84, 127–190.

Shiotsu, T and Weir, C J (2007) The relative significance of vocabulary breadth and syntactic knowledge in the prediction of reading comprehension test performance, *Language Testing* 24 (1), 99–128.

Shizuka, T (2000) *The Validity of Incorporating Reading Speed and Response Confidence in Measurement of EFL Reading Proficiency,* unpublished PhD thesis, University of Reading.

Shizuka, T (2004) *New Horizons in Computerized Testing of Reading*, Suita, Japan: Kansai University Press.

Siegel, L S (1994) Working memory and reading: a life-span perspective, *International Journal of Behavioral Development* 17 (1), 109–124.

Siegel, L S and Ryan, E B (1989) The development of working memory in normally achieving and subtypes of learning disabled children, *Child Development* 60, 973–980.

Smith, F (1971) *Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read*, New York: Holt.

Spolsky, B (1995) *Measured Words: The Development of Objective Language Testing*, Oxford: Oxford University Press.

Stanovich, K E (1980) Toward an interactive-compensatory model of individual differences in the development of reading fluency, *Reading Research Quarterly* 16 (1), 32–72.

Stanovich, K E (1991) Word recognition: changing perspectives, in Barr, R, Kamil, M L, Mosenthal, P and Pearson, P D (Eds) *Handbook of Reading Research*, *Volume II*, Mahwah, NJ: Lawrence Erlbaum, 418–452.

Stanovich, K E, Cunningham, A E and Cramer, B B (1984) Assessing phonological awareness in kindergarten children: issues of task comparability, *Journal of Experimental Child Psychology* 38, 175–190.
Swanson, H L (1992) Generality and modifiability of working memory among skilled and less skilled readers, *Journal of Educational Psychology* 84 (4), 473–488.
Swanson, H L and Alexander, J E (1997) Cognitive processes as predictors of word recognition and reading comprehension in learning-disabled and skilled readers: revisiting the specificity hypothesis, *Journal of Educational Psychology* 89 (1), 128–158.
Taillefer, G F (1996) L2 reading ability: further insight into the short-circuit hypothesis, *The Modern Language Journal* 80 (iv), 461–477.
Tunmer, W E, Nesdale, A R and Wright, A D (1987) Syntactic awareness and reading acquisition, *British Journal of Developmental Psychology* 5, 25–34.
Turner, M L and Engle, R W (1989) Is working memory capacity task dependent? *Journal of Memory and Language* 28, 127–154.
Ulijn, J M (1981) Conceptual and syntactic strategies in reading a foreign language, in Hopkins, E and Grotjahn, R (Eds) *Studies in Language Teaching and Language Acquisition*, Bochum, Germany: Brockmeyer, 129–166.
Ulijn, J M (1984) Reading for professional purposes: psycholinguistic evidence in a cross-linguistic perspective, in Pugh, A K and Ulijn, J M (Eds) *Reading for Professional Purposes*, London: Heinemann, 66–81.
Urquhart, A H and Weir, C J (1998) *Reading in a Second Language: Process, Product, and Practice*, New York: Longman.
van Gelderen, A, Schoonen, R, de Glopper, K, Hulstijn, J, Simis, A, Snellings, P and Stevenson, M (2004) Linguistic knowledge, processing speed and metacognitive knowledge in first and second language reading comprehension: a componential analysis, *Journal of Educational Psychology* 96 (1), 19–30.
van Gelderen, A, Schoonen, R, de Glopper, K, Hulstijn, J, Snellings, P, Simis, A, and Stevenson, M (2003) Roles of linguistic knowledge, metacognitive knowledge and processing speed in L3, L2 and L1 reading comprehension: a structural equation modeling approach, *International Journal of Bilingualism* 7 (1), 7–25.
Wade-Woolley, L (1999) First language influences on second language word reading: all roads lead to Rome, *Language Learning* 49 (3), 447–471.
Walter, C (2000) *The Involvement of Working Memory in Reading in a Foreign Language,* unpublished PhD thesis, University of Cambridge.
Wang, M and Koda, K (2005) Commonalities and differences in word identification skills among learners of English as a second language, *Language Learning* 55 (1), 71–98.
Wang, M, Koda, K and Perfetti, C A (2003) Alphabetic and nonalphabetic L1 effects in English word identification: a comparison of Korean and Chinese English L2 learners, *Cognition* 87 (2), 129–149.
Waters, G S and Caplan, D (1996) The measurement of verbal working memory capacity and its relation to reading comprehension, *The Quarterly Journal of Experimental Psychology, A, Human Experimental Psychology* 49 (1), 51–75.
Weir, C J (1983) *Identifying the Language Needs of Overseas Students in Tertiary Education in the United Kingdom,* unpublished PhD thesis, Institute of Education, University of London.

Weir, C J (1993) *Understanding and Developing Language Tests*, New York: Prentice Hall.
Weir, C J, Yang, H and Jin, Y (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes*, Cambridge: UCLES/Cambridge University Press.
West, M (1953) *A General Service List of English Words*, London: Longman.
West, R and Stanovich, K E (1978) Automatic contextual facilitation in readers of three ages, *Child Development* 49, 717–727.
Yamashita, J (1999) *Reading in a First and a Foreign Language: A Study of Reading Comprehension in Japanese (the L1) and English (the L2),* unpublished PhD thesis, Lancaster University.
Yamashita, J (2001) Transfer of L1 reading ability to L2 reading: an elaboration of the linguistic threshold, *Studies in Language and Culture: The Journal of Nagoya University Graduate School of Languages and Cultures* 23 (1), 189–200.
Yang, H and Weir, C (1998) *Validation Study of the College English Test*, Shanghai: China National College English Testing Committee.
Yorio, C A (1971) Some sources of reading problems for foreign language learners, *Language Learning* 21 (1), 107–115.
Yoshida, H, Yoshida, S and Kobayashi, T (1992) Eibun direct reading CAI no kaihatsu to jissen [The development and practice of CAI for direct reading in English], *CAI Gakkaishi* 9, 147–157.
Yoshida, S (2001) Konpyuutaa riyoo gaikokugo gakushuu ni okeru dokkai shidou [Teaching of reading with the use of CALL], in Kadota, S and Noro, T (Eds) *How the Mind Works in EFL Reading*, Tokyo: Kuroshio-shuppan.

Author index

Subject index